Federal Ministry
Republic of Austria
Social Affairs, Health, Care
and Consumer Protection

The Proposal for an Artificial Intelligence Act COM(2021) 206 from a Consumer Policy Perspective

Univ.-Prof. Dr. Christiane Wendehorst, LL.M.
Professor of Law at the University of Vienna

Vienna, 14 December 2021

Imprint and Copyright

Media owner and publisher:
Federal Ministry of Social Affairs, Health, Care and Consumer Protection
Stubenring 1, 1010 Vienna, Austria

Disclaimer:
The opinions expressed in this document are the sole responsibility of the author and do not necessarily represent the official position of the Federal Ministry for Social Affairs, Health, Care and Consumer Protection.

Publishing and production location: Vienna
Vienna, 2021

Note that this Study has been commissioned in two Parts:
Part I: The qualification of AI systems and practices as prohibited or high-risk
Part II: Individual rights, liability, and enforcement

Part I was already published separately on 4 October 2021. This document includes an updated version of Part I

Table of Contents

Executive Summary

The Proposal for an Artificial Intelligence Act (AIA) of 21 April 2021, COM(2021) 206 final, is a landmark document for the regulation of AI worldwide. In the light of the massive risks for consumers (alongside opportunities) that come with the mass roll-out of AI systems it is of utmost importance to make sure this future instrument, together with other instruments, provides an adequate level of protection for consumers in the AI context.

At first sight, the AIA Proposal seems to turn a blind eye on consumer interests as such. This impression is created, e.g., by the fact that the prohibition of manipulation by subliminal techniques and of exploitation of vulnerabilities is restricted to cases where such AI practices cause, or are likely to cause, physical or psychological harm (excluding mere economic harm). The same impression is created by the way the list in Annex III on high-risk AI systems is currently phrased, as this list includes credit scoring, but not the evaluation of factors similar to creditworthiness (such as complaints history), risk assessment by insurance companies or personalised pricing in general.

The list of prohibited AI practices

Clearly, those who drafted the AIA Proposal have done so with the intention to fill gaps in existing legislation, but at the same time to avoid any sort of overlap with existing legislation. The omission of manipulation by subliminal techniques and exploitation of group-specific vulnerabilities that causes mere economic harm in Article 5, for instance, can be explained by the desire on the part of the drafters not to come close to the domain of the Unfair Commercial Practices Directive (UCPD).

In the case of Article 5 AIA Proposal the policy of avoiding overlap comes at the price of many gaps, for instance with regard to manipulation or exploitation of individuals acting as MSME and not qualifying for consumer protection, manipulation or exploitation of other than transactional decisions in a commercial context, and with regard to the placing on the market of relevant AI systems as such.

The policy of avoiding overlap with existing legislative regimes also comes at the price of a regulatory regime that looks, at least at first sight, rather arbitrary in its policy choices

(e.g. appearing to neglect consumer interests). As Article 5 cannot be properly understood without analysing it within the wider framework of existing data protection law, non-discrimination law, consumer protection law and competition law there is a risk that the AIA will not be properly understood and applied by stakeholders throughout Europe. Also, one should not forget that the AIA has the potential of becoming a global role model for the regulation of AI applications, and in order to fulfil this role, it must be easy to understand and reflect the underlying policy choices and assumptions in a consistent manner. A piece of legislation which is understood only by very few experts worldwide, because in order to understand it one has to have a very profound knowledge of the remaining acquis and the scope of application of various other legal instruments, will not easily become a legal instrument from which other regions in the world draw inspiration. Keeping the scope and regulatory focus of different legal instruments apart makes sense where overlap would create the risk of inconsistencies and/or of unnecessary cumulative effects of varying sets of requirements. However, within a blacklist of prohibited practices, it is not necessary to avoid overlap. Rather, it is entirely acceptable (and in fact an indication of coherence and consistency of the acquis) to have harmful and manipulative practices banned by not only one, but by two or several legal instruments at EU level.

It is therefore recommended to remove the restriction to 'physical or psychological' harm in Article 5 (1) (a) and (b) and to replace it by 'material and unjustified' harm. This would both make sure that economic interests of consumers are duly captured and avoid overreaching application of the prohibition beyond what is intended. Apart from that, it is important to extend the prohibition of the exploitation of group-specific vulnerabilities to vulnerabilities resulting from natural person's economic and social situation. Going one step further, it is recommended to extend the provision from group-specific vulnerabilities to individual vulnerabilities. In the consumer context, the exploitation of very individual vulnerabilities (such as characteristic weaknesses and addictions) that are disclosed through extensive data collection and profiling are becoming more and more problematic. It is also recommended to extend the prohibition of social scoring to social scoring activities conducted by private parties, such as where a gatekeeper platform rates consumers on the basis of their social behaviour in a range of different contexts.

Beyond the consumer context, it would be advisable to add comprehensive or total surveillance in an individual's private life or at the workplace to the list of prohibited practices. The same holds true for the specific technical processing of brain data through brain-computer-interfaces (BCIs) in order to read or manipulate a person's thoughts

against that person's will, in a manner that causes or is likely to cause that person material and unjustified harm (cf. the right to mental privacy and integrity).

In any case, it is of the essence that reference be made to prohibitions following from other laws, including data protection law, non-discrimination law, consumer protection law, and competition law. Also, the Commission must be empowered to adopt delegated acts to update the list of prohibited practices where new practices emerge that pose a similar threat to fundamental rights and European values as posed by the practices explicitly listed at this point.

Restrictions on biometric techniques

Already at first sight, the provisions on 'real-time' remote biometric identification in Article 5 (1) (d) and (2) to (4) seem to be an alien element within the wider framework of prohibited AI practices. This is because those practices are not per se prohibited as obviously incompatible with fundamental rights and European values, but rather restricted and only permitted when particular substantive and/or procedural requirements are met. It would therefore be preferable for provisions on biometric techniques to be moved to a separate new Title IIa on 'Restricted Artificial Intelligence Practices'.

For very similar reasons as have already been given in the context of manipulation by subliminal techniques and exploitation of vulnerabilities, it would be important to clarify better the relationship between the AIA and Article 9 of the General Data Protection Regulation (GDPR) as well as Article 10 of the Law Enforcement Directive (LED). It is less than ideal to have a central EU legal instrument, that is supposed to be understood and applied by lawyers and non-lawyers across Europe and to serve as a role model worldwide, if the interplay with EU data protection law and the considerations behind the concrete way in which the restrictions have been phrased is clear only to very few experts who are familiar with the entire acquis.

More importantly, the fact that the provisions in the AIA on biometric techniques rely on the definition of 'biometric data' provided by the GDPR is very problematic, notably in the consumer context. As this definition requires that the relevant data must allow or confirm the unique identification of a particular natural person it fails to capture many 'second-generation biometrics', such as voice, keystroke or gait patterns, as well as 'soft

biometrics', such as facial expressions, movements or body shape. However, it is these data that are the basis of many biometric techniques applied in the consumer context, such as biometric categorisation and emotion recognition. It is therefore of the essence to change some of the definitions, including by introducing a new category of 'biometrics-based data'. Given that the GDPR as it currently stands does not provide for sufficient safeguards with regard to the processing of such 'biometrics-based data' it is furthermore of the essence to introduce additional restrictions or biometric techniques other than real-time remote biometric identification, including for emotion recognition systems and biometric categorisation systems.

It is also recommended that the definitions of 'real-time' and 'remote' are modified in order to express more clearly what is probably intended. Last but not least, there should be an explicit provision on decisions based on biometric techniques, inspired by Article 22 GDPR and integrating an improved version of the somewhat misguided provision in Article 14 (5) AIA Proposal. This should, in particular, restrict the use of emotion recognition or biometric categorisation systems as legal evidence.

The list of high-risk AI systems

The classification of AI systems as 'high-risk' AI systems may rely on the fact that they serve as a safety component of other products, or are themselves products, covered by New Legislative Framework (NLF) product safety legislation and are required to undergo third-party conformity assessment under that legislation. As far as this is the case results are mostly convincing, but there are also cases where the fact that triggers a requirement of third-party conformity assessment under NLF product safety legislation has little to do with the specific risks posed by AI. For instance, a very small vacuum cleaner robot would be considered a high-risk AI system, but not so a computer game or chat bot intended for children and potentially influencing a child's personal development to a significant extent.

As far as the list of high-risk AI systems in Annex III is concerned a strict limitation to the areas listed in Points 1 to 8 in Annex III should be reconsidered. In particular, a new area should be added that addresses AI systems intended to be used by children and similar vulnerable groups as well as AI systems to be used in situations that create specific vulnerabilities, such as virtual assistants used by consumers for taking important decisions.

From a consumer perspective, Point.5 of Annex III as it currently stands seems to be clearly insufficient. While the inclusion of credit scoring is certainly positive, there seems to be no convincing reason for not also including individual risk assessment of natural persons in the context of access to essential private and public services, in particular insurance. Likewise, the evaluation of aspects such as complaint history or the likelihood that a consumer will exercise statutory or contractual rights should also be covered where it is used to influence future access to private or public services (including sale of any products). Last but not least, AI systems used for personalised pricing within the meaning of Article 6 (1) (ea) Consumer Rights Directive (CRD) should also be included in the list of high-risk AI systems. The exception for small scale providers putting the AI system into service exclusively for their own purposes should apply.

Beyond the consumer context, some of the other Points in Annex III should also be reconsidered or slightly reformulated, including Point 1 on biometric techniques, Point 2 on management and operation of critical infrastructure, and Point 4 on employment, workers management and access to self-employment.

Individual rights

As the AIA Proposal currently stands, individual rights (such as with regard to automated decision making) are entirely left to the GDPR. This is suboptimal for various reasons. Article 22 GDPR and the respective information duties in Articles 13 to 15 GDPR are restricted to fully automated decisions and fail to capture AI systems that recommend decisions to humans, i.e. where there is a meaningful degree of human intervention. Definitely, the GDPR provisions do not capture situation where the data that are being processed relate to a legal person, such as a microenterprise. Also, subject to clarification being provided by the Court of Justice, the information duties might not include a genuine right to receive an explanation for decisions taken. There is little prospect that this will change in the near future because there does not seem to be much appetite at political level for touching the GDPR. In addition, these individual rights are anyway an alien element in the GDPR because the problem is not so much the processing of input data relating specifically to the affected person as a data subject. Rather, the problem lies in the output data, which may have been generated with the help of (training) data relating to very different data subjects, or with the help of non-personal data. This is why it is suggested to give Title IV a new focus on individual rights where AI systems presenting either a transparency or a fairness risk are deployed.

The existing Article 52 on transparency obligations could – with the necessary editorial adjustments – largely remain as it is. However, a new provision on transparency obligations with regard to social bots should be included as there seems to be no plausible reason to have such an obligation for AI systems that interact with natural persons (such as chat bots) and for deep fakes, but not bots that (merely) generate content (which are, however, mentioned in Recital 70).

More importantly, it is recommended to introduce additional provisions in Title IV on an individual right to independent scrutiny of individual decision-making and to an explanation of individual decision-making. Major benefits for affected persons would include that these individual rights do not only apply for fully automated decisions, but also to decisions recommended to humans, and that the right to receive an explanation would be much more explicit and include, in particular, the main parameters of decision-making and their relative weight as well as an easily understandable explanation of inferences drawn if the inference itself is a main parameter. Also, the details, including with regard to appropriate exceptions, would provide much more legal certainty to both affected persons and users of AI systems.

Liability

Liability for AI systems should not primarily be dealt with in the AIA itself, but be largely a matter for product liability law, national tort law and/or a new EU regime of AI liability. However, as these liability regimes are focussed on traditional safety risks (e.g. personal injury, property damage) and are not well suited to address harm caused by fundamental rights risks (e.g. discrimination, manipulation, exploitation), it is advisable to insert two liability provisions in the AIA itself, one on vicarious liability and one liability for lack of 'fundamental rights safety'. The former would help overcome existing uncertainties with regard to vicarious liability under national law (such as §§ 1313a, 1315 ABGB or §§ 278, 831 BGB), the latter with regard to doctrines such as Schutzgesetzverletzung (cf. § 1311 ABGB or § 823 (2) BGB).

Enforcement

From a consumer policy perspective, it seems important that the AIA, or selected relevant provisions thereof, is included in the list of legal instruments in Annex I to the Representative Actions Directive (RAD).

In addition, it is recommended to include a new enforcement mechanism with regard to systemic risks that arise where a high-risk AI system, while complying with the AIA as such, has the potential of significantly changing our societies and economies. This may be the case where an AI system, together with other systems building on that AI system, exceeds a defined threshold of market coverage, causing characteristic features and smaller deficiencies (that may be acceptable in an AI system when seen in isolation) turn into a systemic risk. For example, bias in a system that is dominant on the relevant market could cause new disadvantaged groups to emerge that are no longer captured by non-discrimination law as it currently exists, or a system could have significant effects on human skills and competences, or on the behaviour of affected groups and the way our societies and economies work. The new enforcement mechanism suggested has been inspired by Articles 25 ff of the proposed Digital Services Act (DSA), and it includes data access for vetted researchers.

In addition to a new enforcement mechanism for systemic risks it is suggested to insert a provision that avoids threats to public and national security interests which could result if national authorities in all 27 Member States had full access to all relevant data and the source code of, e.g., AI systems that are safety components in critical infrastructure (such as AI systems used to detect attacks on power grids within the Union).

Zusammenfassung

Der Vorschlag für ein Gesetz über künstliche Intelligenz (Artificial Intelligence Act, AIA) vom 21. April 2021, COM(2021) 206 final, stellt einen Meilenstein in der Regulierung von KI weltweit dar. Angesichts der massiven Risiken für Verbraucher:innen (neben vielen Chancen), die mit der großflächigen Einführung von KI-Systemen einhergehen, muss sichergestellt werden, dass dieses Instrument (in Kombination mit anderen Instrumenten) ein angemessenes Schutzniveau für Verbraucher:innen gewährleistet.

Auf den ersten Blick scheint der AIA-Vorschlag Verbraucherinteressen als solche weitgehend auszublenden. Dieser Eindruck entsteht z.B. dadurch, dass das Verbot der Manipulation durch unterschwellige Techniken und der Ausnutzung von Vulnerabilitäten auf Fälle beschränkt ist, in denen solche KI-Praktiken physischen oder psychischen Schaden verursachen oder wahrscheinlich verursachen werden (d.h. rein wirtschaftlicher Schaden ist nicht erfasst). Denselben Eindruck erweckt die derzeitige Formulierung der Liste in Anhang III über KI-Systeme mit hohem Risiko, da diese Liste zwar KI zur Prüfung der Kreditwürdigkeit umfasst, nicht aber KI zur Bewertung von Faktoren, die der Kreditwürdigkeit ähnlich sind (wie z. B. die Beschwerdehistorie), KI zur Risikobewertung betreffend natürlicher Personen durch Versicherungsunternehmen oder ganz allgemein die personalisierte Preisgestaltung.

Die Liste der verbotenen KI-Praktiken

Die Verfasser:innen des AIA-Vorschlags waren offensichtlich bestrebt, Lücken in den bestehenden Rechtsvorschriften zu schließen, gleichzeitig aber jede Art von Überschneidung mit bestehenden Rechtsvorschriften zu vermeiden. Dass etwa Artikel 5 nicht die Manipulation durch unterschwellige Techniken oder die Ausnutzung gruppenspezifischer Vulnerabilitäten erfasst, wenn diese einen reinen Vermögensschaden verursachen, lässt sich durch den Wunsch der Verfasser:innen erklären, nicht in die Nähe der Richtlinie über unlautere Geschäftspraktiken (UGP-RL) zu gelangen.

Im Fall von Artikel 5 des AIA-Vorschlags wird diese Strategie der Vermeidung von Überschneidungen allerdings durch zahlreiche Lücken erkauft, z. B. in Bezug auf die Manipulation oder Übervorteilung von natürlichen Personen, die als KKMU handeln und

nicht unter den Verbraucherschutz fallen, in Bezug auf die Manipulation oder Ausnutzung von Vulnerabilitäten bei anderen als transaktionsbezogenen Entscheidungen oder in Bezug auf das Inverkehrbringen der betreffenden KI-Systeme als solches.

Das Bestreben, Überschneidungen mit bestehenden Rechtsvorschriften zu vermeiden, geschieht auch um den Preis einer Regelung, die zumindest auf den ersten Blick in ihren rechtspolitischen Entscheidungen recht willkürlich erscheint (z. B. indem sie Verbraucherinteressen scheinbar übersieht). Da Artikel 5 nicht richtig verstanden werden kann, ohne ihn im größeren Rahmen des bestehenden Datenschutz-, Antidiskriminierungs-, Verbraucherschutz- und Wettbewerbsrechts zu analysieren, besteht die Gefahr, dass der AIA von den betroffenen Verkehrskreisen in ganz Europa nicht richtig verstanden und angewendet wird. Außerdem darf nicht vergessen werden, dass der AIA das Potenzial hat, ein globales Vorbild für die Regulierung von KI-Anwendungen zu werden. Um dieser Rolle gerecht zu werden, muss er leicht verständlich sein und die zugrundeliegenden rechtspolitischen Entscheidungen und Annahmen in kohärenter Weise widerspiegeln. Ein Rechtsakt, der weltweit nur von sehr wenigen Expert:innen verstanden wird, weil man zu seinem Verständnis eine sehr profunde Kenntnis des übrigen Acquis und des Anwendungsbereichs verschiedener anderer Rechtsinstrumente haben muss, wird nicht ohne weiteres zu einem Rechtsinstrument werden, von dem sich andere Regionen der Welt inspirieren lassen. Die Trennung von Anwendungsbereich und Regelungsgehalt verschiedener Rechtsinstrumente ist dort sinnvoll, wo Überschneidungen die Gefahr von Unstimmigkeiten und/oder unnötigen Kumulationseffekten mit sich bringen würden. Innerhalb einer schwarzen Liste verbotener Praktiken ist es jedoch nicht notwendig, Überschneidungen zu vermeiden. Vielmehr ist es durchaus hinnehmbar (und in der Tat ein Zeichen für die Kohärenz und Widerspruchsfreiheit des Acquis), dass schädliche und manipulative Praktiken nicht nur durch ein, sondern gleich durch zwei oder mehrere Rechtsinstrumente auf EU-Ebene verboten werden.

Es wird daher empfohlen, die Beschränkung auf ‚physischen oder psychischen' Schaden in Artikel 5 Absatz 1 Buchstaben a und b aufzuheben und etwa durch ‚signifikanten und ungerechtfertigten' Schaden zu ersetzen. Dies würde sowohl sicherstellen, dass die wirtschaftlichen Interessen der Verbraucher:innen angemessen berücksichtigt werden, als auch eine überschießende Anwendung des Verbots über das intendierte Ziel hinaus vermeiden. Es ist ebenfalls wichtig, das Verbot der Ausnutzung gruppenspezifischer Vulnerabilitäten auf Vulnerabilitäten auszuweiten, die sich aus der wirtschaftlichen und sozialen Situation natürlicher Personen ergeben. Noch einen Schritt weiter gehend wird empfohlen, die Bestimmung von gruppenspezifischen Vulnerabilitäten auf individuelle

Vulnerabilitäten auszuweiten. Im Verbraucherkontext wird die Ausnutzung sehr individueller Schwächen (z. B. starke Neigungen oder Abhängigkeiten), die durch umfangreiche Datenerfassung und Profiling offengelegt werden, zu einem immer größeren Problem. Ebenso wird empfohlen, das Verbot des Social Scoring auf Social-Scoring-Aktivitäten auszuweiten, die von privaten Parteien durchgeführt werden, z. B. wenn eine Gatekeeper-Plattform Verbraucher:innen auf der Grundlage ihres Sozialverhaltens in einer Reihe von verschiedenen Lebensbereichen bewertet.

Über den Verbraucherkontext hinaus wäre es ratsam, die umfassende oder totale Überwachung im Privatleben oder am Arbeitsplatz in die Liste der verbotenen Praktiken aufzunehmen. Gleiches gilt für die gezielte technische Verarbeitung von Gehirndaten durch Brain-Computer-Interfaces (BCI), um die Gedanken einer Person gegen deren Willen in einer Weise zu lesen oder zu manipulieren, die dieser Person einen signifikanten und ungerechtfertigten Schaden zufügt oder zufügen kann (vgl. das Recht auf psychische Privatsphäre und Integrität).

In jedem Fall ist es von wesentlicher Bedeutung, dass auf Verbote verwiesen wird, die sich aus anderen Gesetzen ergeben, einschließlich des Datenschutzrechts, des Antidiskriminierungsrechts, des Verbraucherschutzrechts und des Wettbewerbsrechts. Außerdem muss die Kommission die Befugnis erhalten, delegierte Rechtsakte zu erlassen, um die Liste der verbotenen Praktiken zu aktualisieren, wenn neue Praktiken auftreten, die eine ähnliche Bedrohung für die Grundrechte und die europäischen Werte darstellen wie die an dieser Stelle ausdrücklich aufgeführten Praktiken.

Einschränkungen für biometrische Verfahren

Schon auf den ersten Blick scheinen die Bestimmungen über die biometrische Fernidentifizierung ‚in Echtzeit' in Artikel 5 Absatz 1 Buchstabe d) und Absätze 2 bis 4 einen Fremdkörper im größeren Rahmen der verbotenen KI-Praktiken darzustellen. Dies liegt daran, dass diese Praktiken nicht per se als mit den Grundrechten und europäischen Werten offensichtlich unvereinbar verboten sind, sondern vielmehr eingeschränkt und nur dann zulässig sind, wenn bestimmte materiellrechtliche und/oder verfahrensrechtliche Anforderungen erfüllt sind. Es wäre daher vorzuziehen, die Bestimmungen über biometrische Verfahren in einen separaten neuen Titel IIa über „Eingeschränkt zulässige Praktiken der künstlichen Intelligenz" aufzunehmen.

Aus ganz ähnlichen Gründen, wie sie bereits im Zusammenhang mit der Manipulation durch unterschwellige Techniken und der Ausnutzung von Vulnerabilitäten genannt wurden, wäre es wichtig, das Verhältnis zwischen dem AIA und Artikel 9 der Datenschutz-Grundverordnung (DSGVO) sowie Artikel 10 der Strafverfolgungsrichtlinie (LED) besser zu verdeutlichen. Für ein zentrales EU-Rechtsinstrument, das von Jurist:innen und Nichtjurist:innen in ganz Europa verstanden und angewandt werden und weltweit als Vorbild dienen soll, ist es suboptimal, wenn das Zusammenspiel mit dem EU-Datenschutzrecht und die Überlegungen, die hinter der konkreten Formulierung der Beschränkungen stehen, nur wenigen Expert:innen, die mit dem gesamten Acquis vertraut sind, erkennbar sind.

Noch wichtiger ist die Tatsache, dass die Bestimmungen des AIA über biometrische Verfahren auf der Definition von „biometrischen Daten“ der DSGVO beruhen, was vor allem im Verbraucherkontext sehr problematisch ist. Da diese Definition voraussetzt, dass die relevanten Daten die eindeutige Identifizierung einer bestimmten natürlichen Person ermöglichen oder bestätigen, werden viele ‚Second Generation Biometrics ‘, wie Stimme, Tastenanschlag oder Gangmuster, sowie ‚Soft Biometrics“, wie Gesichtsausdruck, Bewegungen oder Körperform, nicht erfasst. Diese Daten sind jedoch die Grundlage vieler biometrischer Verfahren, die im Verbraucherkontext angewandt werden, wie etwa die biometrische Kategorisierung und die Emotionserkennung. Daher ist es von entscheidender Bedeutung, einige der Definitionen zu ändern, unter anderem durch die Einführung einer neuen Kategorie ‚biometriegestützter Daten‘. Da die Datenschutz-Grundverordnung in ihrer derzeitigen Fassung keine ausreichenden Garantien für die Verarbeitung solcher „biometriegestützten Daten“ vorsieht, ist es darüber hinaus von entscheidender Bedeutung, zusätzliche Einschränkungen für andere biometrische Techniken als die biometrische Fernidentifizierung in Echtzeit einzuführen, insbesondere auch für Systeme zur Emotionserkennung und biometrischen Kategorisierung.

Es wird auch empfohlen, die Definitionen von ‚in Echtzeit‘ und ‚aus der Ferne‘ zu ändern, um deutlicher zum Ausdruck zu bringen, was vermutlich beabsichtigt war. Nicht zuletzt sollte es eine ausdrückliche Bestimmung über Entscheidungen auf der Grundlage biometrischer Verfahren geben, die sich an Artikel 22 der Datenschutz-Grundverordnung orientiert und die etwas misslungene Bestimmung in Artikel 14 Absatz 5 des AIA-Vorschlags in verbesserter Form integriert. Diese Bestimmung sollte insbesondere auch die Verwendung von Systemen zur biometrischen Kategorisierung oder Emotionserkennung als Beweismittel erfassen.

Die Liste der Hochrisiko-KI-Systeme

Die Einstufung von KI-Systemen als ‚Hochrisiko'-KI-Systeme kann darauf beruhen, dass sie als Sicherheitskomponente anderer Produkte dienen oder selbst Produkte sind, die unter die Produktsicherheitsvorschriften des New Legislative Framework (NLF) fallen und gemäß diesen Vorschriften einer Konformitätsbewertung durch Dritte unterzogen werden müssen. Soweit dies der Fall ist, sind die Ergebnisse meist überzeugend, aber es gibt auch Fälle, in denen die Tatsache, die eine Konformitätsbewertung durch Dritte gemäß den NLF-Rechtsvorschriften zur Produktsicherheit auslöst, wenig mit den spezifischen Risiken zu tun hat, die von KI ausgehen. So würde beispielsweise ein sehr kleiner Staubsaugerroboter als Hochrisiko-KI-System gelten, nicht aber ein Computerspiel oder ein Chatbot, der für Kinder bestimmt ist und möglicherweise die persönliche Entwicklung eines Kindes in erheblichem Maße beeinflusst.

Was die Liste der Hochrisiko-KI-Systeme in Anhang III betrifft, so sollte eine strikte Beschränkung auf die in Anhang III unter den Nummern 1 bis 8 aufgeführten Bereiche überdacht werden. Insbesondere sollte ein neuer Bereich hinzugefügt werden, der sich mit KI-Systemen befasst, die für die Nutzung durch Kinder und ähnlich schutzbedürftige Gruppen bestimmt sind, sowie mit KI-Systemen, die in Situationen eingesetzt werden, die eine besondere Gefährdungslage hervorrufen, wie z. B. virtuelle Assistenten, die von Verbraucher:innen für das Treffen wichtiger Entscheidungen genutzt werden.

Aus der Sicht von Verbraucher:innen scheint Anhang III Nummer 5 in seiner derzeitigen Fassung eindeutig unzureichend zu sein. Während die Einbeziehung von KI zur Kreditwürdigkeitsprüfung sicherlich positiv ist, scheint es keinen überzeugenden Grund dafür zu geben, nicht auch die individuelle Risikobewertung natürlicher Personen im Zusammenhang mit dem Zugang zu wichtigen privaten und öffentlichen Dienstleistungen, insbesondere Versicherungen, mit einzubeziehen. Ebenso sollte die Bewertung von Aspekten wie der Beschwerdehistorie oder der Neigung zur Ausübung von Verbraucherrechten ebenfalls erfasst werden, wenn sie dazu dient, den künftigen Zugang zu privaten oder öffentlichen Dienstleistungen (einschließlich des Verkaufs von Produkten) zu beeinflussen. Nicht zuletzt sollten auch KI-Systeme, die für die personalisierte Preisgestaltung im Sinne von Artikel 6 Absatz 1 Buchstabe ea Verbraucherrechte-RL verwendet werden, in die Liste der Hochrisiko-KI-Systeme aufgenommen werden. Die Ausnahmeregelung für kleine Anbieter, die das KI-System ausschließlich für eigene Zwecke entwickeln und in Betrieb nehmen, sollte weiterhin gelten.

Über den Verbraucherkontext hinaus sollten auch einige der anderen Punkte in Anhang III überdacht oder leicht umformuliert werden, darunter Punkt 1 über biometrische Verfahren, Punkt 2 über die Verwaltung und den Betrieb kritischer Infrastrukturen und Punkt 4 über Beschäftigung, Personalmanagement und Zugang zur Selbständigkeit.

Individualrechte

In der derzeitigen Fassung des AIA-Vorschlags werden Individualrechte (z. B. in Bezug auf die automatisierte Entscheidungsfindung) vollständig der DSGVO überlassen. Dies ist aus verschiedenen Gründen suboptimal. Artikel 22 DSGVO und die entsprechenden Informationspflichten in den Artikeln 13 bis 15 DSGVO beschränken sich auf vollautomatische Entscheidungen und erfassen keine KI-Systeme, die menschlichen Akteuren Entscheidungen empfehlen, d.h. bei denen ein relevantes Maß an menschlichem Einfluss auf die Entscheidungsfindung gegeben ist. Die DSGVO erfasst definitiv nicht Situationen, in denen sich die verarbeiteten Daten auf eine juristische Person, wie z. B. ein Kleinstunternehmen, beziehen. Vorbehaltlich einer Klärung durch den Gerichtshof dürften die Informationspflichten der DSGVO auch kein echtes Recht auf eine Erklärung getroffener Entscheidungen beinhalten. Es besteht wenig Aussicht, dass sich dies in naher Zukunft ändern wird, da auf politischer Ebene kein großes Interesse daran zu bestehen scheint, die DSGVO anzutasten. Außerdem sind diese individuellen Rechte ohnehin ein Fremdkörper in der DSGVO, weil das Problem nicht so sehr in der Verarbeitung von Eingabedaten liegt, die sich speziell auf die betroffene Person beziehen. Das Problem liegt vielmehr in den Ausgabedaten, die möglicherweise mit Hilfe von (Trainings-)Daten, die sich auf ganz andere Personen beziehen, oder mit Hilfe von nicht-personenbezogenen Daten erzeugt worden sind. Deshalb wird vorgeschlagen, in Titel IV einen neuen Schwerpunkt auf Individualrechte im Zusammenhang mit KI-Systemen zu legen, die entweder ein Transparenz- oder ein Fairness-Risiko darstellen.

Der bestehende Artikel 52 zu den Transparenzpflichten könnte – mit den notwendigen redaktionellen Anpassungen an anderweitig erfolgende Umformulierungen – weitgehend unverändert bleiben. Allerdings sollte eine neue Bestimmung über Transparenzpflichten in Bezug auf Social Bots aufgenommen werden, da es keinen plausiblen Grund zu geben scheint, eine solche Verpflichtung für KI-Systeme zu haben, die mit natürlichen Personen interagieren (wie Chatbots) und für Deep Fakes, nicht aber für Bots, die (lediglich) Inhalte generieren (und die auch in Erwägungsgrund 70 erwähnt werden).

Vor allem aber wird empfohlen, in Titel IV zusätzliche Bestimmungen über das Recht des Einzelnen auf eine unabhängige Prüfung individueller Entscheidungen und auf eine Erklärung individueller Entscheidungen aufzunehmen. Zu den wesentlichen Vorteilen für die Betroffenen würde gehören, dass diese individuellen Rechte nicht nur für vollautomatisierte Entscheidungen gelten, sondern auch für Entscheidungen, die menschlichen Akteuren empfohlen werden, und dass das Recht auf Erklärung sehr viel expliziter wäre und insbesondere die Hauptparameter der Entscheidungsfindung und ihr relatives Gewicht sowie eine leicht verständliche Erläuterung von Ableitungen umfassen würde, wenn die Ableitung selbst ein Hauptparameter ist. Die Einzelheiten, einschließlich sachgerechter Ausnahmen, würden sowohl den Betroffenen als auch den Nutzern von KI-Systemen viel mehr Rechtssicherheit bieten.

Haftung

Die Haftung für KI-Systeme sollte nicht in erster Linie im AIA selbst geregelt werden, sondern weitgehend dem Produkthaftungsrecht, dem nationalen Deliktsrecht und/oder einer neuen EU-Regelung über KI-Haftung vorbehalten sein. Da sich diese Haftungsregelungen jedoch auf herkömmliche Sicherheitsrisiken (z.B. Personen- und Sachschäden) konzentrieren und nicht gut geeignet sind, um durch Grundrechtsrisiken (z.B. Diskriminierung, Manipulation, Ausbeutung) verursachte Schäden anzugehen, ist es ratsam, zwei Haftungsbestimmungen in den AIA selbst aufzunehmen, und zwar eine über die Erfüllungsgehilfenhaftung und eine über die Haftung für fehlende „Grundrechts-Sicherheit“ von KI-Systemen Erstere würde dazu beitragen, bestehende Unsicherheiten im Hinblick auf die Gehilfenhaftung nach nationalem Recht (etwa §§ 1313a, 1315 ABGB oder §§ 278, 831 BGB) zu überwinden. Letztere würde sich an Artikel 82 DSGVO orientieren und europaweit klarstellen, dass bei Verletzung bestimmter Schutzgesetze innerhalb des AIA für die Verursachung materieller wie immaterieller Schäden gehaftet wird (vgl. § 1311 ABGB oder § 823 (2) BGB).

Rechtsdurchsetzung

Aus verbraucherpolitischer Sicht erscheint es wichtig, dass der AIA bzw. ausgewählte relevante Bestimmungen davon in die Liste der Rechtsinstrumente in Anhang I der neuen Verbandsklagen-Richtlinie (RAD) aufgenommen werden.

Darüber hinaus wird empfohlen, einen neuen Durchsetzungsmechanismus für systemische Risiken aufzunehmen, die entstehen, wenn ein Hoch-Risiko-System zwar den Anforderungen des AIA als solchem entspricht, aber dennoch das Potenzial hat, unsere Gesellschaften und Volkswirtschaften erheblich zu verändern. Dies kann der Fall sein, wenn ein KI-System zusammen mit anderen Systemen, die auf diesem KI-System aufbauen, einen bestimmten Schwellenwert der Marktabdeckung überschreitet, so dass charakteristische Merkmale und kleinere Mängel (die bei einem KI-System für sich betrachtet akzeptabel sein mögen) zu einem systemischen Risiko werden. So könnten beispielsweise Verzerrungen in einem System, das den betreffenden Markt beherrscht, dazu führen, dass neue benachteiligte Gruppen entstehen, die nicht mehr von den derzeitigen Antidiskriminierungsgesetzen erfasst werden. Auch könnte ein System erhebliche Auswirkungen auf den Erhalt menschlicher Fähigkeiten und Kompetenzen oder auf das Verhalten betroffener Gruppen oder die Funktionsweise unserer Gesellschaft und Wirtschaft entwickeln. Der vorgeschlagene neue Durchsetzungsmechanismus orientiert sich an den Artikeln 25 ff. des vorgeschlagenen Gesetzes über digitale Dienste (DSA) und beinhaltet einen Datenzugang für zugelassene Forscher.

Zusätzlich zu einem neuen Durchsetzungsmechanismus für systemische Risiken wird vorgeschlagen, eine Bestimmung einzufügen, die Gefahren für öffentliche und nationale Sicherheitsinteressen vermeidet, die entstehen könnten, wenn nationale Behörden in allen 27 Mitgliedstaaten vollen Zugang zu allen relevanten Daten und dem Quellcode von z.B. KI-Systemen hätten, die Sicherheitskomponenten in kritischen Infrastrukturen darstellen (wie KI-Systeme, die zur Erkennung von Angriffen auf Stromnetze in der Union eingesetzt werden).

1 Introduction: Why the AIA matters for consumers in Europe

The Proposal for an Artificial Intelligence Act (AIA Proposal),[1] published by the European Commission on 21 April 2021, is another landmark step in the Union's endeavour to make European law fit for the digital age and at the same time to take a leading role globally on innovative regulatory models. Other legislative initiatives for the digital age that are currently in the pipeline include the Digital Services Act,[2] the Digital Markets Act,[3] the Data Governance Act[4] and the Data Act[5] as well as a potential revision of the GDPR[6] and the replacement of the E-Privacy Directive by an E-Privacy Regulation[7]. In addition, a range of legal instruments without a digital focus are also being adapted to the digital age, including to AI, for instance by replacing the existing Machinery Directive by a modernised Machinery Regulation (MR)[8] and the General Product Safety Directive (GPSD) by a modernised General Product Safety Regulation (GPSR)[9].

[1] Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, COM(2021) 206 final.
[2] Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC, COM(2020) 825 final.
[3] 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.
[4] Proposal for a Regulation of the European Parliament and of the Council on European data governance (Data Governance Act), COM(2020) 767 final.
[5] See schedule at https://www.europarl.europa.eu/legislative-train/theme-a-europe-fit-for-the-digital-age/file-data-act.
[6] See, e.g. European Parliament resolution of 25 March 2021 on the Commission evaluation report on the implementation of the General Data Protection Regulation two years after its application (2020/2717(RSP)).
[7] See mandate for negotiations with EP of 10 February 2021, Proposal for a Regulation of the European Parliament and of the Council concerning the respect for private life and the protection of personal data in electronic communications and repealing Directive 2002/58/EC (Regulation on Privacy and Electronic Communications), Council Document no. 6087/21.
[8] Proposal for a Regulation of the European Parliament and of the Council on machinery products, COM(2021) 202 final.
[9] Proposal for a Regulation of the European Parliament and of the Council on general product safety, amending Regulation (EU) No 1025/2012 of the European Parliament and of the Council, and repealing Council Directive 87/357/EEC and Directive 2001/95/EC of the European Parliament and of the Council, COM(2021) 346 final.

The AIA Proposal divides AI systems into different risk levels. Particular AI systems posing an unacceptable risk are prohibited altogether under Article 5. The main part of the AIA Proposal is devoted to 'high-risk' AI systems. This is where the full set of mandatory requirements listed under Title III, such as concerning data governance, transparency or human oversight, and a requirement of ex ante conformity assessment apply. The AI systems that qualify as high-risk systems are partly items covered by other product safety legislation, and partly listed in Annex III. Some few AI systems (such as bots or deep fakes) give rise to specific information duties.

It is beyond doubt that the AIA as a central piece of product safety legislation is of utmost significance for consumers in Europe. This is why it is all the more surprising that the AIA Proposal seems to turn a blind eye on consumer interests as such, at least at first sight.[10] This impression is created, e.g., by the fact that the prohibition of manipulation by subliminal techniques and of exploitation of vulnerabilities is restricted to cases where such AI practices cause, or are likely to cause, physical or psychological harm (excluding mere economic harm). The same impression is created by the way the list in Annex III on high-risk AI systems is currently phrased. It is therefore of utmost importance to analyse the AIA Proposal from the perspective of consumer policy and to make sure, during the coming months or maybe even years of negotiations at EU level, that the consumer perspective is fully considered.

Consumers may be affected by the AIA where they buy consumer goods with embedded AI or goods that work in combination with AI, or where the AI as such qualifies as an item that is intended for consumers or can, under reasonably foreseeable conditions, be used by consumers.[11]

Illustration 1

A lawnmower robot used by consumers (whose AI components, inter alia, make sure the robot stops when touching a person and thus qualify as safety components) would qualify as a 'high-risk' AI system (see below at 7.1.2.1). The third-party conformity assessment under the MR would therefore have to cover

[10] BEUC, EU proposal for artificial intelligence law is weak on consumer protection, press release, 21 April 2021.
[11] See Article 1 point 1 of the proposed GPSR.

the requirements under the AIA, such as an appropriate level of accuracy and robustness (Article 15 AIA), making sure consumers using lawnmower robots do not get injured.

Illustration 2

An autonomous shopping assistant used by consumers (assisting consumers in identifying the best offers and in placing orders) would not qualify as a high-risk AI system (see below at 7.3.1) although the risks for consumer protection are high (e.g. the risk that consumers are systematically pushed into expensive deals). This is so for want of a legal regime on software safety[12] that would require third-party assessment for such a software or for such software being listed in Annex III of the AIA Proposal.

Due to the fact that the majority of provisions in the AIA Proposal only apply to high-risk AI systems and that the majority of high-risk AI systems will be deployed by professional users, consumers are more often affected by AI systems because businesses use AI systems in their B2C relations.

Illustration 3

A credit scoring AI system used by the majority of banks in a particular region for deciding about whether, or on which terms, to grant credit to consumers qualifies as a high-risk AI system according to Article 6(2) with Annex III No. 5(b) AIA Proposal.

[12] For recommendations see *Wendehorst*/*Duller*, Safety and Liability Related Aspects of Software, European Commission (2021), p. 7, 89 f.

Illustration 4

An AI system used for risk assessment by insurance companies, assisting insurance companies in deciding about whether, or on which terms, to offer insurance to consumers does not qualify as a high-risk AI system because it is not mentioned in Annex III No. 5 AIA Proposal. Therefore, the AIA Proposal does not contain any requirements with regard to such AI systems, although they are used for making decisions with significant importance for an individual's life.

Illustration 5

Likewise, AI systems for personalised pricing do not qualify as high-risk AI systems under the AIA Proposal, irrespective of the scale at which these systems are being used. Thus, it is possible that particular consumers are put at a massive disadvantage, considerably lowering their standard of living.

The illustrations both for AI systems intended for consumers and for AI systems intended for professional users (or public authorities) but affecting consumers have demonstrated that AI systems creating a significant risk for consumers are captured by the AIA Proposal to a very different extent. While some 'high-consumer-risk' AI systems are certainly qualified as 'high-risk' AI systems within the meaning of the AIA Proposal (see the examples of the lawnmower robot or the credit scoring system), other 'high-consumer-risk' AI systems are not captured by the AIA Proposal at all (see the examples of the shopping assistant, the risk assessment system for insurance companies, and the personalised pricing system). It is therefore important to analyse the AIA Proposal, including its relationship with other laws, in order to find out whether there are any gaps or deficiencies from a consumer policy perspective.

2 General Regulatory Approach of the AIA Proposal

2.1 Product safety law approach

The general regulatory approach taken by the AIA Proposal is a safety law approach. This means that, in terms of addressees, basic concepts, structure, enforcement mechanisms etc. the AIA Proposal is largely modelled on traditional product safety law, in the modernised form we see, e.g., in the new Proposals for the MR[13] or the GPSR[14] and that relies on strong enforcement on the basis of the Market Surveillance Regulation.[15] Although the AIA Proposal also contains, in particular in its Articles 29 and 52, duties for the users of AI systems (who are usually businesses or public authorities), the vast majority of provisions in the AIA are addressed at those who design and develop AI and place it on the market, be it as self-standing AI or as AI components in other products ('providers').

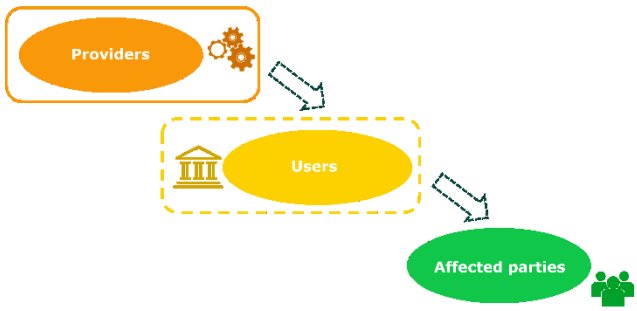

Figure 1: The AIA as product safety legislation

2.2 Risk-based approach

As has already been explained in the Introduction, the AIA Proposal takes a risk-based approach, dividing AI systems into four different risk levels: unacceptable risk (Title II), high risk (Title III), transparency risk (Title IV), and other AI systems which do not require specific legislation in the light of the minimal degree of risk they pose.

[13] See above (fn. 8).
[14] See above (fn. 9).
[15] Regulation (EU) 2019/1020 of the European Parliament and of the Council of 20 June 2019 on market surveillance and compliance of products and amending Directive 2004/42/EC and Regulations (EC) No 765/2008 and (EU) No 305/2011, OJ L 169, 25.6.2019, p. 1–44.

Particular AI systems posing an unacceptable risk are included in the list of 'prohibited AI practices' and dealt with under Article 5. The main part of the AIA Proposal is devoted to 'high-risk' AI systems. This is where the full set of mandatory requirements listed under Title III and the requirement of ex ante conformity assessment apply. The AI systems that qualify as high-risk systems are partly defined by Article 6(1) in conjunction with particular product safety legislation listed in Annex II, and partly in Annex III, which may be extended or otherwise modified according to criteria explained in Article 7. Some few AI systems are being mentioned by Article 52 as posing a particular 'transparency risk'. In essence, the user of bots, emotion recognition and biometric categorisation systems, or deep fakes must normally inform those exposed to it of the operation of the AI system.

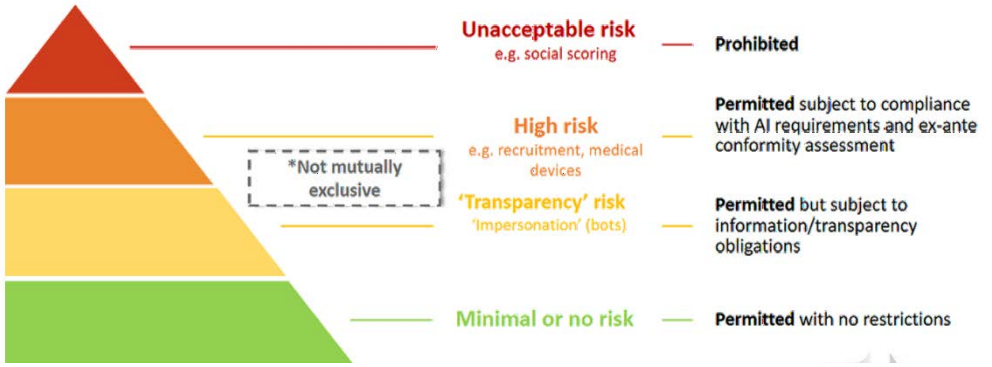

Figure 2: Risk-based approach of the AIA Proposal (Source and ©: European Commission)

It is important to stress that the risk levels are not mutually exclusive. This means, for instance, that a real-time remote biometric identification system dealt with under Title II (and not already prohibited by that Title) must, at the same time, fulfil all the requirements under Title III. In a similar vein, emotion recognition systems and biometric categorisation systems are normally only subject to Title IV, but where they qualify, in the light of their concrete purpose, as a high-risk system, they must also fulfil the requirements listed in Title III.

2.3 Safety risks and fundamental rights risks

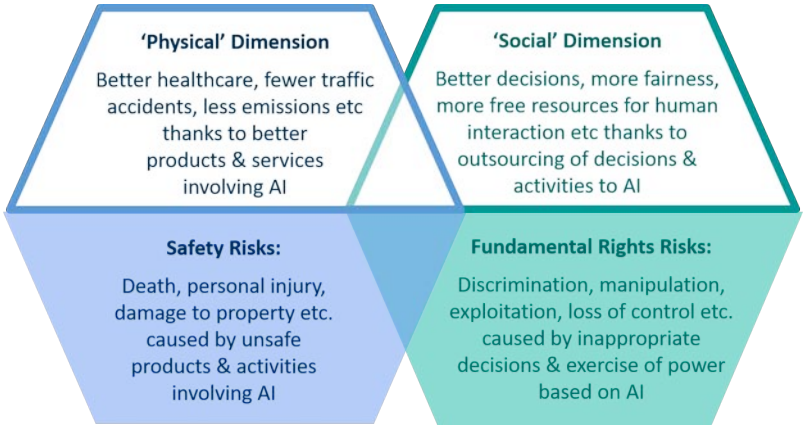

Figure 3: Safety risks and fundamental rights risks

The characteristic challenges posed by AI systems – such as opacity ('black box-effect'), complexity, and partially 'autonomous' and unpredictable behaviour – are similar irrespective of the sector in which, and the purpose for which, the AI system is deployed. However, the potential risks associated with AI systems usually appear to be falling into either of two dimensions: 'safety risks' and 'fundamental rights risks'.[16] These two types of risks are just the downside of our expectations of AI and of the promises made by those developing and deploying the technology, i.e. that AI will both help improving health, saving lives and the climate, and assist us in making better decisions, enhancing fairness and developing into a better society.

[16] In previous publications, I have referred to the two types as 'physical' and 'social' risks, see, e.g., *Schneider*/*Wendehorst*, Response to the Public Consultation on the White Paper: On Artificial Intelligence – A European Approach to Excellence and Trust COM(2020) 65 final (2020); *Wendehorst*/*Duller* (fn. 12) p. 26 ff.; *Wendehorst*, Strict Liability for AI and Other Emerging Technologies, Journal of European Tort Law (JETL) 2020, p. 150, 161 ff.

Safety risks and fundamental rights risks are not mutually exclusive, as any risk to life, health or property is at the same time the risk for certain fundamental rights, and as discrimination, manipulation or exploitation will often coincide with deprivation and on caused to health and overall well-being. This is also why the AI a does not strictly differentiate between the two types of risks. While one could say that particular provisions in the AIA put a particular focus on one of the two types of risks it is neither necessary nor helpful to draw a clear line between the two. One could even say that, by creating a new piece of product safety legislation that clearly includes all sorts of risks for the fundamental rights of individuals, the AIA has created a significantly extended notion of 'safety'.

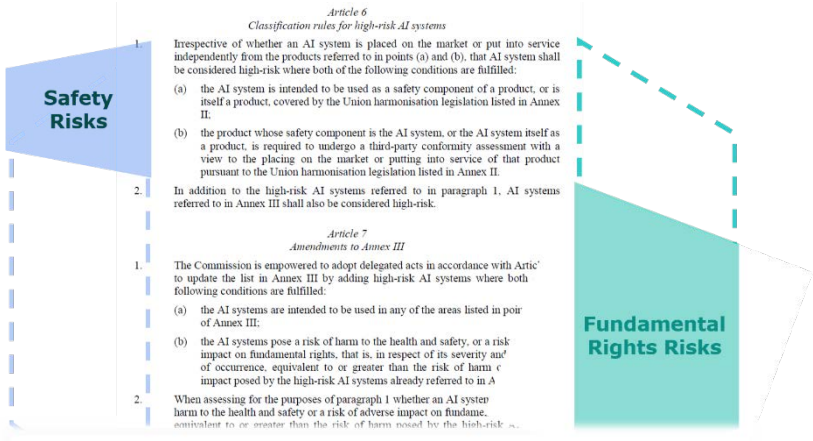

Figure 4: AI addressing both safety and fundamental rights risks

3 Relationship between the AIA and other Legal Instruments

The AIA will not be established on a vast and empty plain. Rather, it will have to be fit into a rather sophisticated system of existing laws many of which will not explicitly address AI systems, but will, without any doubt, capture a wide range of activities that make use of AI systems. According to the Explanatory Memorandum[17] the AIA Proposal is without prejudice to the rules laid down in such other laws, notably data protection law, consumer protection law, and non-discrimination law.

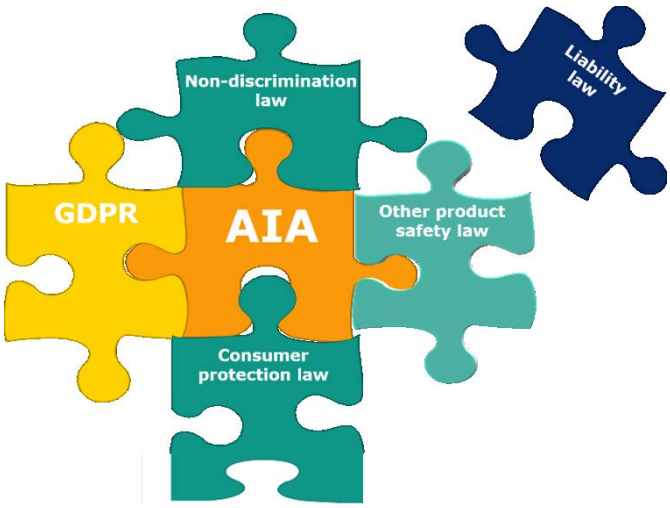

Figure 5: The jigsaw puzzle of legislation affecting AI practices

3.1 Product safety law

As has been explained further above (see 2.1) the general regulatory approach taken by the AIA Proposal is a product safety law approach.

Existing product safety law that follows the New Legislative Framework (NLF) approach[18] consists of the GPSD (in the future: the GPSR) as a horizontal horizontal fall-back regime, of a variety of strictly sectoral product safety regimes (such as regimes specifically focusing on medical devices, toys or personal protective equipment) and a number of regimes that are usually also referred to as 'sectoral' but that are really rather 'semi-horizontal' in nature because of their broad scope of application and of the fact that they

[17] COM(2021) 206 final, p. 4.
[18] Decision No 768/2008/EC of the European Parliament and of the Council of 9 July 2008 on a common framework for the marketing of products, and repealing Council Decision 93/465/EEC, OJ L 218, 13.8.2008, p. 82–128.

often apply in conjunction with the strictly sectoral regimes.[19] Examples for such 'semi-horizontal' regimes are the MR and the RED. The AIA Proposal is more 'horizontal' in nature than the MR or the RED because it covers also standalone software systems that are not (yet) covered by any other product safety regime and because it builds, in its Article 6 (1), on the existing framework of the MR and RED alongside the strictly sectoral frameworks.

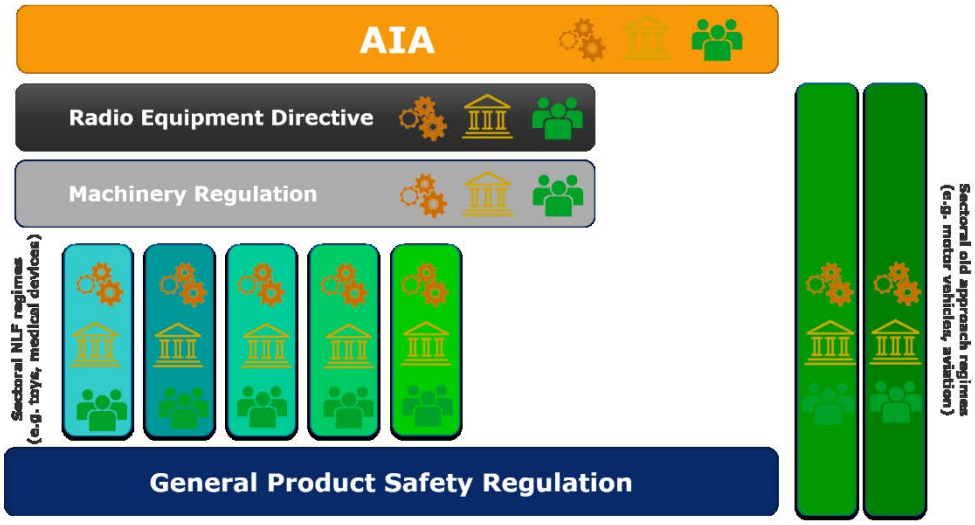

Figure 6: The AIA as a horizontal regime within the wider landscape of product safety legislation (simplified and incomplete)

The AIA both complements and builds on existing product safety legislation. It complements existing product safety legislation insofar as, where the products covered by such legislation use AI systems as safety components, the specific requirements set out by the AIA (e.g. in terms of data governance, record-keeping or human oversight) need to be met. The AIA builds on existing product safety legislation for the purpose of risk classification, i.e. apart from the AI systems that are classified as high-risk AI systems by the AIA itself under Articles 6 (2), 7 and Annex III, the AIA also 'adopts' risk classification made by other legal instruments. However, it must be stressed that the AIA, for 'adopted' risk classification, does not look at whether other product safety legislation listed in Annex II applies the label 'high-risk' to a particular product but rather at whether other product safety legislation subjects the relevant product to a third-party conformity assessment procedure.[20] So, for example, medical devices classified only as 'medium-risk' devices

[19] See Commission, The 'Blue Guide' on the implementation of EU product rules (2016).
[20] See Recital 31 AIA Proposal.

under the Medical Device Regulation (MDR)[21] still count as 'high-risk' AI systems under the AIA because, under the MDR, they require third-party conformity assessment.

3.2 Digital services law

An emerging area of the law that is not explicitly mentioned in the explanatory memorandum is digital services law, in particular the proposed Digital Services Act (DSA)[22] and Digital Markets Act (DMA)[23]. While the latter mainly addresses matters traditionally addressed by competition law (but introduces, in a by-the-way manner, a series of new duties, such as to allow for far reaching data portability), the former includes a number of provisions that address, at least indirectly, AI systems. This concerns, in particular, content moderation systems. While the definition of 'content moderation' given in Article 2 (p) DSA Proposal is not restricted to automated means, and even less to AI, it is clear that content moderation will in reality have to make ample use of AI systems if a platform is to cope with the quantity of content to deal with. Apart from content moderation, a number of further systems mentioned in the DSA will, in reality, be driven by AI, such as complaint-handling systems, search engines, or systems involved in online advertising. The legislative procedure on the DSA is meanwhile quite advanced, with a Council General Approach published on 18 November 2021.[24]

The DSA has its own risk-based approach, which mainly focusses on the size of a platform. The larger a platform the more obligations it has to fulfil and the closer is the scrutiny of its activities. Content moderation systems are not qualified as 'high-risk' AI systems in Annex III of the AIA Proposal as it currently stands, and there is no counterpart of Title III of the AIA Proposal in the DSA, but the DSA provides for an elaborate risk-management system and a range of procedural safeguards that exists in parallel to the AIA and may not require additional application of the AIA rules on 'high-risk' AI systems. However, there is

[21] Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC, OJ L 117, 5.5.2017, p. 1.
[22] Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC, COM(2020) 825 final.
[23] 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.
[24] Council Document ST_13203_2021_INIT

no general exception for digital services, so the provisions of Title II or Title IV of the AIA Proposal also apply to AI systems used by or on online platforms.

3.3 Consumer protection law

3.3.1 Areas of consumer protection law in the broader and narrower sense

The AIA Proposal is not explicitly qualified as consumer protection law, and its Title XII does not propose an explicit addition of the AIA to the list in Annex I of the Representative Actions Directive (RAD),[25] so that this Annex I addresses the AIA only indirectly via Articles 3 and 5 of the GPSD[26] and the future GPSR. However, even without such an explicit reference, it is beyond doubt that the AIA as a central piece of product safety legislation (see above at 2.1) plays an important role for consumer protection.

Within consumer protection law in the wider sense, there are areas and legal instruments that focus on risks for the life, health and property of consumers and, most recently with the AIA, risks for their fundamental rights. Broadly speaking, these areas and legal instruments can be described as addressing the safety of products and services on the one hand and liability for products and services on the other. By way

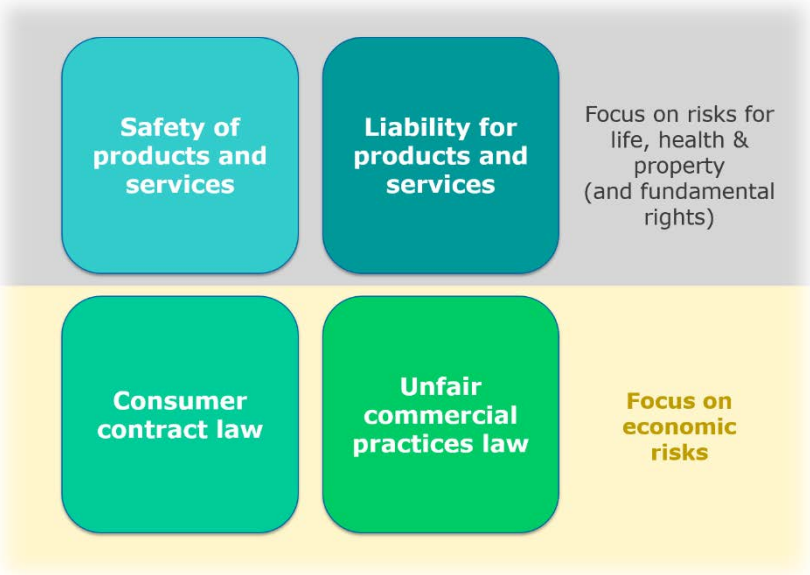

Figure 7: Different areas of (substantive) consumer protection law

[25] Directive (EU) 2020/1828 of the European Parliament and of the Council of 25 November 2020 on representative actions for the protection of the collective interests of consumers and repealing Directive 2009/22/EC, OJ L 409, 4.12.2020, p. 1–27.

[26] Articles 3 and 5 listed as No. 8 in Annex I.

of contrast, there are also areas and legal instruments that focus on economic risks for consumers, and these areas and legal instruments are often referred to as consumer protection law in the narrower sense. The bulk of this consumer protection law in the narrower sense consists of consumer contract law, such as the Consumer Rights Directive (CRD)[27] or the Unfair Contract Terms Directive (UCTD)[28] while, in particular, the Unfair Commercial Practices Directive (UCPD)[29] also addresses commercial relationships without a direct contractual bond between the relevant business and a consumer.

3.3.2 Points of contact between the AIA and consumer protection law

3.3.2.1 Identification of the main points of contact

While the relationship and the points of contact between the AIA Proposal and other product safety law as well as product liability law are rather straightforward (see above at 3.1), the relationship and points of contact between the AIA Proposal and consumer protection law in the narrower sense, such as the UCTD and the UCPD, is more complicated.

Given that consumer protection law in the narrower sense focuses on contractual or other commercial B2C relationships, it is more focused on concrete AI activities and practices that directly affect consumers and less so with the abstract safety of AI systems placed on the market. However, there are obvious overlaps. One of those obvious overlaps is Article 5 AIA Proposal, which focuses on certain AI practices rather than on certain AI systems.

---

[27] Directive 2011/83/EU of the European Parliament and of the Council of 25 October 2011 on consumer rights, amending Council Directive 93/13/EEC and Directive 1999/44/EC of the European Parliament and of the Council and repealing Council Directive 85/577/EEC and Directive 97/7/EC of the European Parliament and of the Council, OJ L 304, 22.11.2011, p. 64–88, as last amended by Directive (EU) 2019/2161 of the European Parliament and of the Council of 27 November 2019 amending Council Directive 93/13/EEC and Directives 98/6/EC, 2005/29/EC and 2011/83/EU of the European Parliament and of the Council as regards the better enforcement and modernisation of Union consumer protection rules, OJ L 328, 18.12.2019, p. 7–28.

[28] Council Directive 93/13/EEC of 5 April 1993 on unfair terms in consumer contracts, OJ L 95, 21.4.1993, p. 29–34, as last amended by Directive (EU) 2019/2161 (fn. 27)

[29] Directive 2005/29/EC of the European Parliament and of the Council of 11 May 2005 concerning unfair business-to-consumer commercial practices in the internal market and amending Council Directive 84/450/EEC, Directives 97/7/EC, 98/27/EC and 2002/65/EC of the European Parliament and of the Council and Regulation (EC) No 2006/2004 of the European Parliament and of the Council (Unfair Commercial Practices Directive) as last amended by Directive (EU) 2019/2161 (fn. 27).

The other obvious overlap is Article 29 AIA Proposal, which is about the obligations of (professional) users of high-risk AI systems.

3.3.2.2 Prohibited AI practices

As far as Article 5 AIA Proposal about prohibited AI practices is concerned, there is a striking absence of any reference to economic concerns. The prohibitions in Article 5 (1) (a) and (b) are explicitly restricted to harm of a physical or psychological nature, and remote biometric identification in Article 5 (1) (d) is by its very nature not about economic risks. The prohibition of social scoring in Article 5 (1) (c) includes a certain concern about economic risks (such as unfavourable treatment in the context of social benefits), but this is only one aspect of a much broader concern that is more related to human dignity and freedoms in general.

It is thus obvious that, by the very way in which the practices listed in Article 5 AIA Proposal have been phrased, the drafters tried to avoid as far as possible any kind of overlap between the AIA and, in particular, the UCPD, because they felt that any practice that would potentially qualify as manipulation by subliminal techniques or as exploitation of vulnerabilities and that causes or is likely to cause a person harm of an economic nature would automatically qualify as an unfair commercial practice and be prohibited under the UCPD.

The prohibition under Article 5 UCPD covers the application of any unfair commercial practices in dealing with consumers. 'Commercial practice' means any act, omission, course of conduct or representation, commercial communication including advertising and marketing, by a trader, directly connected with the promotion, sale or supply of a product to consumers.[30] Commercial practices are considered to be unfair if they are contrary to the requirements of professional diligence and materially distort or are likely to materially distort the economic behaviour of the average consumer whom it reaches or to whom it is addressed, or of the average member of the group when a commercial practice is directed to a particular group of consumers. Commercial practices which are likely to materially distort the economic behaviour only of a clearly identifiable group of consumers who are particularly vulnerable to the practice or the underlying product because of their mental or physical infirmity, age or credulity in a way which the trader

[30] Article 2(d) UCPD.

could reasonably be expected to foresee, are assessed from the perspective of the average member of that group. Specific provisions, which particularise the general prohibition under Article 5 (1) to (3) are in place with regard to commercial practices which are misleading or aggressive. Annex I to the UCPD contains a fully harmonised list of commercial practices which are in all circumstances to be regarded as unfair.

As will be explained in more detail below (5.1.1 and 5.1.2), the decision to avoid overlap with the UCPD is problematic and should be reconsidered.

3.3.2.3 Obligations of (professional) users of AI systems

The other obvious overlap between the AIA Proposal and consumer protection law in the narrower sense is Article 29 AIA. This proposed provision includes a number of obligations of the users of high-risk AI systems, notably the obligation:

- to use the high-risk AI systems in accordance with the instructions of use provided by the provider;
- to ensure, to the extent the user exercises control over the input data, that input data is relevant in view of the intended purpose of the high-risk AI system;
- to monitor the operation of the high-risk AI system on the basis of the instructions of use;
- to suspend the use of the system when they have reasons to consider that the use may result in the AI system presenting a risk at national level and to inform the provider or distributor accordingly;
- to interrupt the use of the AI system when they have reasons to consider that the AI system poses a risk of serious incidents or any malfunctioning which constitutes a breach of obligations under Union law intended to protect fundamental rights and to inform the provider or distributor accordingly;
- to keep the logs automatically generated by the high-risk AI system to the extent that such logs are under their control, for a period that is appropriate in the light of the intended purpose of the high-risk AI system and applicable legal obligations under Union or National law;
- where applicable, to use the information provided by the provider under Article 13 AIA Proposal to comply with their obligation to carry out a data protection impact assessment under Article 35 GDPR.

Article 29 (2) clarifies that these obligations are without prejudice to other user obligations under Union or national law and to the user's discretion in organising resources and activities for the purpose of implementing the human oversight measures indicated by the provider. This means that the obligations of businesses following from consumer protection law remain entirely unaffected by Article 29 AIA Proposal. Such obligations following from consumer protection law may be AI-specific, such as the newly introduced obligation of traders under Article 6 (1) (ea) CRD to inform about any personalised pricing, the likewise newly introduced obligation of online marketplaces under Article 6a (1) (a) CRD to inform about the main parameters determining ranking of offers presented to the consumer as a result of a search query and the relative importance of those parameters as opposed to other parameters, or the related provision on misleading omissions in Article 7 (4a) UCPD. More often, such obligations following from consumer protection law will not be AI-specific, though, and apply irrespective of whether the business uses AI for its operations or not.

3.4 Non-discrimination law

According to the Explanatory Memorandum,[31] the AIA Proposal complements existing Union law on non-discrimination with specific requirements that aim to minimise the risk of algorithmic discrimination, in particular in relation to the design and the quality of data sets used for the development of AI systems complemented with obligations for testing, risk management, documentation and human oversight throughout the AI systems' lifecycle. Union law on non-discrimination exists, in particular, in the employment sector, but also with regard to general access to goods and services in mass transactions.[32] It is to be expected that future Union law will put a stronger focus on non-discrimination in a variety of sectors.[33]

While the Recitals make ample reference to risks of algorithmic discrimination and to all the necessity to ensure non-discrimination and gender equality, the term '(non-

[31] COM(2021) 206 final, p. 4.
[32] Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services, OJ L 373, 21.12.2004, p. 37–43; Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin, OJ L 180, 19.7.2000, p. 22–26.
[33] See, e.g., Article 6 of the recent Proposal for a Directive of the European Parliament and of the Council on consumer credits, COM(2021) 347 final.

)discrimination‘ is not used even once throughout the operative normative text of the AIA Proposal. In the light of the dominant role which algorithmic discrimination plays in the public debate about the ethical and legal implications of AI, it may come as a surprise that discrimination is not even mentioned in the list of prohibited AI practices in Article 5.

Again, the drafters obviously wanted to avoid any overlap between the list of prohibited AI practices and existing non-discrimination law, but also to avoid any extension of non-discrimination law specifically with regard to the use of AI.

3.5 Data protection law

3.5.1 AI as a data driven technology

According to the Explanatory Memorandum[34] the AIA Proposal is without prejudice to the rules laid down in data protection law. This means in essence that any processing of personal data that occurs in the context of the development or deployment of AI has to comply with the GDPR and other relevant data protection law, such as the EUDPR[35] and national provisions implementing the Law Enforcement Directive (LED)[36] and the E-Privacy Directive[37].

[34] COM(2021) 206 final, p. 4.

[35] Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data, and repealing Regulation (EC) No 45/2001 and Decision No 1247/2002/EC, OJ L 295, 39-98.

[36] Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA, OJ L 119, 4.5.2016, p. 89–131.

[37] Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications), OJ L 201, 37-47.

Artificial Intelligence is a data-driven technology.[38] Data plays a decisive role already in the development of AI systems, in particular as training, validation and testing data. Much of what is popularly known as AI and has been included as a technology in Annex I of the AIA Proposal has not been fully programmed by traditional coding techniques and is not rule-based, but has learned how to execute tasks through a process that improves performance through experience and requires large amounts of data.

However, AI is also an algorithmic system. Once it has been developed and is deployed for a variety of tasks, it is used for processing input data in order to obtain particular output data, such as a classification, prediction or recommendation. Output data may immediately trigger a reaction by physical actuators (in which case we tend to speak of 'robotics'), or a non-physical reaction of some kind such as the placing of orders in high-frequency trading (in which case we often use the term 'autonomous agents'), or merely serve as a basis for human decision-making (e.g. recommender systems). There exists a spectrum as to the division of tasks between human and machine, and as to the degree of automation. Output data may again become input or training data to the same or a different system.

3.5.2 The data perspective and the algorithm perspective

The 'data perspective' and the 'algorithm perspective' are not just two sides of the same coin. Instead, they represent two different ethical discourses, which both complement each other and are contingent upon each other, and which are typically also reflected in different acts of legislation.[39]

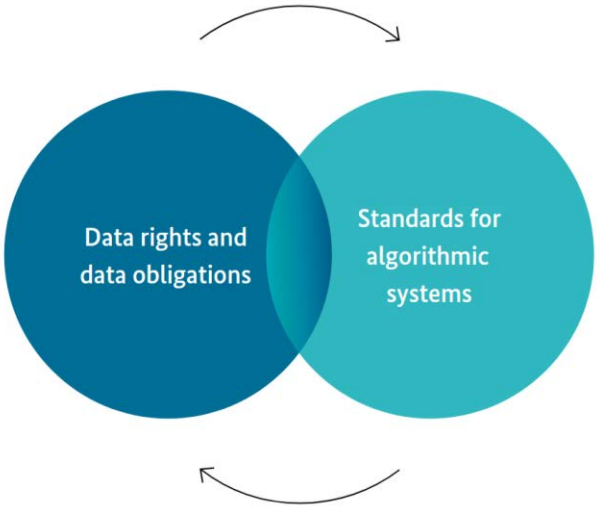

Figure 8: Data perspective and algorithm perspective
(Source and ©: Data Ethics Commission 2019)

[38] On the role of data in AI see *Wendehorst* et al, Framework paper for GPAI's work on data governance (2020), Global Partnership on AI (GPAI), p.6.
[39] Opinion of the Data Ethics Commission (2019), p. 77.

The data perspective focuses on the context of meaning and the semantics of data, in particular the origin of data and the potential impact their processing may have on parties to whom the information coded in the data refers. A central distinction in this context is that between personal and non-personal data,[40] and central debates are those around data subjects' rights, rights in co-generated data[41] (which is arguably a concept that is to be preferred over the concept of 'data ownership rights') and data sharing. Pieces of legislation specifically addressing the data perspective include the GDPR, EUDPR, LED and E-Privacy Directive. Upcoming pieces of legislation include the Data Governance Act[42] and the Data Act[43].

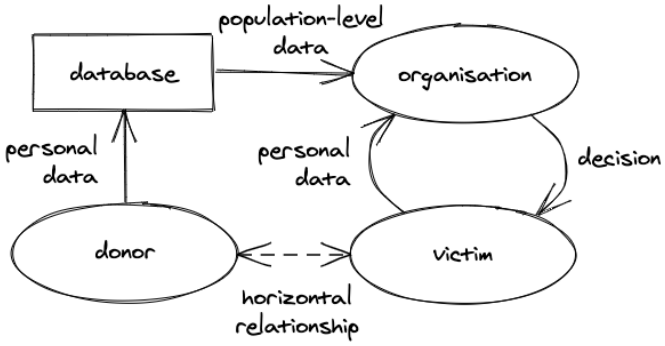

Figure 9: The multirelational nature of data use (Source and ©: Jeni Tennison 2020)

By way of contrast, the algorithm perspective focuses on the architecture of data-driven algorithmic systems, their dynamics and the systems' impacts on individuals and society. Central concerns are automation as such and the outsourcing of increasingly complex operational and decision-making processes to machines.[44] The algorithms perspective differs from the data perspective in that the individuals affected by the system may not necessarily have anything to do with the original training, validation and testing data, and the link between those subjects and any input data or output data may be of a more coincidental nature and may not be at the focus of concern. Central

[40] See Article 29 Working Party, Opinion 4/2007 on the Concept of Personal Data, WP 136.

[41] This concept was created, and the term coined, by the ALI-ELI Principles for a Data Economy, for the most current draft see ALI-ELI Principles for a Data Economy – Data Transactions and Data Rights, ELI Final Council Draft (August 2021). The concept has meanwhile gained widespread recognition, see COM(2020) 66 final p. 8, 13; European Commission, Inception impact assessment, Ares(2021)3527151, p. 1; Opinion of the Data Ethics Commission (fn. 39), p. 85 ff.

[42] Proposal for a Regulation of the European Parliament and of the Council on European data governance (Data Governance Act), COM(2020) 767 final.

[43] See schedule at https://www.europarl.europa.eu/legislative-train/theme-a-europe-fit-for-the-digital-age/file-data-act.

[44] On the risks of outsourcing decision-making in the context of employment see *Allhutter* et al, Der AMS-Algorithmus (2020), iTA.

debates are those around algorithmic fairness, non-discrimination, or human oversight. The AIA proposal aims to address the algorithm perspective in a horizontal manner.

The line between the data perspective and the algorithm perspective is blurred, and one and the same use of an AI system often raises concerns both from the point of view of the data perspective and from the point of view of the algorithm perspective.

Illustration 6

Company V operates an online video game. When users such as G are playing the game, an AI system in the background analyses every single of their reactions to a broad variety of situations, meticulously measuring all sorts of behavioural traits, resulting in an extremely granular behavioural profile. The creation of the behavioural profile of G raises massive concerns from the point of view of the data perspective, and with the protection of G as the person to whom the information recorded in the data refers (the 'data subject') in mind.

Illustration 7

Company V may not only use the data for creating a profile of G. Rather, the data activities may also affect third parties sharing a number of characteristics (such as age, profession, family situation, shopping habits, browsing history etc.) with G, because inferences are drawn from G's behaviour to the behaviour of such third parties, assuming that they will react in a given situation in very much the same manner as G. These assumptions may be fed into AI used for recruitment, for personalised pricing, for credit rating and all sorts of other sensitive purposes. The use of such AI raises massive concerns from the point of view of the algorithm perspective, and the focus is on the protection of any individual affected by the system in the labour or consumer context, more or less without regard to the concrete personal input data of these individuals (such as age, profession etc.) that are processed during such use.

4 Analysing the Interplay between the AIA and the GDPR (and LED)

4.1 Interplay between Articles 6-9 GDPR (8-10 LED) and the AIA

4.1.1 General observations

Any processing of personal data that occurs in the context of the development or deployment of AI has to comply with the GDPR and/or other relevant data protection law, such as national law implementing the LED or E-Privacy Directive. As this study puts a focus on the consumer perspective it will mainly concentrate on analysing the interplay between the AIA and the GDPR. However, as a number of provisions in the AIA Proposal that are within the scope of this study specifically address AI practices by law enforcement authorities also the LED will have to be taken into account.

It is one of the cornerstones of European data protection law that processing of personal data is prohibited unless it is justified by a legal ground. Under the GDPR, any processing of personal data must be based on a legal ground within the meaning of Article 6 GDPR and, in the case of special categories of data referred to in Article 9 GDPR, one of the grounds listed in Article 9 (2) GDPR. In the private sector, the most important legal grounds within Article 6 (1) GDPR are consent (a), contract (b) and legitimate interests (f). In the public sector, the most important legal grounds are consent (a), protection of vital interests (d) and fulfilment of tasks in the public interest (e). The legal ground of fulfilment of legal obligations (c) may become relevant both in the private and public sector, but hardly ever plays a predominant role.

When it comes to the processing of specific categories of data within the meaning of Article 9 (1) GDPR (e.g. biometric data) processing of such data must, in addition, fall under one of the grounds listed in (2).[45] Controllers from the private sector are mostly restricted to explicit consent (a), manifest publication by the data subject (e), or research (j), unless they fall under one of the specific categories, such as data processing in the

[45] See Recital 51 GDPR.

medical context (h) or employment context (b) or for the establishment, exercise or defence of legal claims (f). For controllers from the public sector, Article 9 (2) includes the general clause of substantial public interest (g) on top of more specific grounds such as protection of public health (h) or statistical purposes (j).

According to Article 8 LED, Member States shall provide for processing to be lawful only if and to the extent that processing is necessary for the performance of a task carried out by a competent authority for law enforcement purposes that it is based on Union or Member State law. Member State law regulating processing for law enforcement purposes must specify at least the objectives of processing, the personal data to be processed and the purposes of the processing.

4.1.2 How Article 5 AIA relies on the GDPR: the case of social scoring

The way in which the AIA Proposal relies on data protection law is best illustrated by the example of Article 5, and in particular the rules on social scoring in Article 5 (1) (c) and on biometric identification in Article 5 (1) (d) and (2) to (4). Even though the wording of Article 5 does not include any explicit reference to the GDPR or the LED it is clear that the way in which the provisions on social scoring and biometric identification have been phrased can only be explained by the existence of data protection law as a background regime.

Starting with social scoring, this would normally be something covered by the GDPR and not by the LED. 'Law enforcement' is defined as ''the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, including the safeguarding against and the prevention of threats to public security' by competent authorities.[46] While social scoring may potentially serve law enforcement purposes by creating incentives for law-abiding behaviour its focus is clearly a different one, which is why social scoring would require justification under the GDPR.

As far as social scoring does not rely on the processing of specifically sensitive categories of personal data within the meaning of Article 9 GDPR it requires justification under Article 6 GDPR. It would be largely impossible to justify social scoring by way of direct

[46] See Article 1 (1) LED and the corresponding exclusion from the scope of the GDPR in Article 2 (2) (d) GDPR.

reliance on the GDPR (i.e. without Member State law creating a link between the activity and the GDPR), in particular for players from the private sector.

Illustration 8

Company G offers a gatekeeper platform service (such as a dominant social network or online marketplace). G defines the conditions for its users, including premium content and services offered and the calculation of prices, according to the users' individual CO2 footprints as expressed in the users' individual 'climate score'. This climate score is calculated on the basis of data collected from a variety of sources and a variety of different contexts.

Such data use by G would not be 'necessary' for the purposes of the legitimate interests pursued by G or by a third party within the meaning of Article 6 (1) (f) GDPR. Even if one accepts that the fight against climate change and global warming is a legitimate interest, this legitimate interest would arguably not justify extensive collection of personal data and detrimental or unfavourable treatment of individuals according to their CO2 footprints in areas that are unrelated to the activities creating emissions. In other words, the legitimate interest would be overridden by the interests and fundamental rights and freedoms of the data subjects. At the end of the day, the only possible legal ground would be consent under Article 6 (1) (a) GDPR, but for consent to be truly 'free' a data subject may not suffer any detriment in the case of refusal or withdrawal of consent where such detriment goes beyond what is strictly necessary to protect the legitimate interests of the controller.[47]

Illustration 9

Public authority P decides about the allocation of social benefits and access to public services, including public transport systems, libraries and many other facilities, according to the citizens' individual CO2 footprints as expressed in each

[47] See Recital 42 GDPR, which does not even make an exception for the legitimate interests of the controller.

citizen's individual 'climate score'. This climate score is calculated on the basis of data collected from a variety of sources and a variety of different contexts.

In illustration 9, P might argue that this data use is 'necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller' within the meaning of Article 6 (1) (e) GDPR. This would require that P is, in the relevant Member State, tasked with combating climate change, which a Member State is free to do at any time because the allocation of tasks within public administration remains within the discretion of Member States. Alternatively, the relevant Member State could introduce a law obliging public authorities and potentially also parties from the private sector, such as gatekeeper platform G in the previous illustration, to collect relevant data and calculate a 'climate score' of citizens, which would mean that the activities could also be based on Article 6 (1) (c) GDPR. Admittedly, Article 6 (3) provides that such Member State law shall meet an objective of public interest and be proportionate to the legitimate aim pursued, and this provision would have to be seen in the light of the general principles relating to the processing of personal data enshrined in Article 5 GDPR (such as the principle of data minimisation) as well as in the light of data protection as a fundamental right under the Charter. However, it is equally clear that Member States have a wide margin of discretion in that regard and that it is very difficult to enforce sanctions for violation of EU primary law against a Member State.

This is why, in the case of public authorities or private parties obliged to engage in such data processing by Member State law, the GDPR does not provide for sufficient safeguards against social scoring, which in turn provides an explanation for the restriction of the prohibition in Article 5 (1) (c) AIA Proposal.

4.1.3 How Article 5 AIA relies on the GDPR: the case of biometric techniques

4.1.3.1 General observations

The way in which the AIA Proposal relies on the GDPR and LED can also be illustrated by the example of biometric identification addressed in Article 5 (1) (d) and (2) to (4). In particular, the restriction to the use of such identification system for the purpose of law enforcement finds its explanation in the fact that the LED tends to give a higher degree of discretion and leeway to public authorities than the GDPR, and that private actors have

hardly any leeway at all under Article 9 (2) GDPR when it comes to the processing of biometric data. Beyond very specific contexts, such as the medical context, private actors normally need to rely on the data subject's explicit consent within the meaning of Article 9 (2) (a) GDPR.

4.1.3.2 Biometric identification

The following Tables[48] illustrate the way in which the AIA proposal and data protection law interact when it comes to the admissibility of particular biometric techniques, indicating whether the GDPR, the LED or the AIA would normally allow a particular practice. Table 1 shows the situation with regard to biometric identification.

Table 1: Admissibility of biometric identification (based on simplified assumptions)

Scenario	GDPR	LED	AIA
Law enforcement authorities use real-time remote biometric identification to search for a known terrorist.	n.a.	Yes	Yes*)
Law enforcement authorities use real-time remote biometric identification in public spaces to detect anomalous behaviour in the population.	n.a.	MS law	No
Private security company in charge of securing an airport uses real-time remote biometric identification on airport premises	MS law	n.a.	Yes
Private owner of residential premises applies remote biometric identification to people passing by on the street	No	n.a.	Yes

n.a.= not applicable MS = Member State *) = but high-risk, i.e. Title III AIA applies

In Scenario 1 (police uses real-time remote biometric identification to trace down a terrorist) and the applicable data protection regime is clearly the LED, which would allow this kind of use of biometric data under Article 10, both because tracing down terrorists is normally necessary to protect the vital interests of potential victims (b) and because tracing down terrorists will certainly be mandated by Union or Member State law (a). Also the AIA would allow that kind of use under Article 5 (d) (ii) or (iii) because tracing down

[48] *Wendehorst/Duller*, Biometric recognition and behavioural detection, European Parliament (2021).

terrorists serves the prevention of terrorist attacks and because being a terrorist means being a perpetrator or suspect of a serious criminal offence of the type referred to in that provision. However, the AIA would still allow the use only under certain conditions mentioned in Article 5 (2) to (4), and the application would count as a 'high-risk' AI system within the meaning of Annex III, i.e. Title II of the AIA would fully apply.

In Scenario 2 (police uses real-time remote biometric identification to detect anomalous behaviour in the population) the applicable data protection regime would again be the LED. Article 10 (a) would allow such a practice is authorised by Member State law, which would have to restrict the use to cases where it is strictly necessary and subject to appropriate safeguards for the rights and freedoms of the data subject. Given that these restrictions are very vague and Member States have a lot of leeway to define what they find strictly necessary or an appropriate safeguard, the AIA has introduced an additional protective regime.

In Scenario 3 (private company uses real-time remote biometric identification on airport premises) the applicable data protection regime is the GDPR. Article 9 (2) (g) GDPR would allow such processing if it is necessary for reasons of substantial public interest, on the basis of Union or Member State law, provided such law is proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject. Whether or not Article 5 AIA addresses this situation depends on whether the activity qualifies as 'law-enforcement' under Article 3 (41) and whether the company qualifies as 'law-enforcement authority' within the meaning of Article 3 (40). This may be the case where, under Member State law, the private company is entrusted to exercise public power for the purpose of the prevention of threats to public security, otherwise the situation is not addressed by Article 5 (but the application would still count as a 'high-risk' AI system).

In Scenario 4 (owner of private premises wants to secure the premises by means of remote biometric identification) the applicable data protection regime would again be the GDPR. However, beyond very specific situations where significant public interests come into play, the only legal ground the private owner could rely on would-be explicit consent under Article 9 (2) (a) GDPR, which is impossible to get from unknown data subjects walking by on the street. So, while this use of remote biometric identification would clearly not be covered by the prohibition under Article 5 AIA it would anyway not be permissible under the GDPR.

4.1.3.3 Biometric categorisation

The situation with biometric categorisation differs from the situation with biometric identification. The reason is that biometric identification relies on the use of 'biometric data' as this term is defined in both the GDPR (and LED) as well as in the AIA. This definition is restricted to data of a person 'which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic data'. However, biometric categorisation will normally rely on 'soft biometrics' that allow the assignment of natural persons to specific categories, such as a specific sex, age, or ethnic origin, but that does not usually allow the unique identification of a natural person.

Table 2: Admissibility of biometric categorisation (based on simplified assumptions)

Scenario	GDPR	LED	AIA
Asylum authorities use biometric categorisation system for age-verification of juvenile migrants	Yes	n.a.	Yes*)
Law enforcement authorities use biometric categorisation for targeted surveillance of individuals with particular ethnic origin	n.a.	No	Yes*)
Provider of online game uses biometric categorisation for age-verification as a child protection measure	Simple consent	n.a.	Yes
Online store uses biometric categorisation of professional customers (e.g. according to age groups) for targeted economic exploitation of vulnerabilities	Simple consent	n.a.	Yes

n.a.= not applicable MS = Member State *) = but high-risk, i.e. Title III AIA applies

In Scenario 5 (asylum authorities use biometric categorisation for age-verification of juvenile migrants) the applicable data protection regime is the GDPR, and authorities can usually process the relevant data on the legal ground established by Article 6 (1) (e) GDPR, i.e. exercise of public authority vested in the asylum authorities. Under the AIA, this system would not fall under any of the prohibited or restricted practices listed in Article 5, but it would qualify as high-risk AI system according to Annex III 7 (d).

In Scenario 6 (targeted surveillance by police based on ethnic origin) the applicable data protection regime is the LED. Under Article 11 (3) LED, profiling that results in discrimination against natural persons on the basis of special categories of personal data

referred to in Article 10, including ethnic origin, is prohibited. So, where there is no other justification for enhanced surveillance of particular individuals and where enhanced surveillance is exclusively based on ethnic origin such practice would, as a discriminatory practice, be prohibited under the LED. If it were not prohibited in any case, it would, under the AIA, qualify as a high-risk system under Annex III 6 (e) because it is a system used by law enforcement authorities for predicting the occurrence of criminal offences based on assessing particular characteristics of groups.

In Scenario 7 (age verification of online gamers for child protection purposes) the applicable data protection regime would be the GDPR. Data processing would be based on the legal ground of legitimate interests in Article 6 (1) (f), if not already fulfilment of a legal obligation within the meaning of Article 6 (1) (c). Currently, this AI system would not be covered by any particular rules of the AIA.

In Scenario 8 (biometric categorisation of customers for exploitation of vulnerabilities) the applicable data protection regime would be the GDPR. The only legal ground for data processing that would potentially be available to the user of that system is the customers' simple consent within the meaning of Article 6 (1) (a). Given that the GDPR does not contain any hard and fast substantive limits to consent and restrict itself to requirements of the more procedural nature, it is possible that data processing is permissible under the GDPR. The AIA Proposal does address the exploitation of-specific vulnerabilities, but Article 5 (1) (b) is restricted to practices that cause physical or psychological harm and does not cover practices that merely cause economic harm. This practice might thus be acceptable under data protection law and the AIA. As the 'victims' are professional customers, they are not even protected by consumer protection law (on which see below 5.1.1.2 and 5.1.2.2).

4.1.3.4 Emotion recognition

Like biometric categorisation, also emotion recognition systems rely on the processing of personal data that do not qualify as 'biometric data' within the meaning of the GDPR.

Table 3: Admissibility of biometric categorisation (based on simplified assumptions)

Scenario	GDPR	LED	AIA
Police uses emotion recognition system during interrogation of suspect	n.a.	Yes	Yes*)
Statistics authorities use emotion recognition in voting booths to find out about people's attitude towards democracy (e.g. anger, satisfaction)	MS law	n.a.	Yes
Q&A chat-bot uses emotion recognition to react appropriately to very dissatisfied customers	Yes	n.a.	Yes
Social network uses emotion recognition (detecting fear, anger and other emotions) for targeted political advertising, exploiting individual vulnerabilities.	Simple consent	n.a.	Yes

n.a.= not applicable MS = Member State *) = but high-risk, i.e. Title III AIA applies

In Scenario 9 (police uses emotion recognition during interrogation of the suspect) the applicable data protection regime is the LED. Given that, again, no biometric data within the strict definition of the GDPR and the LED are at stake, processing of data would be based on the general clause of Article 8 LED. The practice is not a prohibited practice under the AIA. While emotion recognition systems are normally only subject to transparency obligation under Article 52 (2) AIA this transparency obligation does not apply where the system is used to investigate criminal offences. The system would qualify as a high risk system under Annex III 6 (b).

In Scenario 10 (statistics authorities analyse the emotional state of voters during elections) the applicable data protection regime is the GDPR. Authorities would most probably rely on Article 6 (1) (e), i.e. performance of tasks carried out in the public interest. While it could seem questionable whether the practice meets the general proportionality test, it also has to be borne in mind that research and statistical purposes enjoy a privileged status under the GDPR in various respects, so there would possibly be little to stop the authorities from applying that practice if Member State law so provides and unless the practice violates primary EU law. Under the AIA the only requirement would be a transparency obligation under Article 52 (2).

In Scenario 11 (Q&A chat bot uses emotion recognition to ensure appropriate reactions) it is again the GDPR that applies. Depending on the situation, data processing could be

justified by (consent or) Article 6 (1) (b) if there is a contract or, more often, by legitimate interests under Article 6 (1) (f). Under the AIA, there would again just be the transparency obligation of Article 52 (2).

Finally, in Scenario 12 (social network uses emotion recognition for exploiting individual vulnerabilities during an election campaign), the applicable data protection regime is again the GDPR. Data processing could only be justified by simple consent under Article 6 (1) (a), which is, however, easily given as most people tend to routinely press any 'OK' buttons presented to them.[49] There is nothing to stop this practice under the AIA, which would, again, only impose the transparency obligation of Article 52 (2).

4.1.4 Conclusions

The brief analysis of the two different illustrations with regard to social scoring in the form of a 'climate score' (above at 4.1.2) and of the 12 different scenarios involving biometric techniques (above at 4.1.3) has demonstrated how Articles 6 to 9 GDPR and 8 to 10 LED on the one hand and Article 5 of the AIA Proposal on the other hand fulfil a very similar function in 'filtering out' undesirable data practices or algorithmic practices by way of a prohibition. However, it has also become clear that this function is not fulfilled in a very consistent manner. There are still gaps when it comes to state activities that do not qualify as law enforcement activities as they are not covered by Article 5 AIA and as Member States have quite some leeway under data protection law, including with regard to practices that pose significant fundamental rights risks and may lead to 'function creep' (see Scenario 10). There are also some scenarios where a highly undesirable practice is not filtered out by either data protection law or the AIA Proposal, usually because it is not specifically addressed by the AIA and because data processing may be justified by simple consent under Article 6 (1) (a) GDPR. Considering that the GDPR largely refrained from imposing any substantive limits on consent and considering also the phenomenon of 'consent fatigue' and the fact that people tend to simply click on any okay button presented to them, this threshold is not very high and is not appropriate (see Scenarios 8 and 12).

[49] *Sartor/Lagioia/Galli*, Regulating Targeted and Behavioural Advertising in Digital Services, European Parliament (2021), p. 103.

4.2 Interplay between Article 22 GDPR (11 LED) and the AIA

4.2.1 General observations on Article 22 GDPR (and 11 LED)

4.2.1.1 Scope and effect of the prohibition in Article 22 GDPR

Article 22 GDPR contains specific individual rights against 'automated individual decision-making, including profiling'. This formulation is remarkable in itself because it seems to combine two different, albeit related, phenomena. According to Article 4 (4) 'profiling' means any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements'. The use of the word 'evaluating' suggests that profiling involves some form of assessment or judgement about a person. A simple classification of individuals based on known characteristics such as their age, sex, and height does, in the eyes of the Article 29 WP,[50] not necessarily lead to profiling. Rather, this depends on the purpose of the classification. The Article 29 WP gives the example that a business wishes to classify its customers according to their age or gender for statistical purposes and to acquire an aggregated overview of its clients without making any predictions or drawing any conclusion about an individual. In this case, the purpose is not assessing individual characteristics and is therefore not profiling.[51]

While there is no separate definition of 'automated individual decision-making' it becomes clear from the wording of Article 22 (1) GDPR that what is meant is a decision that is based solely on automated processing of data.[52] Automated decisions can be made with or without profiling; profiling can take place without making automated decisions.[53] Therefore, Article 22 may apply even where the automated decision does not involve any

[50] Article 29 Data Protection Working Party, Guidelines on Automated individual decision-making and profiling for the purposes of Regulation 2016/679, last revised and adopted 6 February 2018, p. 7.
[51] Article 29 WP (fn. 50), p. 7.
[52] *Hladjk* in *Ehmann/Selmayr* DSGVO, Art. 22 para. 6.
[53] Article 29 WP (fn. 50), p. 8.

form of assessment or judgment about a person and is merely the result of 'mechanical' calculation of factors previously defined by a human.

Article 22 GDPR leaves open what it means by 'decision'. While some tend to take the view that any kind of output data that potentially trigger some kind of reaction, either directly or indirectly, amounts to a decision[54] this understanding is arguably too broad for an appropriate understanding of Article 22. The explicit reference to legal effects seems to indicate that, in particular, the mere triggering of physical actuators is not included. Where a care robot or connected car 'decides' to make a particular movement that movement may have severe consequences for individuals (in the worst case: kill them) but this is arguably not what is meant by Article 22 GDPR, for otherwise this provision would mean a prohibition of robotics in many contexts. Rather, it seems that Article 22 is more focused on decisions that are similar to the types of decisions that tend to have legal effects, such as provision or denial of service, making or denial of an offer, price calculation etc.[55] It is for decisions of that kind that Article 22 (1) provides that an individual has the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

Article 22(1) establishes a general prohibition for decision-making based solely on automated processing. This prohibition applies whether or not the data subject takes an action regarding the processing of their personal data.[56] The WP29 has stated clearly that the controller cannot avoid the Article 22 provisions by 'fabricating' human involvement, such as by having an employee routinely confirm automatically generated decisions without any actual influence on the result. To qualify as human involvement, the controller must 'ensure that any oversight of the decision is meaningful, rather than just a token gesture. It should be carried out by someone who has the authority and competence to change the decision. As part of the analysis, they should consider all the relevant data.'[57] Any processing likely to result in a high risk to data subjects requires the controller to carry out a Data Protection Impact Assessment (DPIA).[58]

[54] Opinion of the Data Ethics Commission (fn. 39), p. 160 ff.
[55] Article 29 WP (fn. 50), p 10; *Hladjk* (fn. 52) DSGVO, Art. 22 para 9.
[56] Article 29 WP (fn. 50), p. 19.
[57] Article 29 WP (fn. 50), p. 21.
[58] Article 29 WP (fn. 50), p. 20.

However, paragraph (2) provides for three exceptions, which are: the data subject's explicit consent, necessity for entering into (or performance of) a contract between the data subject and a data controller, or authorisation by Union or Member State law which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests. In the first two cases (consent or contract) the data controller must implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.

None of the exceptions may be invoked for decisions that are based on special categories of personal data referred to in Article 9 (1), such as biometric data, unless this is based on the data subject's explicit consent or justified by a substantial public interest, on the basis of Union or Member State law which shall be proportionate to the aim pursued and respect the essence of the right to data protection, and where suitable and specific measures are in place to safeguard the fundamental rights and the interests of the data subject.

4.2.1.2 Significance of Article 22 GDPR for information rights

Article 22 GDPR is also relevant for the data subjects' information and data access rights. Under Articles 13 and 14 GDPR the controller must actively provide the data subject with particular information. Where personal data are directly collected from the data subject the information duties follow from Article 13, otherwise they follow from Article 14. According to Article 15(1), the data subject has a right to obtain particular information against a controller of personal data at any time.

Items of information that need to be provided are very similar. In particular, they include the purposes of the processing for which the personal data are intended as well as the legal basis for the processing (Articles 13 (1) (c), 14 (1) (c) and 15 (1) (a)) and, where the processing is based on legitimate interests, the concrete legitimate interests pursued if the data is collected from the data subject (Article 13 (1) (d)). Even more importantly in the AI context, information duties include 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 (Articles 13 (2) (f), 14 (2) (g) and 15 (1) (h)). Under Article 14, there are various exceptions from the information duty, notably when the provision of such information proves impossible or would involve

a disproportionate effort, in particular for processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes.

Under Article 13, the information must be provided when the data are obtained from the data subject, whereas under Article 14 the controller must provide the information within a reasonable period after obtaining the personal data, but at the latest within one month, having regard to the specific circumstances in which the personal data are processed. If the personal data are to be used for communication with the data subject, information under Article 14 must be given at the latest at the time of the first communication, and if a disclosure to another recipient is envisaged, information must be given at the latest when the personal data are first disclosed. Where the data subject actively requests information under Article 15 (1) the controller must act, according to Article 12 (3), without undue delay and in any event within one month of receipt of the request; that period may be extended by two further months where necessary, taking into account the complexity and number of the requests. All information must be provided free of charge, but where requests from a data subject are manifestly unfounded or excessive the controller may either charge a reasonable fee or refuse to act on the request.

In contrast with the data subject's right to obtain information under Article 15 (1), Article 15 (3) establishes a right of the data subject to receive a copy of the personal data undergoing processing, unless to the extent that this would adversely affect the rights and freedoms of others. For any further copies requested by the data subject, the controller may charge a reasonable fee based on administrative costs. Where the data subject makes the request by electronic means, and unless otherwise requested by the data subject, the information must be provided in a commonly used electronic form.

This right to receive a copy is similar to the portability right under Article 20, but the latter is different in several respects. Most notably, pursuant to Article 15(3), the controller has a duty to make available the data used as input to create the profile as well as access to information on the profile and details of which segments the data subject has been placed into.[59] Under Article 20 the controller only needs to communicate the data provided by the data subject or observed by the controller and not the profile itself. While Recital 63 provides some protection for controllers concerned about revealing trade secrets or

[59] WP29 (fn. 50), p. 17.

intellectual property, controllers cannot rely on the protection of their trade secrets as an excuse to deny access or refuse to provide information to the data subject. [60]

4.2.1.3 Article 22 GDPR as a misplaced provision

Given that the focus of concerns raised by automated decision-making is clearly on the algorithm perspective and not on the data perspective (see above at 3.5.2), Article 22 GDPR is, strictly speaking, misplaced, as it would make much more sense to have this and further standards for algorithmic decision making in the AIA. After all, the individuals affected by automated decision-making within the meaning of Article 22 GDPR are very often not identical with the individuals primarily subjected to data collection and/or profiling activities.[61] The failure of the GDPR to address properly the multi-relational nature of data processing and use has often been criticised, but it is arguably not possible to address this multi-relational nature within the framework of the GDPR or any other legislative instrument that is focused on the data perspective. In a similar vein, the controller of personal data is not the proper addressee of the prohibition under Article 22 GDPR because the point is not processing of the data that is actually being controlled, but rather the way in which – in a context not specifically related to data – decisions are being made. In other words, the specific risks inherent in automated individual decision-making are largely unrelated to the type and quantity of personal data relating to an affected individual that is being processed for the purposes of the automated decision-making. Quite on the contrary, automated individual decision-making may be the more problematic the less personal data of the affected individual are actually being processed.

Illustration 10

Applications for a vacant position are analysed by HR software with a view to creating a shortlist of 20 applicants whose applications will be evaluated by HR department staff. This is a fully automated decision covered by Article 22 GDPR. The specific sensitivity of the situation is entirely unrelated to traditional privacy

[60] WP29 (fn. 50), p. 17.

[61] For an explanation see only *Tennison* Individual, collective and community interests in data, Open Data Institute (2020).

concerns.[62] For instance, the sensitivity does not change where applicants' personal data are subjected to encryption and strong pseudonymisation measures, or where the data processed are reduced to a minimum. Quite on the contrary, where data processed are reduced to a minimum (e.g. the system only looks at the type font used and discards from the outset all applicants using Times Roman or Arial) this may be particularly problematic and totally unacceptable.

There may even be extreme cases where an individual is subjected to automated decision-making without any personal data of that individual being processed in a way triggering the application of the GDPR.

Illustration 11

During the COVID-19 pandemic, gates to an airport building automatically deny access to anyone wishing to enter, irrespective of whether or not that individual holds a ticket or urgently needs a flight, once the maximum number of individuals allowed on the premises has been reached. In the light of the potential significance of that decision for an individual this decision should normally be covered by Article 22 GDPR. However, as the automated decision-making system closes the gate without regard to any personal data relating to the individual to whom entry is denied, the GDPR does not apply at all to this decision.

So, ideally, Article 22 GDPR should be transferred to the AIA or, in order to avoid gaps for algorithmic systems not qualifying as AI systems under the AIA, be derogated by a more specific provision to be inserted in the AIA. The author is, of course, aware of the fact that this may prove to be unrealistic.

[62] This is why it is hardly convincing that Article 29 WP (fn. 50), p. 23 requires in this case that the HR department consider whether a "less privacy-intrusive measure" could be adopted – privacy is not at stake here.

4.2.2 How Article 14 AIA relates to Article 22 GDPR

4.2.2.1 General relationship

Theoretically, there is no overlap between Article 14 AIA and Article 22 GDPR, as Article 22 GDPR exclusively addresses the controller, who is usually identical with what the AIA calls the 'user', while Article 14 AIA addresses those players who design and develop high-risk AI systems. Therefore, the counterpart of Article 22 GDPR within the AI is not Article 14, but Article 29 on the obligations of users of high-risk AI systems. Technically speaking, the prohibitions enshrined in Article 22 GDPR qualify as 'other user obligations under Union or national law' within the meaning of Article 29 (2) AIA. However, practically speaking, there is of course overlap and potential friction between the two provisions. This is so because those designing and developing high-risk AI systems should hardly be allowed to design and develop AI systems in a way that makes the users of the AI system, when they follow the instructions issued under Article 13 (3) (d) AIA, violate Article 22 GDPR.

4.2.2.2 Scenarios with meaningful human involvement

Article 14 AIA is insofar much broader in its scope than Article 22 GDPR as it covers AI systems without regard to the degree of human involvement in a decision. Notably, Article 14 AIA also covers recommender systems where a human takes the recommendation by the AI system only as a basis for its own free decision, i.e. where the human involvement is so meaningful and substantial that Article 22 GDPR does not apply.

The following table shows how the two provisions apply to five different applications of recommender systems.

Table 4: Recommender systems (based on simplified assumptions)

Scenario	Article 22 GDPR		Article 14 AIA	
	Covered by scope	Admissible with appropriate safeguards	Covered by scope	Requirements may be met
Recommender system for (internal) price calculation with ultimate decision by human employee	–	n.a.	–	n.a.

Scenario	Article 22 GDPR		Article 14 AIA	
	Covered by scope	Admissible with appropriate safeguards	Covered by scope	Requirements may be met
Recommender system for deciding about cancellation of a contract for rental of residential premises (private sector) with ultimate decision by human employee	–	n.a.	–	n.a.
Recommender system for student admission with ultimate decision by human employee	–	n.a.	+	+
Recommender system for passport control based on remote biometric identification with verification/decision by one(!) human employee	–	n.a.	+	–
Recommender system for targeted passport control based on remote biometric identification with verification and decision by two human employees	–	n.a.	+	+

As such applications are not covered by Article 22 GDPR the user is only subject to the obligations under Article 29 AIA, which include, via the instructions for use that have to comply with Article 13 (3) (d), the requirements of Article 14 AIA. Whether or not Article 14 AIA applies at all depends on whether or not the AI system at hand qualifies as a high-risk AI system within the meaning of Article 6 (2) and Annex III. This is the case in Scenarios 3 to 5 above, but not in Scenarios 1 and 2.

Attention should be drawn to a rather peculiar provision in Article 14 (5) AIA. According to that provision, AI systems intended to be used for any remote biometric identification of natural persons (as defined in Article 3 (36)) must not lead to any action or decision by the user on the basis of the identification resulting from the system unless this has been verified and confirmed by at least two natural persons. This provision does not seem to make much sense because it is both insufficient and overreaching.[63] 'No action or decision' would mean that not even identity control (such as requesting a passport) may follow from a high matching score, which would turn biometric identification by AI close to completely useless. This is aggravated by the fact that, in a situation where immediate action is of the essence, two confirmations may not be possible. On the other hand, the rule in Article 14 (5) is insufficient because it does not give any guidance as to the

[63] *Wendehorst/Duller* (fn. 48), p. 82.

independence of the two natural persons, nor on the training they have received or on the means they use.

4.2.2.3 Scenarios without meaningful human involvement

Where there is no meaningful human involvement in the decision, Article 22 GDPR potentially applies. Whether or not this is the case depends on whether or not the decision 'produces legal effects' for or 'similarly significantly affects' the data subject.

The WP29 lists as examples for legal effects the cancellation of a contract, the entitlement to or denial of a particular social benefit granted by law (such as child or housing benefit), and refused admission to a country or denial of citizenship.[64]

The following table illustrates how Article 22 GDPR and Article 14 AIA deal with situations where a fully automated decision produces legal effects.

Table 5: Fully automated decisions with legal effect (based on simplified assumptions)

Scenario	Article 22 GDPR		Article 14 AIA	
	Covered by scope	Admissible with appropriate safeguards	Covered by scope	Requirements may be met
Fully automated cancellation of online gaming contract (based on default with payments) without further human intervention	+	+	–	n.a.
Fully automated cancellation of contract about rental of residential premises (based on default with payments) without further human intervention	+	–/(+)	–	n.a.
Fully automated denial of social benefit without human intervention	+	–/MS	+	+
Fully automated refusal of admission to country based on biometric identity verification without human intervention	+	–/MS	+	+

[64] Article 29 WP (fn. 50), p. 21.

Scenario	Article 22 GDPR		Article 14 AIA	
	Covered by scope	Admissible with appropriate safeguards	Covered by scope	Requirements may be met
Fully automated refusal of admission to country based on remote biometric identification without human intervention	+	–/MS	+	–

All decisions producing legal effects automatically trigger the application of Article 22 GDPR, irrespective of the significance for the affected individual. However, that significance may come into play at a different level, e.g. when it comes to whether automated decisions in a contractual context are 'necessary' (e.g. this will be difficult to argue in scenario 7 as rental of residential premises is not a mass transaction). In scenarios 8 to 10 admissibility would depend on Member State law (assuming that explicit consent is not an option as it would be difficult to argue that consent was 'free'). From the perspective of Article 14 AIA everything would again depend on whether the application at hand qualifies as a high-risk system under Article 6 (2) and Annex III AIA, with the specific prohibition in Article 14 (5) concerning remote biometric identification.

For a decision that does not produce legal effects to similarly significantly affect the data subject, the decision must, according to the WP29, have the potential to significantly affect the circumstances, behaviour or choices of the individuals concerned, have a prolonged or permanent impact on the data subject, or at its most extreme, lead to the exclusion or discrimination of individuals. This means in essence, that decisions taken must be of a 'high- risk' nature. Examples given by the WP29 include decisions that affect someone's financial circumstances, such as their eligibility to credit, decisions that affect someone's access to health services, decisions that deny someone an employment opportunity or put them at a serious disadvantage, and decisions that affect someone's access to education, for example university admissions. Similarities with Annex III of the AIA are obvious, but Annex III of the AIA takes a more sectoral approach and does not put so much weight on the concrete effects of the individual decision.

The following table illustrates how Article 22 GDPR and Article 14 AIA deal with situations where a fully automated decision does not produce legal effects.

Table 6: Fully automated decisions with other than legal effect

Scenario	Article 22 GDPR		Article 14 AIA	
	Covered by scope	Admissible with appropriate safeguards	Covered by scope	Requirements may be met
Fully automated targeted advertising of the type 'other customers who bought … often also bought …'	–	n.a.	–	n.a.
Fully automated personalised pricing on a particular online marketplace (margin of +/- 2%)	–/(+)	n.a./(+)	–	n.a.
Fully automated personalised pricing across the bigger online marketplaces (margin of +/- 30%)	+	(+)	–	n.a.
Fully automated denial of credit for residential premises (based on poor credit score) without human intervention	+	–/(+)	+	?
Fully automated decision on targeted passport control based on remote biometric identification without human intervention	–	n.a.	+	– (!)

Whether or not Article 22 GDPR applies depends on the significance of the decision for the individual, with much more weight given to economic aspects than under Annex III AIA (see Scenarios 12 and 13), as Annex III AIA seems to largely ignore economic aspects and specific consumer concerns with some few exceptions, see scenario 14 (for details see below under 6). Again, attention must be drawn to the peculiar solution in Article 14 (5) according to which the confirmation by two humans is required in cases of remote biometric identification irrespective of the significance of the measure, i.e. even where this is just about targeted passport control.

4.2.3 Conclusions

The brief analysis of the 15 case scenarios has demonstrated how Article 22 GDPR (and a similar analysis could be made for Article 11 LED) interacts with Article 14 of the AIA Proposal. Generally speaking, and with the exception of the rather misguided rule in Article 14 (5), Article 14 AIA seems to take a much more nuanced approach as compared with Article 22 GDPR, in particular as Article 14 AIA also includes situations where an AI system is used only to prepare and support a human decision. However, Article 22 GDPR allows for the consideration of economic concerns, while Annex III AIA, which defines

Article 14's scope of application, currently turns much of a blind eye on economic risks and consumer interests (see above at 3.3.2).

Considering that Article 22 GDPR is anyway misplaced (above at 4.2.1.2) a provision mirroring (but possibly adapting and improving) Article 22 GDPR should be inserted in the AIA. Article 22 GDPR should, however, remain applicable for the individual right it affords. The same holds true with regard to provisions parallel (but possibly adapted and improved) to Articles 13 (2) (f), 14 (2) (g) and 15 (1) (h) GDPR. For details see the discussion on individual rights (below at 4).

5 The List of Prohibited AI Practices in Article 5(1)(a) to (c)

5.1 Analysis of the proposed prohibitions

5.1.1 Manipulation by subliminal techniques

5.1.1.1 Practices addressed by the prohibition

Article 5 (1) (a) prohibits the placing on the market, putting into service or use of an AI system that deploys subliminal techniques beyond a person's consciousness in order to materially distort a person's behaviour in a manner that causes or is likely to cause that person or another person physical or psychological harm.

Recital 16 clarifies that research for legitimate purposes in relation to such AI systems should not be stifled by the prohibition, if such research does not amount to use of the AI system in human-machine relations that exposes natural persons to harm and such research is carried out in accordance with recognised ethical standards for scientific research.

The concept of 'subliminal techniques' has already been referred to by Article 9 (1) (b) Audiovisual Media Services Directive (AVMSD).[65] Recital 16 AIA Proposal describes such techniques as 'subliminal components individuals cannot perceive'. Apart from operating beyond a person's consciousness the techniques must be applied in order to materially distort a person's behaviour. According to Recital 16, such an intention may not be

[65] Directive 2010/13/EU of the European Parliament and of the Council of 10 March 2010 on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive) (codified version), OJ L 95, 15.4.2010, p. 1–24, as last amended by Directive (EU) 2018/1808 of the European Parliament and of the Council of 14 November 2018 amending Directive 2010/13/EU on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive) in view of changing market realities, OJ L 303, 28.11.2018, p. 69–92.

presumed if the distortion of human behaviour results from factors external to the AI system which are outside of the control of the provider or the user. Also, the practice must cause or be likely to cause a person physical or psychological harm.

Illustration 12

Company V provides an online video game. It is in V's commercial interest that users keep playing the game for as much time as possible. This is why V applies a number of measures whose purpose it is to make sure that users spend close to all their free time on this game. The measures applied by V include an online marketplace for gaming equipment creating incentives for users to engage in speculative trading, and features making sure that once a user shows signs of fatigue and seems to be planning to leave the game that user receives an attractive offer to continue (such as a free premium weapon), and/or experiences a victory, causing users to play for excessive hours.

While offering an online marketplace for gaming equipment would not be beyond the users' consciousness, the AI features reacting in a very targeted way to signs of fatigue remain hidden to the users (who are made to believe they have just been 'lucky') and would therefore qualify as subliminal techniques. Given that spending excessive hours on playing online video games, or even addiction to such games , may have serious effects on an individual's health and overall well-being, this would count as a prohibited AI practice.

5.1.1.2 Not addressed: infliction of other than physical or psychological harm

As has been explained above (at 3.3.2.2), the restriction to physical or psychological harm and thus the exclusion of mere economic harm is to be understood against the backdrop of the prohibition of unfair commercial practices under the UCPD.[66]

[66] On the application of the UCPD to manipulation by AI see *Hacker*, Manipulation by Algorithms. Exploring the Triangle of Unfair Commercial Practice, Data Protection, and Privacy Law, ELJ (forthcoming)

Illustration 13

Company V in the previous illustration makes additional profits by way of in-game sales, such as the sale of premium weapons or other virtual equipment against real money. V applies a range of subliminal techniques beyond the users' consciousness in order to instigate users to buy such equipment. Given that the immediate effect of this would not be physical or psychological harm, but rather economic harm, this practice would not be prohibited by Article 5 (1) (a) AIA Proposal.

While this practice would not be captured by the prohibition of subliminal techniques under Article 9 (1) (b) AVMSD as communication in online video games is not normally covered by the scope of the AVMSD it would qualify as an aggressive commercial practice under Article 8 UCPD. According to this provision, a commercial practice shall be regarded as aggressive if, in its factual context, taking account of all its features and circumstances, by harassment, coercion, including the use of physical force, or undue influence, it significantly impairs or is likely to significantly impair the average consumer's freedom of choice or conduct with regard to the product and thereby causes him or is likely to cause him to take a transactional decision that he would not have taken otherwise.

However, due to the restriction of the prohibition in Article 5 (1) (a) AIA Proposal to the infliction of physical or psychological harm the mere infliction of economic harm is not covered by any specific prohibition under Union law where the AI practice does not qualify as a commercial practice that is covered by the scope of the UCPD. According to Article 3 (1) with Article 2 (d) the UCPD covers only 'business-to-consumer commercial practices', which are defined as any 'act, omission, course of conduct or representation, commercial communication including advertising and marketing, by a trader, directly connected with the promotion, sale or supply of a product to consumers.' Therefore, the infliction of economic harm by way of subliminal techniques to parties other than consumers is not covered by the UCPD.

Illustration 14

Company S operates a social network used by consumers as well as by businesses that includes a news feed and various advertising features. Micro and small business users are specifically targeted with news items that elaborate on high fines that were imposed on businesses of a similar branch by data protection authorities or tax authorities. Such news items often come in conjunction with advertising for expensive consulting services on data protection or tax matters. Still shocked and frightened by the news items they have just read, a number of micro and small business users are instigated to buy those services although there would really have been no plausible risk of being fined by authorities.

While this commercial practice might potentially be prohibited under national unfair competition law, there is no prohibition at Union level because the UCPD does not apply.

Likewise, the UCPD does not cover activities that are not directly connected with the promotion, sale or supply of a product, even if such activities affect consumers.

Illustration 15

Company N runs a 'free' navigation app. Users of the app have given their consent to the processing of their personal data for a range of defined purposes, in line with the GDPR. An AI system in the background prevents users from closing the app and switching off the collection of geolocation data by way of targeted messaging, either distracting the user from their apparent plans to switch off data collection or making them switch on the data collection again through targeted offers (such as granular weather forecasts) that require the processing of geolocation data.

The UCPD, according to its Recital 7, does not address commercial practices carried out primarily for purposes other than influencing consumers' transactional decisions relating to products, including for example commercial communication

aimed at investors, such as annual reports and corporate promotional literature. It is at least uncertain whether the practice applied by N would be covered by the UCPD[67] as the practice is not about users making transactional decisions with regard to whether or not to buy a service but decisions about their privacy settings.

Article 5 AIA is also needed on top of the UCPD because, by prohibiting not only the use but also the placing on the market of particular AI systems.

Illustration 16

Assume that software development company D develops the AI system using subliminal techniques as described in illustration 13 and sells the AI system to providers of online games worldwide. The mere placing on the market of such an AI system would, as such, not be captured by the UCPD and therefore not be prohibited. This is why a prohibition in Article 5 AIA is required in addition to the UCPD and possibly other law that may also prohibit the use of such AI systems in dealings with consumers.

5.1.1.3 Why the restriction to physical and psychological harm must be reconsidered

The illustrations have demonstrated that the policy of avoiding overlap between Article 5 and the UCPD comes at the price of many gaps and of a regulatory regime that looks, at least at first sight, rather arbitrary in its policy choices (e.g. appearing to neglect consumer interests).

As Article 5 cannot be properly understood without analysing it within the wider framework of existing consumer protection law there is a risk that the AIA will not be properly applied by courts and stakeholders throughout Europe.

[67] See Opinion of the Data Ethics Commission (fn. 39), p. 97.

Also, one should not forget that the AIA has the potential of becoming a global role model for the regulation of AI applications, and in order to fulfil this role, it must be easy to understand and reflect the underlying policy choices and assumptions in a consistent manner. A piece of legislation which is understood only by very few experts worldwide, because in order to understand it one has to have a very profound knowledge of the remaining acquis and the scope of application of various other legal instruments, will not easily become a legal instrument from which other regions in the world draw inspiration.

It is therefore recommended to remove the restriction to 'physical or psychological' harm in Article 5 (1) (a) and to replace it by 'material and unjustified' harm (without any further restriction).[68] This would both make sure that economic interests of consumers are duly captured and avoid overreaching application of the prohibition beyond what is intended.[69]

5.1.2 Exploitation of group-specific vulnerabilities

5.1.2.1 Practices addressed by the prohibition

Article 5 (1) (b) AIA Proposal prohibits the placing on the market, putting into service or use of an AI system that exploits any of the vulnerabilities of a specific group of persons due to their age, physical or mental disability, in order to materially distort the behaviour of a person pertaining to that group in a manner that causes or is likely to cause that person or another person physical or psychological harm. This prohibition is closely aligned with the prohibition of manipulation by subliminal techniques. In particular, the requirements of an intention to materially distort a person's behaviour and of causing physical or psychological harm (or a likely risk to cause such harm) coincide in both Article 5 (1) (a) and (b).

[68] For a similar proposal see Federation of German Consumer Organisations (vzbv), Artificial Intelligence needs real-world regulation, 5 August 2021, p. 14. For very critical remarks on scope see also *Veale/Borgesius*, Demystifying the Draft EU Artificial Intelligence Act, Computer Law Review International, 2021, p. 99; BEUC, Regulating AI to Protect the Consumer (2021), p 10 ff.

[69] Risks of overreaching application should not be underestimated, and it could be advisable to rely on authorities and courts for tackling potential abuse and circumvention. This is also why the author is sceptical towards removing any kind of reference to intentionality as suggested, e.g., by vzbv (fn. 68), p. 14, 15; see already *Veale/Borgesius* (fn. 68), p. 99 ff., who also address situations where harm is caused by third parties or the affected person.

The prohibition in (b) aims at AI systems that exploit vulnerabilities of children and people due to their age, physical or mental incapacities.

Illustration 17

Company M operating a messenger app uses an AI system with addictive and compulsive design specifically tailored for attracting younger children, such as by systematically sending them push-notifications even when the app is closed on the child's smartphone and putting them under peer-pressure to react immediately to incoming messages with a score rating the children according to how fast they react and how many messages they send. Using the app may have the effect that children no longer concentrate on school, family life and their other activities, leading to stress and anxiety and ultimately to psychological and/or physical harm.

5.1.2.2 Not addressed: infliction of other than physical or psychological harm

Again, the infliction of mere economic harm is not captured by Article 5 (1) (b) AIA Proposal, because the drafters felt that this would be covered already, in particular by the UCPD.

Illustration 18

If company M in illustration 17 uses an addictive design also for pushing children into contracts, such as by providing a separate ad and news feed, systematically offering attractive new ringtones, games or video clips at prices even younger children can normally afford with their pocket money.

While this practice is not addressed by Article 5 (1) (b) AIA Proposal it is clearly to be qualified as an aggressive commercial practices under Article 8 UCPD. 'Including in an advertisement a direct exhortation to children to buy advertised products or persuade their parents or other adults to buy advertised products for them' is even a blacklisted practice under Point 28 of Annex I UCPD.

Again, the question arises whether the provision in Article 5 (1) (b) AIA Proposal leaves too many gaps in terms of unacceptable practices that are neither covered by the AIA Proposal nor by the UCPD or other Union legislation. Such gaps could exist with regard to (in particular micro and small) business users, see illustration 14, who may also be of different age groups, different degrees of digital literacy, etc. Such gaps could equally exist with regard to other than transactional decisions, see illustration 15. And, finally, the UCPD would not as such prohibit the placing on the market of relevant technology, see illustration 16, which is why the AIA is required on top of the UCPD even where a case is within the scope of the UCPD.

5.1.2.3 Not addressed: exploitation of vulnerabilities not related to age etc

However, there is also the question whether, even within the narrow ambit of 'physical or psychological harm', the current version of Article 5 (1) (b) AIA Proposal is consistent. For instance, a group-specific economic or social situation may make individuals belonging to that group as vulnerable as the factors currently mentioned in Article (5) (1) (b) AIA Proposal.

On a more general note, it could be questioned whether the traditional 'vulnerable groups' are still a helpful concept in the digital age[70] or whether we are either all vulnerable or at least new 'vulnerable groups' have emerged that have nothing to do with age, physical or mental disabilities.

Illustration 19

Assume that Company V in illustration 12 does not use subliminal techniques for their features that make sure users do not leave the game even when over-tired. Instead, once a gamer shows signs of fatigue, V openly displays a sign saying: 'We realise you are tired. However, if you quit now, it may be more difficult for you to collect the Elo-points necessary to reach the next level. Are you sure you want to quit?' While many gamers would not be distracted by this, gamers tending towards a subjugation schema as a general emotional and behavioural pattern

[70] *Helberger/Micklitz/Sax/Strycharz*, Surveillance, consent and the vulnerable consumer. Regaining citizen agency in the information economy, in: BEUC, EU Consumer Protection 2.0. Structural asymmetries in digital consumer markets (2021), p. 3, 5 ff..

(but usually not displaying a disorder that would qualify as ‘mental disability’) are prompted to stay on the game for excessive hours, causing physical and psychological harm.
This practice would not be prohibited by Article 5 as the techniques are neither subliminal nor targeting a vulnerable group defined by age, physical or mental disability (unless a court would apply a very broad understanding of ‘disability’, basically reading it as ‘disposition’).

Going one step further, the question arises whether the exploitation of group-specific vulnerabilities is the only problematic case or whether the exploitation of very individual vulnerabilities that have been detected with the help of extensive data collection and data analytics is not at least as problematic and unacceptable.

Illustration 20

Company V in illustration 19 collects a broad range of personal data from anyone playing the online game, analysing every single of their reactions to a broad variety of variety of situations, meticulously measuring all sorts of behavioural traits, resulting in an extremely granular behavioural profile (see illustration 6). Once a gamer shows signs of fatigue, V openly displays a message that is perfectly tailored to the individual personality of the relevant gamer, addressing him or her in a manner that is optimally suited to keep him or her playing for excessive hours, causing physical and psychological harm.

It should be noted that a problem we see very often is the exploitation of very individual vulnerabilities in a way that causes economic harm, e.g. offering expensive loans to individuals in financial distress. However, these situations are indeed covered again by the UCPD. According to Article 9 (c) UCPD, in determining whether a commercial practice uses harassment, coercion, including the use of physical force, or undue influence, as an aggressive commercial practice account shall be taken of, inter alia, the exploitation by the trader of any specific misfortune or circumstance of such gravity as to impair the consumer's judgement, of which the trader is aware, to influence the consumer's decision with regard to the product.

5.1.2.4 Why the scope of the prohibition needs to be reconsidered

For very similar reasons as have been put forward above (at 5.1.1.3) with relation to the prohibition of manipulation the scope also of the prohibition of exploitation should be broadened.[71] Again, the illustrations have demonstrated that the policy of avoiding overlap between Article 5 and the UCPD comes at the price of many gaps. Again it must be noted that a regulatory regime that looks arbitrary in its policy choices and whose interplay with the UCPD is not obvious at first sight is likely not to be understood and applied properly by courts and stakeholders throughout Europe. And again, one should not forget that the AIA has the potential of becoming a global role model for the regulation of AI applications, and that it must be convincing in itself in order to fulfil this role.

It is therefore recommended also with regard to exploitation to remove the restriction to 'physical or psychological' harm also in Article 5 (1) (b) and to replace it by 'material and unjustified' harm (without any further restriction). This would both make sure that economic interests of consumers are duly captured and avoid overreaching application of the prohibition beyond what is intended.

It is also recommended to remove the restriction to the particular group-specific vulnerabilities currently addressed. In any case, exploitation of individual vulnerabilities should be added as a prohibition (which would, strictly speaking, make separate mentioning of the group-specific vulnerabilities superfluous, but the European legislator may wish to keep that prohibition for symbolic reasons). If the European legislator wants to mention group-specific vulnerabilities specifically they should be extended to groups defined by their social or economic situation.

5.1.3 Social scoring

5.1.3.1 Practices addressed by the prohibition

According to Article 5 (1) (c) it is prohibited to place on the market, putting into service or use, by or on behalf of public authorities, AI systems for the evaluation or classification of the trustworthiness of natural persons over a certain period of time based on their social behaviour or known or predicted personal or personality characteristics, where the social

[71] See vzbv (fn. 68), p. 16, 17; BEUC (fn. 68), p. 12 f.

score leads either to detrimental or unfavourable treatment of certain natural persons or whole groups thereof in social contexts which are unrelated to the contexts in which the data was originally generated or collected[72] or to detrimental or unfavourable treatment of certain natural persons or whole groups thereof that is unjustified or disproportionate to their social behaviour or its gravity. Recital 17 explains that such AI systems may lead to discriminatory outcomes and the exclusion of certain groups, violate the right to dignity and non-discrimination and the values of equality and justice.

5.1.3.2 Not addressed: social scoring by private parties

As has been demonstrated in further detail above (at 4.1.2), the restriction of the prohibition in Article 5 (1) (c) to social scoring conducted by public authorities or on their behalf can be explained by the specific interplay between the AIA and the GDPR.

Illustration 8 (gatekeeper platform service defining conditions for users according to their individual 'climate score' calculated from a range of various sources) has shown that it would be largely impossible for players from the private sector to justify social scoring by way of direct reliance on the GDPR. The only possibility for such players to engage in social scoring would be consent under Article 6 (1) (a) GDPR, but for consent to be truly 'free' a data subject may not suffer any detriment in the case of denial or withdrawal of consent where such detriment goes beyond what is strictly necessary to protect the legitimate interests of the controller. By contrast, given that Member States have a wide margin of discretion in defining objectives of public interest and legitimate aims there is a greater danger that public authorities might, based on Member State law or on a task entrusted by Member State law, engage in social scoring activities lawfully.

5.1.3.3 Why the scope of the prohibition should be broadened

Even though the restriction to social scoring conducted by public authorities or on their behalf finds a certain justification in the interplay between the AIA and the GDPR, it is still problematic that the restriction looks arbitrary at first sight and is only understandable against the background of the thorough legal analysis that requires in-depth knowledge of the legal situation in Europe. It is not helpful for a central piece of EU legislation to look

[72] For a critical assessment of the 'contexts' element, see *Veale/Borgesius* (fn. 68), p. 100.

arbitrary in its policy choices, and only at first sight, because this may lead to confusion and reduce acceptance by stakeholders throughout Europe and even more so in other regions of the world.

There seems to be no benefit in avoiding overlap between the AIA and the GDPR, in particular not when it comes to the list of prohibited AI practices. If certain AI practices are prohibited because they are seen as a violation of fundamental European values and fundamental rights there is no problem whatsoever in having this prohibition enshrined not only in one, but in two or even several legal instruments at EU level. This is why it is recommended to abolish the restriction to social scoring activities that are conducted by public authorities or on their behalf.[73]

5.2 Prohibited practices missing

5.2.1 Total or comprehensive surveillance

The list of prohibited AI practices as it currently stands in Article 5 (1) AIA Proposal is surprisingly short. While it may be neither possible nor desirable to create a very granular list that covers any objectionable AI practice, we can possibly imagine there are certain general practices that can be seen as violating fundamental European values and fundamental rights and that are so far missing on the list.

One example for this is total or very comprehensive surveillance in individuals' private or work life. One may even say that it is the fear of such total or comprehensive surveillance that is the more general consideration underlying both the prohibition of social scoring in Article 5 (1) (c) and the restriction of real-time remote biometric identification in Article 5 (1) (d) and (2)-(4).[74] In line with the prohibitions of manipulation and exploitation, the prohibition of total or comprehensive surveillance of natural persons in their private or work life should be restricted to surveillance that occurs to an extent or in a manner that causes or is likely to cause those persons material and unjustified harm (see above at 5.1.1.3 and 5.1.2.4 for arguments why this formulation is preferable to the current

[73] See EDPB-EDPS Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), n. 29; vzbv (fn. 68), p. 17 f.; BEUC (fn. 68), p. 14.
[74] *Wendehorst/Duller* (fn. 48), p. 47 ff, 72.

restriction to physical and psychological harm, even though the restriction to physical and psychological harm might be upheld for the prohibition of total or comprehensive surveillance).

Illustration 21

Company A deploys an AI system that closely monitors and evaluates every single move of their employees, including when and for how long they visit the restrooms, as well as any conversations and other forms of social contact with their fellow employees during the working day. The feeling of being constantly watched and monitored creates stress and anxiety with the employees, and several of them develop serious stress symptoms and/or leave the company without finding new employment.

This practice may or may not be incompatible with national labour and/or tort law. Depending on the relevant Member State law, the practice may or may not be incompatible with data protection law, given that Member States have a lot of leeway with regard to data protection in the employment context under Article 88 GDPR. Whether or not there exists a clear prohibition at national level there are good arguments for having a prohibition of such practices at EU level.

5.2.2 Violation of mental privacy and integrity

While the AIA Proposal certainly considers the specificities of biometric techniques and has introduced a number of specific provisions to deal with such techniques, it is surprising that the Proposal seems to turn a blind eye on more recent developments with regard to brain-computer-interfaces (BCIs). Such systems, most of which involve the use of AI, have a significant potential for infringing mental privacy and integrity.[75] This may occur, in particular, by direct or remote measurement and/or manipulation of brain data.

[75] See European Group on Ethics in Science and Technology (EGE), Opinion on the Ethical Aspects of ICT Implants in the Human Body (2005), p. 31.

Illustration 22

In addition to monitoring their employees, company A in illustration 21 obliges their employees to undergo weekly 'corporate spirit sessions' during which their brain waves are stimulated while they are being exposed to certain corporate promotional videos. This is in order to support them in developing a positive attitude towards their employer, company A, and its activities.

While this practice may again be prohibited already under Member State labour and/or general tort law and while, therefore, there may not necessarily be a gap, this is such a central point with such significant potential for harm that it might be advisable to include an extra prohibition in the list.

This does not automatically mean a ban on polygraphs and other biometric detection systems that are used for inferring a person's thoughts or intentions, but only where such systems use specific technical processing of brain data, such as brain waves. Needless to say, where such measurement or manipulation occurs for medical reasons (e.g. for the steering of exoskeletons that assist a person in moving limbs), for research purposes or otherwise in accordance with the person's (free) will it cannot be covered by a prohibition. This is why the prohibition should only apply where the use occurs against the relevant person's will or in a manner that causes or is likely to cause that person or another person material and unjustified harm, it will usually be of the psychologically or physically nature. It should be noted that 'against the will' is not identical to 'without the will'. Hence, the use of BCI for treating unconscious patients would of course not be prohibited.

5.3 Clarifications and flexibility elements missing

5.3.1 Allowing for flexible adaptation of the list of prohibited AI practices

Given the fast pace at which technology is developing it strikes as somewhat odd that there is inbuilt flexibility in most of the provisions of the AIA Proposal, but not with regard to the prohibited AI practices in Article 5. For most of the central parts of the AIA, including with regard to the definition of artificial intelligence system (Annex I), the list of AI systems covered by safety legislation and posing a high safety risk (Annex II), and the

list of other high-risk AI systems (Annex III), the European Commission may adapt the instrument to changes in the technological landscape, without having to initiate a regular legislative procedure. There seems to be no justification for 'carving in stone' (i.e. allowing for changes only in a regular legislative procedure) precisely the list of prohibited AI practices in Article 5.[76]

5.3.2 Clarifying the relationship with prohibitions following from other laws

As has been demonstrated in various places throughout this Study, those who drafted the AIA Proposal have done so with the intention to fill gaps in existing legislation, but at the same time to avoid any sort of overlap with existing legislation. While it has been suggested to broaden the scope of the prohibitions in Article 5 (1) (a) to (c) (see above at 5.1.1.3, 5.1.2.4 and 5.1.3.3) and to add further prohibitions (see above at 5.2) it is still clear that the list in Article 5 can never be exhaustive. The vast majority of prohibitions will follow from law that is not at all AI-specific and will apply to practices irrespective of whether or not AI systems are involved. This is why it is of the essence to include, in Article 5, a reference to other law, in particular data protection law, non-discrimination law, consumer protection law and competition law. While one could argue that such reference is superfluous because those other bodies of the law apply in any case and continue to apply even when the AIA is enacted, the prohibitions currently listed in Article 5 cannot be properly understood without analysing them within the wider framework of existing Union law. In order to avoid misunderstandings it is advisable to add a reference to other law for the sake of clarification.

5.4 Recommendations as to the overall regulatory technique

5.4.1 Option 1: Reducing Article 5 to the minimum and focus on a new Annex Ia

Concerning the overall regulatory technique one possibility would be to reduce Article 5 to the bare minimum, i.e. to a reference to prohibitions by other law, combined with a

[76] See also *Ebers et al*, The European Commission's Proposal for an Artificial Intelligence Act — A Critical Assessment by Members of the Robotics and AI Law Society (RAILS), Multidisciplinary Scientific Journal, 2021, p. 593.

reference to a new Annex Ia which lists AI practices not properly covered by other law or that ought to be added for the sake of consistency and in order to avoid arbitrary choices. In addition, there should be an authorisation of the European Commission to extend the list in Annex Ia by way of delegated acts where the need for such extension arises.

The benefit of this regulatory technique would be that it put more stress on the fact that AI is just one tool among many, and that what counts is the purpose for which it is used, but that using AI for an evil purpose is neither better nor worse than using other tools for an evil purpose. E.g., manipulation by subliminal techniques that cause a person physical or psychological harm as well as exploitation of group-specific vulnerabilities that cause a person such types of harm need to be prohibited and banned in Europe irrespective of whether the technical means used fulfil the definition of 'AI system'. It is just by coincidence (or rather: due to the current 'AI hype') that the only horizontal regulatory instrument currently on the table that lends itself for such a prohibition is an instrument on AI systems.

The downside of this regulatory approach is that it reduces the symbolic value of having a 'strong' list of red lines in terms of AI practices in the main text of the AIA.

If this regulatory approach were taken, Article 5 could read as follows:

TITLE II

PROHIBITED ARTIFICIAL INTELLIGENCE PRACTICES

Article 5

1. Artificial intelligence practices are prohibited if they violate any Union or national law that addresses the relevant practice as such and without regard to the use of AI. Such law includes, but is not limited to, data protection law, non-discrimination law, consumer protection law, and competition law.

2. The AI practices referred to in Annex Ia, which may or may not be fully prohibited by existing Union or national law, shall in any case be considered prohibited as incompatible with fundamental rights and European values.

3. The Commission is empowered to adopt delegated acts in accordance with Article 73 to update the list in Annex Ia by adding prohibited AI practices where the AI practices poses a significant threat to fundamental rights and European values, as enshrined, in particular, in the European Charter of Fundamental Rights.

ANNEX Ia
PROHIBITED AI PRACTICES REFERRED TO IN ARTICLE 5(2)

Prohibited AI practices pursuant to Article 5(2) are:

1. the placing on the market, putting into service or use of an AI system that deploys subliminal techniques beyond a person's consciousness in order to materially distort a person's behaviour in a manner that causes or is likely to cause that person or another person ~~physical or psychological~~ **material and unjustified** harm;

2. the placing on the market, putting into service or use of an AI system that exploits any of the vulnerabilities of

(a) a specific group of persons due to their age, physical or mental disability **or social or economic situation; or**

(b) **an individual whose vulnerabilities are characteristic of that individual's known or predicted personality or social or economic situation** ~~in order to materially distort the behaviour of a person pertaining to that group~~ in a manner that causes or is likely to cause that person or another person ~~physical or psychological~~ **material and unjustified** harm;

3. **the putting into service or use of an AI system for the comprehensive surveillance of natural persons in their private or work life to an extent or in a manner that causes or is likely to cause those persons or another person material and unjustified harm**;

4. **the placing on the market, putting into service or use of an AI system for the specific technical processing of brain data in order to read or manipulate a person's thoughts against that person's will or in a manner that causes or is likely to cause that person or another person material and unjustified harm.**

5. the placing on the market, putting into service or use of AI systems ~~by public authorities or on their behalf~~ for the evaluation or classification ~~of~~ the trustworthiness of natural persons over a certain period of time based on their social behaviour or known or predicted personal or personality characteristics, with the social score leading to either or both of the following:

(a) detrimental or unfavourable treatment of certain natural persons or whole groups thereof in social contexts which are unrelated to the contexts in which the data was originally generated or collected;

(**b**) detrimental or unfavourable treatment of certain natural persons or whole groups thereof that is unjustified or disproportionate to their social behaviour or its gravity;

[…]

5.4.2 Option 2: Combination of list, reference to other law, and flexibility clause

In order to maintain the strong symbolic value of having a list of clearly formulated red lines in the main text of the AIA it might be preferable to go for a combination of the list of prohibitions in the main text (with the individual prohibitions broadened in scope as described above and possibly further prohibitions added), a reference to other law, and a flexibility clause that allows the European Commission to extend the list in the new Annex Ia by way of delegated acts.

If this regulatory approach is chosen, Article 5 could read as follows:

TITLE II

PROHIBITED ARTIFICIAL INTELLIGENCE PRACTICES

Article 5

1. The following artificial intelligence practices shall be prohibited:

(a) the placing on the market, putting into service or use of an AI system that deploys subliminal techniques beyond a person's consciousness in order to materially distort a person's behaviour in a manner that causes or is likely to cause that person or another person ~~physical or psychological~~ **material and unjustified** harm;

(b) the placing on the market, putting into service or use of an AI system that exploits any of the vulnerabilities of

(i) a specific group of persons due to their age, physical or mental disability **or social or economic situation; or**

(ii) an individual whose vulnerabilities are characteristic of that individual's known or predicted personality or social or economic situation

~~in order to materially distort the behaviour of a person pertaining to that group~~ in a manner that causes or is likely to cause that person or another person ~~physical or psychological~~ **material and unjustified** harm;

(ba) the putting into service or use of an AI system for the comprehensive surveillance of natural persons in their private or work life to an extent or in a manner that causes or is likely to cause those persons or another person material and unjustified harm;

(bb) the placing on the market, putting into service or use of an AI system for the specific technical processing of brain data in order to read or manipulate a person's thoughts against that person's will or in a manner that

causes or is likely to cause that person or another person material and unjustified harm.

(c) the placing on the market, putting into service or use of AI systems ~~by public authorities or on their behalf~~ for the evaluation or classification ~~of~~ the trustworthiness of natural persons over a certain period of time based on their social behaviour or known or predicted personal or personality characteristics, with the social score leading to either or both of the following:

(i) detrimental or unfavourable treatment of certain natural persons or whole groups thereof in social contexts which are unrelated to the contexts in which the data was originally generated or collected;

(ii) detrimental or unfavourable treatment of certain natural persons or whole groups thereof that is unjustified or disproportionate to their social behaviour or its gravity;

[…]

1a. In addition to the prohibited AI practices referred to in paragraph (1), AI practices referred to in Annex Ia shall also be considered prohibited. The Commission is empowered to adopt delegated acts in accordance with Article 73 to update the list in Annex Ia on the basis of a similar threat to fundamental rights and European values as posed by the practices listed in paragraph (1).

1b. Paragraphs (1) and (1a) are without prejudice to prohibitions that apply where an artificial intelligence practice violates other laws, including data protection law, non-discrimination law, consumer protection law, and competition law.

6 Biometric Techniques as 'Restricted' AI Practices

In another study commissioned by the JURI and PETI Committees of the European Parliament the author of this Study has provided an in-depth analysis of biometric techniques under the AIA Proposal.[77] The following findings include a summary of selected aspects of that study.

6.1 Differentiating between per se-prohibitions and restrictions

The rules on real-time remote biometric identification seem to be an alien element within Title II, which is about 'prohibited AI practices'. While Article 5 (1) (a) to (c) address AI practices that are clearly incompatible with European values and that should therefore be prohibited under all circumstances, Article 5 (1) (d) and (2) to (4) on real-time remote biometric identification is not about a per se prohibition. Rather, those provisions contain a number of significant restrictions as well as conditions under which the use of real-time remote biometric identification systems for law enforcement purposes in publicly accessible spaces should be allowed.[78] It is therefore suggested to remove paragraphs (1) (d) and (2) to (4) from Article 5 and to include them in a new and separate Title IIa which should be devoted to 'restricted artificial intelligence practices'.

[77] *Wendehorst/Duller* (fn. 48).

[78] It is, of course, possible to say that any kind of mandatory restriction or mandatory requirement amounts to a prohibition of practices that are not in compliance with restrictions or requirements. However, this is true for most regulatory regimes, and it is clearly not the spirit in which Article 5(1)(a) to (c) have been formulated.

6.2 Biometric identification

6.2.1 Limitations of the prohibition/restriction

6.2.1.1 Limitation to identification (as contrasted with authentication/verification)

Biometric identification is a method of identifying or confirming a person's identity based on the individual's unique physical, physiological or behavioural characteristics. In the narrower sense of the term, 'identification' is to be distinguished from 'authentication'. Authentication is a 'one-to-one' comparison, matching the live template of a particular person who claims to have a particular identity with the stored template in a template database that is connected with that identity in order to verify whether the claim is true. By contrast, identification in the narrower sense is defined as a 'one-to-many' comparison where the persons identified do not claim to have a particular identity but where that identity is otherwise established – often without the conscious cooperation of these persons or even against their will – by matching live templates with templates stored in a template database. The AIA Proposal only includes provisions on identification in the narrow sense. This is expressed in the definition in Article 3 (36), which restricts the relevant definition to cases of identification 'without prior knowledge of the user of the AI system whether the person will be present and can be identified'. In other words, biometric border controls, where the person presenting a biometric passport and claiming to have a particular identity, would not be covered at all, nor would the many biometric authenticaion methods used for securing premises and devices against unauthorised access. This limitation seems to be justified because the ethical issues raised by biometric identification are much more serious than ethical issues raised in cases where a person already claims him- or herself to have a particular identity.

6.2.1.2 Limitation to biometric identification

As far as the limitation of the scope of (current) Article 5 (1) (d) and (2) to (4) of the AIA Proposal to biometric identification techniques is concerned, the restriction is not fully convincing from an ethical point of view. What counts most in terms of fundamental rights concerns is the fact that someone (e.g. law enforcement authorities) holds a biometric template of a particular person and is thus able to identify and trace that person anywhere on the globe. Where this is the case, it is only of secondary importance whether large-scale remote identification actually occurs by using the biometric templates or by

some other means, such as by tracking people's mobiles. In other words, what is ethically problematic is (a) storing people's biometric templates in a way that potentially allows to trace those people, and (b) mass surveillance of people, whereas, if both (a) and (b) is fulfilled, the fact that mass surveillance occurs by biometric means is not the decisive point.

However, there may nevertheless be good reasons for limiting the provision to identification by biometric means. First of all, it is self-evident that the more biometric techniques are used, the more encouragement there will be for the storing and refining of biometric templates. Limiting the use of biometric identification techniques may thus indirectly discourage the investment in biometric templates of the whole population. The author also realises that there is a lot of public anxiety about biometric techniques and that, from a political point of view, it may be advisable to introduce a rule specifically on biometric identification. Still, the author would like to suggest considering whether the relevant provision in the AIA could include other forms of real-time remote identification (such as by tracking mobile phone signals) while still stressing biometric identification techniques in a prominent way.

6.2.1.3 Limitation to real-time identification

One of the most striking points about the prohibition in Article 5 (1) (d) is its limitation to 'real-time' biometric identification as contrasted with 'post' biometric identification. Article 3 (37) defines a 'real-time' remote biometric identification system as a remote biometric identification system whereby the capturing of biometric data, the comparison and the identification all occur without a significant delay. The definition further clarifies that this comprises not only instant identification, but also limited short delays in order to avoid circumvention. However, whether or not there is a delay, and the length of that delay, can hardly be the decisive factor when it comes to the extent to which an identification technique should be restricted.[79]

One can easily conceive a range of situations where biometric identification occurs with a significant delay, but still there is a significant threat to fundamental rights.

[79] vzbv (fn. 68), p. 18.

Illustration 23

Video surveillance is in action at various points on High Street, which is a popular place for large demonstrations to be held, including demonstrations against the government. When such a demonstration takes place, video recordings are produced and stored that capture all individuals that have taken part in the demonstration. For logistical reasons, the video material is usually analysed, and all the individuals identified by way of biometric identification techniques, only one or two days after the recording has taken place. As Article 5 (1) (d) and the definition of 'real-time' in Article 3 (37) currently stand, this practice would not be covered even though, evidently, it poses a significant threat to fundamental rights.

Conversely, there are certain forms of biometric identification where the time lag between collection of live templates and matching is minimal but where biometric identification that occurs only on a very small scale and only when triggered by a particular incident. Such forms of biometric identification pose a significantly lower risk to fundamental rights and need not be restricted to the same extent by the AIA.

Illustration 24

Video surveillance is in action at various points on High Street. Video material is stored for 24 hours and then deleted. It is streamed in real time to police headquarters, where policemen can watch the scenes on High Street if required, and analysed in real time by an AI system trained to recognise incidents (such as violence or accidents) that require police action. In particular where a crime has been committed, the police would analyse the video material with the help of biometric identification techniques, checking whether the offender's live template matches with any template in an existing database. This should not qualify as 'real-time' remote identification even where the delay between recognition of the incident and biometric identification of the offender is minimal.

What seems to be the decisive factor for the degree to which biometric identification techniques should be restricted by the AIA is whether surveillance by means of biometric identification occurs on a continuous basis or otherwise on a large scale over a period of time and without focus on a particular past incident. The definition of 'real-time' in Article 3 (37) should therefore be amended.

If the definition of 'real-time' is changed it could be advisable to also amend the definition of 'remote'. The drafters of this definition obviously found the absence of prior knowledge whether the person identified will be present or not to be the decisive factor that makes identification techniques different from authentication or verification techniques. However, it seems questionable whether this is in fact the decisive factor. Arguably, one can speak about identification (i.e. a 'one to many' matching exercise) also where the person using the AI system knows that a particular person identified will be present. It should therefore be considered to focus more on the fact whether or not the persons to be identified consciously cooperate for authentication purposes, e.g. by putting their thumb onto the fingerprint scanner at the entrance of a room.

To summarise, it is suggested to modify definitions as follows:

Article 3
Definitions

For the purpose of this Regulation, the following definitions apply:

[…]

(36) 'remote biometric identification system' means an AI system for the purpose of identifying natural persons at a distance through the comparison of a person's biometric data with the biometric data contained in a reference database, and without **the conscious cooperation of the persons to be identified** ~~prior knowledge of the user of the AI system whether the person will be present and can be identified~~ ;

(37) ''real-time' remote biometric identification system' means a remote biometric identification system whereby the capturing of biometric data, the comparison and the identification ~~all~~ occur **on a continuous or large-scale basis over a period of time and without limitation to a particular past incident (such as a crime recorded by a video camera);** ~~without a significant delay. This comprises not only instant identification, but also limited short delays in order to avoid circumvention~~.

While there may be a justification for having stricter rules for more biometric identification techniques that qualify as 'real-time' within the meaning of a definition that

has been amended along the lines of the proposal above, this is not to say that there should be no restrictions at all when it comes to biometric techniques that qualify as 'post' biometric identification. It has been demonstrated above (at 4.1.3) that, where biometric identification is applied, Article 9 GDPR and Article 10 LED provide for a rather high level of protection. However, as in other cases, the interplay between the AIA Proposal and the GDPR does not become sufficiently clear. In order to produce a legal instrument that is better understandable for those who have to apply it and it looks less arbitrary in its policy choices at first sight it is highly advisable to clarify the relationship with Article 9 GDPR and Article 10 LED as well as integrate 'post' biometric identification in the legislative framework of the AIA (for a proposal how this could be implemented, see below at 6.3.2.3).

6.2.1.4 Limitation to publicly accessible spaces

Very similar remarks as have just been made with regard to 'post' biometric identification can be made with regard to remote biometric identification that occurs in places other than publicly accessible spaces. There seem to be good reasons for having stricter provisions where remote biometric identification occurs in spaces that are accessible to the public. Where spaces are not publicly accessible, but accessible only to a limited number of persons (such as the employees of a company, or the inmates of a prison) fewer people are affected and the use of biometric identification techniques is much closer to biometric authentication. Where spaces are accessible to an indefinite number of people, but the spaces are not physical (but online) spaces, Article 9 GDPR defines rather strict limits for the use of biometric data in the narrow sense, so the level of protection is already quite high. Usually, the only possible justification will be explicit consent within the meaning of Article 9 (2) (a) GDPR. However, for reasons of consistency it seems advisable to clarify the relationship with the GDPR and to integrate such use of remote biometric identification in the legislative framework of the AIA (for a proposal how this could be implemented again below at 6.3.2.3).

6.2.1.5 Limitation to law enforcement purposes

Another striking limitation of Article 5 (1) (d) AIA Proposal is the limitation to law enforcement purposes. Generally speaking, law enforcement should be a privileged purpose when compared with other purposes (such as data collection for public planning activities). Article 10 LED allows the processing of biometric data for purposes of identification where this is strictly necessary for law enforcement purposes, subject to

appropriate safeguards for the rights and freedoms of the data subject, and only where further requirements are met, such as that the processing of biometric data is authorised by Union or Member State law. Under Article 9 (2) (g) GDPR, national authorities may process biometric data, including for identification purposes, where the processing of biometric data is necessary for reasons of substantial public interest, on the basis of Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject. There is thus not much difference between the leeway for public authorities under the LED on the one hand and the GDPR on the other. So if there is a danger of excessive use of real-time remote biometric identification by law enforcement authorities, the same danger exists with regard to public authorities that process biometric data on the basis of the GDPR for purposes other than law enforcement.

Illustration 25

Municipality M would like to get a better idea of who lives in the city or visits the city, what are the citizens' habits, and how they move around during the day. The data is to be used for planning purposes, e.g. for improvement of the public transport system, or for the management of crowds and assemblies in public spaces. On the basis of national law, which includes details as to pseudonymisation and other safeguards, M uses real-time remote biometric identification for collecting the data. This practice may be considered disproportionate to the aim pursued and thus not to be in conformity with Article 9 GDPR. However, also disproportionate use of biometric techniques for law enforcement purposes would theoretically also be prohibited by the LED. It is therefore difficult to understand why the one is dealt with under the AIA Proposal, but not the other.

While there is no comparable urgency to also regulate private use of remote biometric identification in publicly accessible spaces (as this is more likely to be fully captured by Article 9 GDPR) there is no harm in including it in the restriction, provided the provision is formulated in a way that is fully consistent with Article 9.

6.2.2 A new regulatory approach

What is recommended for the treatment of biometric techniques within the AIA is an entirely new regulatory approach.[80] This new regulatory approach would no longer list the use of biometric techniques among the per se prohibition is in Article 5 but rather move the existing provisions on real-time remote biometric identification to a new Title IIa on 'Restricted AI Practices' and a separate Article Sa. This new Article 5a should be phrased in a way that is as closely aligned as possible with, in particular, Article 9 GDPR. This could be achieved by listing, in the new Article 5a AIA, the legitimate purposes for which real-time remote biometric identification is permissible, following as far as possible the structure and wording of Article 9 (2) GDPR. Law enforcement would then rightly be treated as a privileged purpose, alongside qualified consent, the use for scientific research purposes, the use for the protection of the vital interests of the person identified, and use for migration, asylum or border control management.

Whether or not additional specifications, beyond the restrictions that already follow from the current text, should be added to the use for migration, asylum or border control management, is a political question. In any case, and assuming that migration, asylum or border control management do not generally/always qualify as 'law enforcement' (see also the division in Annex III), there is currently no restriction at all in the AIA Proposal.

It is recommended that the new Article 5a on real-time remote biometric identification (or, as the case may be, also other forms of real-time remote identification) clarify that other areas of the law apply to fill the gaps, in particular data protection law and non-discrimination law.

In addition, the new paragraph should stress explicitly the controller's duty not to collect any data beyond what is strictly necessary to achieve the purpose on which biometric identification is based, and to erase any personal data collected from biometric identification as soon as these data are no longer strictly necessary to achieve the purpose for which the data have been collected.

---

[80] See, however, EDPB-EDPS Joint Opinion 5/2021 (fn. 73) n. 32 and BEUC (fn. 68), p. 15 calling even for a complete ban of biometric identification in publicly accessible spaces; for a similar opinion see vzbv (fn. 68), p. 17 ff.; in Resolution 2020/2016(INI) the European Parliament calls for a ban of biometric techniques in public places, but wants to allow the use facial recognition for the identification of victims of crime. The use of facial recognition for other law enforcement purposes shall only be allowed if an appropriate legal framework is in place.

Illustration 26

Police has received information from intelligence services that individuals X and Y are planning a terrorist attack on a Christmas market. This is why real-time remote biometric identification is used to trace and stop any of X or Y in case one of them were to be seen in the city or even in the vicinity of the market. In a situation such as this, use of real-time remote biometric identification would be justified. However, it is only necessary to compare the live templates of people walking by (such as that of innocent bystander B) with the stored templates of X and Y. By contrast, it would not be permissible to fully identify B by way of comparison with any stored template of B because this is not necessary for achieving the purpose. In the given situation it may be justified to store video recordings for a longer period than usual (as there may be situations where, ex post, it turns out that previously unknown person Z was cooperating with X and Y and exploring the area to prepare for an attack). However, unless such a situation arises later and B was acting in a suspicious manner, it would not be necessary to identify B.

The newly structured Article 5a might read as follows:

TITLE IIA

RESTRICTED ARTIFICIAL INTELLIGENCE PRACTICES

CHAPTER 1

BIOMETRIC TECHNIQUES

Article 5a
'Real-time' remote biometric [Opt.: or other] identification

1. AI systems may be used for 'real time' remote biometric identification *[Opt.: or other 'real time' remote identification]* in publicly accessible spaces only when such surveillance is limited to what is strictly necessary for:

(a) the use for a specific purpose to which the persons identified have given their explicit consent within the meaning of Article 9 (2)(a) of Regulation (EU) 2016/679;

(b) the use for purposes and under conditions referred to in Article 9 (2)(b) and (j) of Regulation (EU) 2016/679;

(c) the use for migration, asylum or border control management;

(d) the use ~~of 'real-time' remote biometric identification systems in publicly accessible spaces~~ for the purpose of law enforcement, ~~unless and~~ in as far as such use is strictly necessary for one of the following objectives:

(i) the targeted search for specific potential victims of crime, including missing children;

(ii) the prevention of a specific, substantial and imminent threat **to public security, in particular** to the life or physical safety of natural persons**,** or of a terrorist attack;

(iii) the detection, localisation, identification or prosecution of a perpetrator or suspect of a criminal offence referred to in Article 2(2) of Council Framework Decision 2002/584/JHA and punishable in the Member State concerned by a custodial sentence or a detention order for a maximum period of at least three years, as determined by the law of that Member State.

2. The use of 'real-time' remote [biometric] identification systems in publicly accessible spaces for the purpose~~s of law enforcement for any of the objectives~~ referred to in paragraph 1 point**s c) and** d) shall take into account the following elements:

(a) the nature of the situation giving rise to the possible use, in particular the seriousness, probability and scale of the harm caused in the absence of the use of the system;

(b) the consequences of the use of the system for the rights and freedoms of all persons concerned, in particular the seriousness, probability and scale of those consequences.

In addition, the use of 'real-time' remote [biometric] identification systems in publicly accessible spaces ~~for the purpose of law enforcement~~ for any of the objectives referred to in paragraph 1 point**s c) and** d) shall comply with necessary and proportionate safeguards and conditions in relation to the use, in particular as regards the temporal, geographic and personal limitations.

3. As regards paragraphs 1, point**s c) and** d) and 2, each individual use ~~for the purpose of law enforcement~~ of a 'real-time' remote [biometric] identification system in publicly accessible spaces shall be subject to a prior authorisation granted by a judicial authority or by an independent administrative authority of the Member State in which the use is to take place, issued upon a reasoned request and in accordance with the detailed rules of national law referred to in paragraph 4. However, in a duly justified situation of urgency, the use of the system may be

commenced without an authorisation and the authorisation may be requested only during or after the use.

The competent judicial or administrative authority shall only grant the authorisation where it is satisfied, based on objective evidence or clear indications presented to it, that the use of the 'real-time' remote biometric identification system at issue is necessary for and proportionate to achieving one of the objectives specified in paragraph 1, point**s (c) and** (d), as identified in the request. In deciding on the request, the competent judicial or administrative authority shall take into account the elements referred to in paragraph 2.

4. A Member State may decide to provide for the possibility to fully or partially authorise the use of 'real-time' remote [biometric] identification systems in publicly accessible spaces ~~for the purpose of law enforcement~~ within the limits and under the conditions listed in paragraphs 1, point**s (c) and** (d), 2 and 3. That Member State shall lay down in its national law the necessary detailed rules for the request, issuance and exercise of, as well as supervision relating to, the authorisations referred to in paragraph 3. Those rules shall also specify in respect of which of the objectives listed in paragraph 1, point**s (c) and** (d), including which of the criminal offences referred to in point **(d)** (iii) thereof, the competent authorities may be authorised to use those systems for the purpose of law enforcement.

5. **Further requirements or restrictions following from other Titles of this Act or from other laws, in particular data protection law and non-discrimination law, remain unaffected. In any case, only such personal data may be collected through remote biometric identification as are strictly necessary to achieve the purpose stated in paragraph (1), and must be erased as soon as they are no longer necessary in relation to this purpose.**

6.3 Emotion recognition and biometric categorisation

6.3.1 Biometric data and biometrics-based data

The definitions of 'biometric identification system', 'biometric categorisation system' and 'emotion recognition system' all build on the definition of 'biometric data'. This definition, in turn, has been copied from Article 4 (14) GDPR and is defined as 'personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of the natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic data'. Recital 51 of the GDPR clarifies that, e.g., the processing of photographs should not systematically be considered to be processing of special categories of personal data as they are covered

by the definition of biometric data only when processed through a specific technical means allowing the unique identification or authentication of a natural person.

While it is certainly essential to stress the requirement of specific technical processing (for, otherwise, almost any everyday activity might potentially be covered), the requirement that the data must allow or confirm the unique identification of a natural person makes the definition far too narrow. It essentially reflects the dominant concepts during times of 'first generation biometric technologies' and fails to keep pace with technological developments.[81] Recently, 'second generation biometric technologies' such as voice, keystroke or gait patterns, and 'soft biometrics' such as facial expressions, movements or body shape, as well as multi-modal biometric techniques has been gaining ground, opening up vast new fields of application, including in the consumer context.[82]

As the definitions of 'biometric categorisation system and 'emotion recognition system' currently stand they would be restricted to techniques that are based on data that would allow or confirm the unique identification of the natural person concerned. By way of contrast, an emotion recognition system based on pulse frequency, body temperature and non-unique facial expressions (such as smiling, raising of the brow or yawning) or non-unique voice signals (such as volume or trembling) would not be covered by the definition of emotion recognition system under the AIA. This is so simply because the data used would not qualify as biometric data in the narrow sense.

It is therefore suggested to include a new definition of 'biometrics-based data'. This definition would largely coincide with the definition of biometric data, but would differ from that definition in that biometrics-based data may or may not allow or confirm the identification of a natural person. It is important to stress that there would still be the requirement of specific technical processing, i.e. a video showing a person who is smiling would not amount to biometrics-based data, but the use of specific analytic tools that tell a smiling person from a person in a different mood would qualify as biometrics-based data.

[81] For details see *Wendehorst/Duller* (fn. 48), p. 12 ff. with further references.
[82] *Peissl/Schaber/Strauß*, Der Körper als Schlüssel? Biometrische Methoden für Konsument:innen, Kammer für Arbeiter und Angestellte für Wien/Institut für Technikfolgenabschätzung (2020).

Illustration 27

When customers call the helpline of company H they are prompted to state the reason why they are calling. Their oral statement is analysed by an AI system that (a) is a natural language processing (NLP) system analysing the content of the statement, such as whether the customer has a question or is complaining, or (b) analyses the customer's voice with regard to pitch, volume, trembling, accent etc. in order to find out about the customer's background and emotions, both (a) and (b) with the aim of allowing a chat-bot to react in a very targeted way. While data used by the AI system in (b) are biometrics-based data, the data used by the AI system in (a) are not.

The Illustration also shows that the fact an AI system uses biometrics-based data says little about the purpose and 'level of criticality' of that data use, i.e. the same or very similar effects as can be achieved with the help of biometrics-based data can often be achieved with the help of other data. However, it is still advisable to create provisions specifically for AI systems using biometrics-based data as otherwise the scope of provisions would become very fuzzy and too much uncertainty would be created.[83] Also, use of biometrics-based data raises very specific and additional ethical concerns, due to the fact that a person cannot easily change such data that justify stricter regulation.[84]

To summarise, it is suggested to phrase the definition as follows:

Article 3
Definitions

For the purpose of this Regulation, the following definitions apply:

[…]

(33) 'biometric data' means personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic data;

[83] Note that vzbv (fn. 68), p. 9 recommends using the term 'personal data' instead. For the reasons stated above this seems not to be advisable as a solution.

[84] On the ethical concerns regarding the so-called 'datafication' of humans, see Alterman, A Piece of Yourself: Ethical Issues in Biometric Identification, Ethics and Information Technology (2003), p. 139.

(33a) 'biometrics-based data' means personal data resulting from specific technical processing relating to physical, physiological or behavioural signals or characteristics of a natural person, such as facial expressions, movements, pulse frequency, voice, keystrokes or gait, which may or may not allow or confirm the unique identification of a natural person;

In line with this new definition, also the definitions of 'emotion recognition' and 'biometric categorisation' systems should be amended. If the definitions are amended, a couple of further (as such less important) editorial changes could be made.

This could mean changing definitions as follows:

Article 3
Definitions

For the purpose of this Regulation, the following definitions apply:

[…]

(34) 'emotion recognition system' means an AI system for the purpose of identifying or inferring emotions**, thoughts** or intentions of natural persons on the basis of their ~~biometric~~**biometrics-based** data;

(35) 'biometric categorisation system' means an AI system for the purpose of assigning natural persons to specific categories such as sex, age, hair colour, eye colour, tattoos, ethnic origin**, health, mental ability, behavioural traits** or sexual or political orientation, on the basis of their ~~biometric~~**biometrics-based** data;

6.3.2 Emotion recognition and biometric categorisation as restricted AI practices

6.3.2.1 Absence of a 'safety net' in Article 9 GDPR

While, in the case of biometric identification systems, there was still Article 9 GDPR and Article 10 LED as a kind of 'safety net' because biometric identification relies on the processing of biometric data within the definition of the GDPR and LED, this is not the case with emotion recognition and biometric categorisation systems. The reason is that those systems do not rely on the processing of biometric data within the meaning of Article 9 GDPR, but only on what has here been called 'biometrics-based data' that may or may not allow or confirm the unique identification of a natural person (see above at 6.3.1). This is why, in the majority of cases, processing of personal data for purposes of emotion

recognition or biometric categorisation will only be subject to Article 6 GDPR, including simple consent within the meaning of Article 6 (1) (a) and other legal grounds available for personal data in general. Of course, it seems hardly convincing that Article 9 GDPR qualifies personal data revealing political opinions, religious beliefs or trade union membership as particularly sensitive categories of data, while emotions, thoughts and intentions (e.g. identified by way of brain-computer-interfaces) only qualify as general personal data. However, unless the GDPR is changed in that respect (which would create other problems), we have to accept this unsatisfactory situation. This means that any protection which fails to be provided by Article 9 GDPR must be provided by the AIA itself.

6.3.2.2 Is the 'Article 9 GDPR regime' too restrictive?

This raises the question whether the protective regime to be established by the AIA should mirror the protective regime established by the GDPR for biometric data in the narrow sense or whether this would be overreaching.[85] The first concern about overreaching effects could arise with regard to robots that require biometric categorisation or emotion recognition in the broader sense for safety reasons.

Illustration 28

A big cleaning robot to be used in public spaces monitors and analyses the behavioural characteristics of people walking by. Where the AI system indicates, for instance, the presence of a child, an elderly person or a person with physical or mental disabilities, the robot would automatically stop or otherwise adapt its movements to that situation in order to take maximum precaution against personal injury. Clearly, it is desirable that such systems are in put place, and it is impossible to seek the affected persons' consent for such use.

In this case, it may even be possible to restrict processing to anonymous data and thus remain outside the definition of 'emotion recognition system' as well as outside the scope of the GDPR.[86] If the GDPR applies, it normally (in the absence

[85] See, however, EDPB-EDPS Joint Opinion 5/2021 (fn. 73) n. 33, 35, calling even for a complete ban of these techniques; for a similar opinion see vzbv (fn. 68), p. 19, however considering some exceptions on p. 20; BEUC (fn. 68), p. 17 follows the recommendation of the EDPB-EDPS.

[86] Cf. a ruling by the German Federal Administrative Court (BVerwG) on the automated registration of motor vehicle number plates, according to which data collection is irrelevant in terms of data protection law if the data are fully erased or anonymised within a system immediately after collection without the possibility of

of processing of biometric data in the narrow sense of the GDPR) requires only a legal ground under Article 6. However, data processing would be justified even under a stricter regime designed along the lines of Article 9 because processing is necessary to protect the vital interests of the relevant individuals. Thus, application of the stricter regime would not cause any undesired effects.

A concern about overreaching effects could also arise with regard to AI systems that are not established to protect 'vital' interests but still other important interests (e.g. economic interests, protection of minors) of the affected individuals.

Illustration 29

Online marketplace M applies biometric categorisation systems as well as emotion recognition systems for age verification and fraud prevention purposes. For instance, where a customer who appears to be a minor attempts to access adult content or order items sale of which is restricted to adults the system would automatically deny conclusion of a contract. Likewise, an AI system analysing keystroke patterns triggers specific fraud prevention measures where the keystroke pattern of the individual placing an order deviates in a conspicuous manner from the usual keystroke pattern of the user under whose name the individual is acting.

Under the GDPR, in the absence of processing biometric data within the narrow sense of the GDPR, such measures are justified by Article 6 (1) (f). If such biometric techniques were, under the AIA, subjected to a stricter regime designed along the lines of Article 9 GDPR, this would mean a significant difference for M as M would have to seek the data subjects' explicit consent for applying these biometric techniques, or there would have to be Member State law that defends a substantial public interest and is proportionate to the aim pursued. Explicit consent could, however, easily be solicited when the user opens the account (and

---

establishing a personal link, see BverwG, Judgment of 22 October 2014 - 6 C 7/13 [ECLI:EN:BverwG:2014:221014U6C7.13.0], with reference to case law of the Federal Constitutional Court (BverfG).

a fraudulent third party claiming to be the user would have to accept being treated as the real user would be).

However, there are clearly situations where soliciting explicit consent would be cumbersome and negatively affect user experience and might seem a disproportionate burden in the light of the fact that the application cannot possibly have a negative impact on the affected natural person's legitimate interests and where personal data is erased or anonymised instantaneously without leaving any trace to the identifiable natural person (which may potentially even mean that data processed does not qualify as 'personal data' at all and that applications are therefore outside of both the scope of the GDPR and of the definitions of 'emotion recognition' or 'biometric categorisation' under the AIA Proposal).

Illustration 30

A voice-controlled Q&A chat bot on the corporate website of company C uses emotion recognition (analysing voice features) to find out in which mood users are approaching C (e.g. whether they are angry or curious) in order to adapt to this mood and find the most appropriate language (see also illustration 27). While the system processes biometrics-based data these data are only used for choosing the right language and once the language has been chosen by the chat bot the data are erased immediately without leaving any trace whatsoever on the system. While it would be theoretically possible to request the user's explicit consent this might seem disproportionate.

An analysis of a number of scenarios has thus demonstrated that subjecting the use of a broader range of biometric techniques to a regulatory regime that is more modelled on Article 9 GDPR than on Article 6 GDPR does not meet with serious objections, provided there is a de minimis exception for trivial cases. In the light of the enhanced risks for fundamental rights that come with the use of biometric techniques and their enhanced sensitivity in ethical terms it is therefore recommended to deal with these techniques in a similar way as with biometric identification (see above at 6.2.2).

An additional argument in this context is the difficulty to draw a clear line between biometric identification and biometric categorisation, which also speaks in favour of a more or less uniform regime.

Illustration 31

Department store D makes ample use of biometric techniques. For instance, when a customer enters the front gate the customer's face is scanned and an AI system immediately compares the live template with templates in a reference database consisting of facial data from customers who have visited the store during the past 24 months. In the event of a match, an AI system analyses the shopping history of this customer. If the shopping history indicates a 'VIP customer', such as a customer who is prepared to buy high-end luxury goods, or large quantities of items, a human shop assistant is instructed to approach this customer and accompany him or her, leading them in a targeted way to particular departments and offers (according to detailed instructions communicated to the shop-assistant in real-time). In addition, a behavioural detection system is in place that indicates movements and other behavioural traits resembling those of VIP customers and indicating a potential intention to buy luxury goods or larger quantities, in order to send human shop assistants also to those 'predicted VIP customers'. In this case it may be not entirely clear which components of the system qualify as biometric 'identification', which as 'biometric categorisation' and which as 'emotion recognition'. If a more or less uniform regime is in place for all of them, this greatly reduces uncertainty and facilitates both compliance and enforcement.

At the end of the day, it is therefore recommended to qualify the use of emotion recognition and biometric categorisation systems (as well as biometric identification that does not qualify as 'real-time' and 'remote') as restricted AI practices, subjecting them to a very similar regulatory approach as real-time remote biometric identification (but without the further restrictions currently found in Article 5 (2) to (4) AIA Proposal).

6.3.2.3 How to design the restrictions

If the regulatory approach is to be modelled on Article 9 GDPR it is still difficult to decide which of the purposes in Article 9 GDPR to include and which to exclude. Definitely, the situations in which the use of emotion recognition systems or biometric categorisation systems is justified will have to be more broadly defined than with regard to 'real-time' remote biometric identification as they include also the full range of medical purposes and many further purposes listed in Article 9 GDPR. While explicit consent as well as purposes such as medical purposes or scientific research purposes must clearly be listed as

admissible, and whereas some are clearly not applicable from the outset, things are less clear, e.g., with 'processing that is necessary for the establishment, exercise or defence of legal claims or whenever courts are acting in their judicial capacity' (Article 9 (2) (f) GDPR). From an ethical point of view such use would raise a number of issues.

Illustration 32

The judiciary of a Member State introduces the use of emotion recognition systems in order to find out whether persons in the courtroom (defendant, witnesses, etc.) are telling the truth. As the AIA Proposal currently stands, this would arguably be qualified as a high-risk application under Point 8 (a) of Annex III, but would otherwise be permissible if based on Member State law. It is highly questionable whether this is the right policy choice.

For the sake of simplicity and clarity of drafting, the transparency provisions, which are currently found in Article 52(2) AIA Proposal, should be inserted in the new Article 5b.

There should likewise be a reminder that Article 5b is without prejudice to further restrictions following from other laws, in particular data protection law.

To summarise, the new Article 5b could be phrased as follows:

Article 5b
Other use of biometric techniques

1. Biometric identification systems not covered by Article 5a, emotion recognition systems and biometric categorisation systems may be used only when such use is limited to what is strictly necessary for:

(a) the use for a specific purpose to which the affected persons have given their explicit consent within the meaning of Article 9 (2)(a) of Regulation (EU) 2016/679;

(b) the use for purposes and under conditions referred to in Article 9 (2)(b), (c), (g), (h), (i) and (j) of Regulation (EU) 2016/679;

(c) the use for the purpose of law enforcement, migration, asylum or border control management in as far as purposes are proportionate to the aim pursued, respect the essence of the fundamental rights and interests affected and provide for suitable and specific measures to safeguard them.

2. Emotion recognition systems and biometric categorisation systems may also be used where processing of the personal data of the natural person concerned is otherwise based on a legal ground under Regulation (EU) 2016/679 and the data are used exclusively for triggering a reaction that can, by its very nature, not have a negative impact on that natural person's legitimate interests and fundamental rights, and the data are erased or fully anonymised instantaneously without leaving any trace to the identifiable natural person.

3. Users of AI systems within the meaning of paragraph (1) shall inform of the operation of the system the natural persons exposed thereto unless this is inconsistent with the purpose within the meaning of paragraph (1) for which the system is used.

4. Further requirements or restrictions following from other Titles of this Act or from other laws, in particular data protection law, non-discrimination law and consumer protection law, remain unaffected.

6.4 Decisions taken on the basis of biometric techniques

Article 14(5) of the AIA Proposal provides in the context of human oversight that, for high-risk AI systems referred to in Point 1 (a) of Annex III, human oversight measures shall be such as to ensure that, in addition, no action or decision is taken by the user on the basis of the identification resulting from the system unless this has been verified and confirmed by at least two natural persons.

Given that this is more than a design requirement this rule should, as a restriction on use, be mirrored in Title IIa. At the same time however, the rule needs to be significantly modified in several respects because it is both insufficient and overreaching (see above at 4.2.2.2). For instance, 'no action or decision' would mean that not even identity control (such as requesting a passport) may follow from a high matching score, which would turn biometric identification by AI close to completely useless. This is why, in line with Article 22 GDPR, actions or decisions should only be captured by the provision if they produce legal effects or similarly significantly affect the natural person concerned. The author of this study is well aware of the fact that Article 22 GDPR is far from perfect and raises a number of difficult issues of interpretation, but creating inconsistency with Article 22 GDPR should likewise be avoided.

The rule in Article 14 (5) is also insufficient because it is restricted to biometric identification (excluding emotion recognition and biometric categorisation) and fails to give any guidance as to the independence of the two natural persons, nor on the training

they have received or on the means they use. This is why a more open provision, including emotion recognition and biometric categorisation, and focussing on the independence and on the reliability and accuracy of the means used for verification, would be preferable.

Illustration 33

Migration authorities use an AI system to analyse the spoken voice of a migrant seeking asylum with the aim of verifying whether the person seeking asylum actually originates from the geographic region from which the person purports to originate. While the result of this analysis may be an important factor, together with other factors, in establishing the relevant facts with regard to the asylum seeker's geographic origin, it must not already in itself count as legal evidence that the asylum seeker in fact originates from the region indicated by the system. Rather, there must be independent means of verification, such as an expert opinion by a human expert in the field. In the light of the significance of the decision that is at stake (i.e. whether or not asylum is granted) for the individual's life standards to be met by this expert opinion are rather high, e.g. it would not be sufficient to call some random two employees of the migration authorities in and have them confirm the result.

Again, it is not easy to decide how to exactly phrase the new restriction on decisions taken. In any case, consistency must be achieved with Article 22 GDPR and Article 11 LED.

Bearing in mind this objective, the new Article 5c could read as follows:

Article 5c
Decisions based on biometric techniques

1. No action or decision which produces legal effects concerning the person exposed to biometric identification, emotion recognition or biometric categorisation, or which similarly significantly affects that person, is taken by the user on the basis of the output from the system unless this has been verified by means that are independent from the system and that provide a degree of reliability and accuracy appropriate to the significance of the action or decision. In particular, emotion recognition systems and biometric categorisation systems must, as such, not be used as legal evidence that the natural person concerned has in fact had the emotions, thoughts or intentions recognised by the system or belongs in fact to the category assigned by the system.

2. Further requirements or restrictions following from other Titles of this Act or from other laws remain unaffected.

7 The List of High-Risk AI Systems

7.1 High-risk AI systems covered by other NLF product safety law

7.1.1 Relationship between the AIA and NLF product safety law

The AIA Proposal is largely modelled on traditional product safety law (see above at 2.1). Although it includes, in particular in its Articles 29 and 52, duties for the users of AI systems (who are usually businesses or public authorities), and prohibitions in Article 5 that area also addressed at users, the vast majority of provisions in the AIA are addressed at those who design and develop AI and place it on the market ('providers').

This raises the question of the relationship between the AIA and other product safety law, both existing and in the pipeline. The relationship is largely expressed in Article 6 (1). According to that provision, an AI system shall be considered high-risk where two conditions are fulfilled cumulatively: the AI system is intended to be used as a safety component of a product, or is itself a product, covered by particular Union legislation listed in Annex II, and the product whose safety component is the AI-system, or the AI system itself as a product, is required to undergo a third-party conformity assessment with a view to the placing on the market or putting into service of that product pursuant to the legislation listed in Annex II. This applies irrespectively of whether an AI system is placed on the market or put into service independently from another product referred to. Although Annex II lists both New Legislative Framework (NLF) legislation in Section A and old approach legislation in Section B, because of the restriction of scope in Article 2 (2), this basically means only third party conformity assessment required under NLF legislation.

While analysing all NLF instruments listed in Annex II Section A would go beyond the scope of this study it may still be interesting to analyse two examples that are particularly

relevant in the consumer context: the new Proposal for a Machinery Regulation (MR Proposal)[87] and the Safety of Toys Directive (STD)[88].

7.1.2 Analysing two examples of NLF product safety legislation

7.1.2.1 The new Proposal for a Machinery Regulation

Under Article 21 of the MR Proposal, the question whether internal conformity assessment is sufficient or third-party conformity assessment is required depends on whether or not the relevant machinery product is qualified as a high-risk machinery product listed in Annex I to the MR Proposal. Where this is the case, the manufacturer or the manufacturer's authorised representative and the person who has carried out a substantial modification to the machinery product shall apply either an EU type-examination procedure (module B) provided for in Annex VII, followed by conformity to type based on internal production control (module C) set out in Annex VIII, or conformity assessment based on full quality assurance (module H) set out in Annex IX. Where the machinery product is not a high-risk machinery product listed in Annex I, the manufacturer or the manufacturer's authorised representative and the person who has made a substantial modification to the machinery product shall apply the internal production control procedure (module A) set out in Annex VI.

Annex I to the MR Proposal lists as Point 24 'Software ensuring safety functions, including AI systems' and as Point 25 'Machinery embedding AI systems ensuring safety functions'. This means that more or less any machinery qualifies as high-risk machinery, in particular where the AI component affects the way in which the machinery moves. While the definitions of 'safety component' in the MR Proposal on the one hand and the AIA Proposal on the other are not identical, and while for the qualification as a high-risk AI system the definition in the AIA Proposal is decisive, the indicative list in Annex II to the MR Proposal may still prove to be used for determining whether or not an AI component fulfils a safety function.

[87] Above (fn. 8).

[88] Directive 2009/48/EC of the European Parliament and of the Council of 18 June 2009 on the safety of toys, OJ L 170, 30.6.2009, p. 1–37, as last amended by Commission Directive (EU) 2021/903.

So, in essence, the MR Proposal and the AIA Proposal have been closely aligned, and it has been ensured that machinery using AI as a safety component qualifies as a high-risk AI system. As the MR Proposal currently stands, this holds true even where the risk associated with a machinery product seems to be rather low.

Illustration 34

The AI embedded in a small vacuum cleaner robot for household purposes would qualify as a safety component of a product or system within the meaning of Article 3 (14) AIA Proposal, because the AI ensures, inter alia, that the robot does not hit and damage the furniture. Under the MR Proposal, the vacuum cleaner robot is subject to third-party conformity assessment because, due to it having embedded AI for safety functions, it is automatically qualified as a high-risk machinery product according to Annex I Point 25 of the MR Proposal. Since a third-party conformity assessment is necessary, the robot also qualifies as high-risk AI under the Article 6 AIA.

7.1.2.2 The Safety of Toys Directive (STD)

According to Article 19 STD, the applicable conformity assessment procedure depends on whether or not the manufacturer has applied harmonised standards, the reference number of which has been published in the Official Journal of the European Union, covering all relevant safety requirements for the toy. Where this is the case, internal production control procedures are sufficient. However, a toy must be submitted to EC-type examination, as referred to in Article 20 STD, together with the conformity to type procedure set out in Module C of Annex II to Decision No 768/2008/EC, where harmonised standards meeting all the above-mentioned requirements do not exist or where they do exist but where the manufacturer has not applied them or has applied them only in part. The same applies where one or more of the harmonised standards has been published with a restriction. Needless to say, the manufacturer should submit a toy to third-party conformity assessment also where none of these cases applies but still the manufacturer considers that the nature, design, construction or purpose of the toy necessitate third party verification.

It is not entirely clear under what conditions toys that include AI systems would qualify as high-risk AI systems under the AIA. The STD does not cover standalone software products

(such as computer games) according to its Annex I Point 15. So the most relevant cases of toys using AI for safety functions are automatically excluded, and it is unclear to what extent the STD is limited to the types of risks listed in Annex II to the STD.

Illustration 35

A speaking doll is equipped with an AI system supporting the chat bot function. When a child starts a conversation with the doll the AI system, which is connected to the Internet, searches for appropriate replies with the help of information found on the Internet. It turns out that quite harmless questions posed by five-year-old children (such as: 'Do you want to have another baby?') trigger replies from the doll with rather explicit language that is entirely inappropriate for that age group. Children react disturbed and confused, and some develop behavioural anomalies.

This kind of risk is not among the traditional risks of toys for which the STD was created, and it is certainly not among the risks addressed by the requirements listed in Annex II to the STD (which focuses on concerns such as whether small parts can be swallowed by under three-year-olds or whether toxic chemicals have been used). It is at best by reference to the general clause in Article 10 (2) STD that the chat bot function might, potentially, be included in the safety concept of the STD. (As the doll is connected to the Internet there may be mandatory third party conformity assessment under Article 17 (4) RED[89], but only if the manufacturer has not fully applied harmonised standards or where such harmonised standards do not exist, and only with regard to the essential requirements described in Article 3 (2) and (3) RED, which are very far from addressing the type of risk at stake here.)

If the chat bot function is not integrated in a speaking doll but marketed as standalone software for children there is clearly no product safety regime in place that would cover this kind of item, and as such software is not listed in Annex III of

[89] Directive 2014/53/EU of the European Parliament and of the Council of 16 April 2014 on the harmonisation of the laws of the Member States relating to the making available on the market of radio equipment and repealing Directive 1999/5/EC, **OJ L 153, 22.5.2014, p. 62–106.**

the AIA Proposal either, a chat bot for children would not be considered a high-risk AI system.

7.2 The notion of 'fundamental rights risks' and criteria for risk classification

7.2.1 General approach to risk classification

As has been explained further above (at 2.3), the AIA Proposal covers both safety risks and fundamental rights risks. While, theoretically, Articles 6 and 7 cover both types of risks in the same way, it is quite clear that Article 6 (1) in conjunction with NLF product safety legislation primarily, if not exclusively, covers safety risks whereas Article 6 (2) in conjunction with Annex III and Article 7 primarily, if not exclusively, fundamental rights risks.

It is only in Article 7 AIA Proposal, which is on amendments to Annex III, that the AIA Proposal gives some indications as to what counts as a high-risk AI system beyond the systems covered by other product safety legislation. Article 7 (1) (b) clarifies that an AI system must pose a risk of harm to the health and safety or a risk of adverse impact on fundamental rights that is, in respect of its severity and probability of occurrence, equivalent to or greater than the risk of harm posed by the high-risk AI systems already referred to in Annex III. When assessing whether that is the case the European Commission must take into account a number of criteria listed in Article 7 (2):

- the intended purpose of the AI system;
- the extent to which an AI system has been used or is likely to be used;
- the extent to which the use of an AI system has already caused harm to the health and safety or adverse impact on the fundamental rights or has given rise to significant concerns in relation to the materialisation of such harm or adverse impact, as demonstrated by reports or documented allegations submitted to national competent authorities;
- the potential extent of such harm or such adverse impact, in particular in terms of its intensity and its ability to affect a plurality of persons;
- the extent to which potentially harmed or adversely impacted persons are dependent on the outcome produced with an AI system, in particular because for

practical or legal reasons it is not reasonably possible to opt-out from that outcome;

- the extent to which potentially harmed or adversely impacted persons are in a vulnerable position in relation to the user of an AI system, in particular due to an imbalance of power, knowledge, economic or social circumstances, or age;
- the extent to which the outcome produced with an AI system is easily reversible, whereby outcomes having an impact on the health or safety of persons shall not be considered as easily reversible;
- the extent to which existing Union legislation provides for:
- effective measures of redress in relation to the risks posed by an AI system, with the exclusion of claims for damages;
- effective measures to prevent or substantially minimise those risks.

7.2.2 Risks for society at large as fundamental rights risks?

In contrast with an undated interim draft of the AIA Proposal that had been leaked at the beginning of 2021, the AIA Proposal as it was finally published on 21 April no longer contains a more detailed description what counts as relevant 'harm' and as 'risk', in particular, whether harm caused not primarily to a particular individual but to society at large should be included or not.[90] The leaked version still included an explicit reference to 'systemic adverse impacts for society at large, including by endangering the functioning of democratic processes and institutions and the civic discourse, the environment, public health, [public security];'. As the notion of fundamental rights as enshrined in the Charter is, on the one hand, primarily one of individual rights, but as, on the other hand, democracy, a liberal society, the environment and public health are of vital importance for individual rights, it is now unclear whether the AI a Proposal is narrower in its focus than the leaked version was. Arguably, the term 'fundamental rights risks' needs to be understood very broadly, and there are no indications why the European Commission would curtail its own discretion in extending the list of high-risk AI systems in Annex III by way of delegated acts to be adopted under Article 7. It is recommended to clarify the inclusion of risks for society at large within the notion of fundamental rights risks in a Recital (see below at 7.2.3).

[90] See *Veale/Borgesius* (fn. 68), p. 99 pointing out loopholes if the harm requirement is limited to individual harm.

7.2.3 Economic risks as fundamental rights risks?

In the consumer context, economic risks are often the most relevant risks. Mere economic loss is only under exceptional circumstances considered as the violation of human rights, and it is in a similar vein that liability under tort law for mere economic loss is only triggered under exceptional circumstances (such as within contractual relationships all where the tortfeasor has been acting intentionally and/or maliciously). It is therefore not clear from the outset to what extent Articles 6 (2) and 7 AIA Proposal also address economic risks.

However, Points 4 and 5 of Annex III, which are about employment, workers management and access to self-employment on one hand and access to and enjoyment of essential private services and public services and benefits on the other, and which include items such as AI systems intended to be used to evaluate the creditworthiness of natural persons to establish their credit score, indicate that economic risks are covered by the notion of fundamental rights risks when they are sufficiently serious.

In order to avoid any doubt or uncertainty, it is recommended to clarify the fact that economic risks are covered by Articles 6 and 7 AIA in a Recital.[91] This could be phrased along the lines of the following:

[…] Whereas: […]

(32) As regards stand-alone AI systems, meaning high-risk AI systems other than those that are safety components of products, or which are themselves products, it is appropriate to classify them as high-risk if, in the light of their intended purpose, they pose a high risk of harm to the health and safety or the fundamental rights of persons, taking into account both the severity of the possible harm and its probability of occurrence and they are used in a number of specifically pre-defined areas specified in the Regulation. The identification of those systems is based on the same methodology and criteria envisaged also for any future amendments of the list of high-risk AI systems. **The notion of fundamental rights risks is understood very broadly and not restricted to risks for rights explicitly mentioned under a separate Article in the Charter as long as there is a sufficient link between the risk and enjoyment of rights under the Charter. Notably, the notion of fundamental rights risk may, depending on the context, include mere economic risks where those risks are sufficiently severe, for instance because they affect access to essential goods or services (such as energy supply, credit or insurance) or because they operate on a large scale and may significantly affect the general standard of living of a natural**

[91] For a similar opinion see vzbv (fn. 68), p. 14; BEUC (fn. 68), p. 10 f.

person (such as large scale personalised pricing). The notion of fundamental rights risks also includes risks for society at large, such as for democratic institutions and a fair and open discourse.

[…]

7.3 Critical assessment of Annex III

7.3.1 General omission of AI intended for use by consumers

What seems striking already at first sight is the fact that the AI systems listed in Annex III are all AI systems that are used by professional users or public authorities. By contrast, they do not include AI systems that are typically to be used by private individuals, notably consumers.

Illustration 36

An autonomous shopping assistant used by consumers (see illustration 2), which may, e.g., come as standalone software, together with devices such as 'Echo Dot', or integrated in household devices such as smart fridges, would not qualify as a high-risk AI system although the risks for consumer protection are high (e.g. the risk that consumers are systematically pushed into expensive deals). This is so for want of a legal regime on software safety that would require third-party conformity assessment for such a software or for such software being listed in Annex III of the AIA Proposal.

Illustration 37

A virtual assistant designed to assist vulnerable adults in managing their daily lives, including taking contractual decisions, reminding them of important duties, and establishing contact with family members, a doctor or first aid response services in cases of a likely emergency, would not qualify as a high-risk system, again because there is no legal regime in place that would require third-party conformity assessment and because the system is not listed in Annex III.

Illustration 38

A computer game or chat bot to be used by children and that may significantly affect the child's personal development (see illustration 35) is not considered a high-risk AI system either, again for want of a legal regime on software safety that would require third-party conformity assessment for such a software or for such software being included in Annex III.

This is a conspicuous gap that needs to be filled. In the medium term, it would be desirable to create a semi-horizontal product safety regimes specifically for software[92] that will then provide for third-party conformity assessment and automatically bring high-risk software products into the scope of Title III of the AIA. In the short term, and until this can be implemented, a new area would have to be added to the list in Annex III.

This new area could be described as follows:

5a. Use by vulnerable groups or in situations that imply vulnerability to fundamental rights risks

(a) AI systems intended to be used by children in a way that may seriously affect a child's personal development, such as by educating the child in a broad range of areas not limited to areas which parents or guardians can reasonably foresee at the time of the purchase;

(b) AI systems, such as virtual assistants, intended to be used by natural persons for taking decisions with regard to their private lives that have legal effects or similarly significantly affect the natural persons;

7.3.2 The way consumer interests are addressed by Point 5

Point 5 of Annex III is the most important one in the consumer context. It is about access to and enjoyment of essential private services and public services and benefits. As Point 5 currently stands, it includes three groups of cases, only one of which addresses typical consumer interests (whereas the others rather address citizens' interests in general):

[92] For recommendations see *Wendehorst*/*Duller* (fn. 12) p. 7, 89 f.

- AI systems intended to be used by public authorities or on behalf of public authorities to evaluate the eligibility of natural persons for public assistance benefits and services, as well as to grant, reduce, revoke, or reclaim such benefits and services;
- AI systems intended to be used to evaluate the creditworthiness of natural persons or establish their credit score, with the exception of AI systems put into service by small scale providers for their own use; and
- AI systems intended to be used to dispatch, or to establish priority in the dispatching of emergency first response services, including by firefighters and medical aid.

7.3.2.1 Credit scoring and the exception for small scale providers

It strikes as somewhat odd that typical consumer interests are only addressed by Point 5 (b) on credit scoring systems.[93] The exception of AI systems put into service by small scale providers for their own use seems to be justified, though.

Illustration 39

Micro business M engages in e-commerce activities with customers who are natural persons. Over the years, M has collected a list of customers that have never paid for goods received. To such customers, M only offers delivery against advance-payment. M codes a simple algorithm to automate the decision whether or not to offer delivery only against advance-payment, taking into account total default, late payments and similar problems with regard to creditworthiness.

In this situation, it would seem exaggerated and unjustified to have M submit the algorithm to full conformity assessment under the AIA.

It should be stressed that the exceptions for small scale providers can only capture such AI systems as have been developed by the small scale providers, i.e. these parties must really be 'providers' within the meaning of the AIA, and not only 'users'. This could be clarified in the wording by adding the words 'of AI systems' after 'providers'.

[93] See also vzbv (fn. 68), p. 10 f.; *Ebers et al* (fn. 76), p. 594.

Illustration 40

Assume that micro business M in illustration 39 has bought AI-driven credit scoring software from big software company S. In this case, it must be clear that the credit scoring software is fully subject to the requirements under the AIA irrespective of the fact that it is (in this particular case) used by a small scale business.

7.3.2.2 Evaluation of factors with a similar effect as the evaluation of creditworthiness

However, the question arises whether the restriction to AI systems used to evaluate the creditworthiness or to establish a credit score makes Point 5 (b) too narrow. There are many other criteria with evaluation has a very similar effect as the evaluation of the creditworthiness and which are not mentioned in Point 5 (b):

Illustration 41

Assume that the software which M has bought from big software company S in illustration 40 does not evaluate the creditworthiness of customers, but factors such as the return rate (i.e. the percentage of withdrawals from purchases based on consumer protection law), the number of complaints about non-conformity of the goods, or the ratings for traders provided by that customer on reputational systems. Depending on the 'customer evaluation score', M might refuse to contract with a customer, pretending goods ordered are out of stock, or charge the customer a higher price.

Despite the fact that this AI system affects customers in at least the same way as a credit scoring system used by an e-commerce retailer the AI system is currently not covered by Point 5 of Annex III

Most notably, AI systems used for individual risk assessments by insurance companies are currently not covered by Point 5 (see already illustration 3):

Illustration 42

Insurance company I uses an AI system for deciding about whether, or on which terms, to offer private health insurance to consumers. The AI system is fed with data from a variety of sources, including (with the consumer's consent) data concerning shopping behaviour, dietary preferences, and physical activities (e.g. a step counting app or data from a fitness bracelet). In the absence of such data the AI system would normally assume that the consumer does not lead a very healthy life. Depending on the score produced by the system consumers may be denied insurance, or be offered insurance only on rather unfavourable conditions, or with restrictions (assume that a human is involved in the decision in a meaningful way so that the prohibition under Article 22 GDPR does not apply).

Although this AI system has a significant impact on individuals' lives it is currently not qualified as a high-risk AI system under the AIA Proposal.

7.3.2.3 Personalised pricing

More generally, personalised pricing (see already illustration 5) is currently not qualified as a high-risk AI system even though it may come with massive potential for discrimination and unfair treatment for consumers:

Illustration 43

Big software company S not only provides AI that rates consumers according to their creditworthiness or other factors (see illustration 40 and illustration 41) but also AI systems for personalised pricing, which aim at predicting the maximum price an individual consumer is prepared to pay for a particular item in order to offer the consumer exactly this price. This system is successful due to the fact that the personalised pricing AI provided by S is dominant on the market, i.e. the majority of retailers, in particular those present on some gatekeeper size online marketplaces, use this personalised pricing AI. Consumers who request erasure of their profiles under Article 17 GDPR are normally offered a higher price because they lose their 'personal discounts', incentivising (or even forcing) them to join the system again. On balance, this means that consumers pay significantly more, and

that there is conspicuous indirect age discrimination because younger and more IT-savvy consumers start using profile optimization AI (i.e. software simulating a browsing and shopping behaviour that tends to yield lower prices). Whether or not there is also indirect gender discrimination is hard to tell because nobody how the algorithm works and has properly tested it.

While it may be overreaching to include also personalised ranking (which is now addressed both by Article 6a (1) (a) of the revised CRD and by the P2B-Regulation) AI systems used for personalised pricing within the meaning of Article 6 (1) (ea) CRD should also be considered high-risk AI systems. As for creditworthiness, an exception for small scale providers using the system exclusively for their own purposes is justified.

7.3.2.4 Recommendations

At the end of the day, Point 5 of Annex III should be significantly extended and reformulated along the lines of the following: [94]

5. Access to and enjoyment of essential private services and public services and benefits**, including access to products**:

   (a) AI systems intended to be used by public authorities or on behalf of public authorities to evaluate the eligibility of natural persons for public assistance benefits and services, as well as to grant, reduce, revoke, or reclaim such benefits and services;

   (b) AI systems intended to be used

      (i) to evaluate the creditworthiness of natural persons or establish their credit score~~,~~

      (ii) to evaluate the behaviour of natural persons such as with regard to complaints or the exercise of statutory or contractual rights in order to draw conclusions for their future access to private or public services,

      (iii) for making individual risk assessments of natural persons in the context of access to essential private and public services, including insurance contracts, or

      (iv) for personalised pricing within the meaning of Article 6 (1) (ea) of Directive 2011/83/EU,

[94] See vzbv (fn. 68), p. 12 for further suggestions of AI systems to be added.

with the exception of AI systems put into service by small scale providers **of AI systems** for their own use;

(c) AI systems intended to be used to dispatch, or to establish priority in the dispatching of emergency first response services, including by firefighters and medical aid.

7.3.3 Selected further observations on Annex III

7.3.3.1 Biometric identification and categorisation of natural persons

The list of high-risk AI systems in Annex III starts with the category 'biometric identification and categorisation of natural persons'. As has been demonstrated above (see 6.3), fundamental rights risks inherent in emotion recognition systems are likewise significant. It is recommended to phrase the description of the category in Point 1 more broadly, referring to 'biometric techniques'. It should be considered whether emotion recognition systems should generally be considered as high-risk AI systems.

6. Biometric **techniques** ~~identification and categorisation of natural persons~~:

(a) AI systems intended to be used for the 'real-time' and 'post' remote biometric identification of natural persons;

(b) **AI systems intended to be used for emotion recognition where that recognition may lead to a decision that produces legal effects for the relevant natural person or similarly significantly affect him or her**;

7.3.3.2 Management and operation of critical infrastructure

According to Point 2 (a) high-risk AI systems including AI systems intended to be used as safety components in the management and operation of road traffic and the supply of water, gas, heating and electricity. This item seems to be somewhat an alien element within the list of high-risk AI systems in Annex III as the specific risk to be captured is of an entirely different nature. It would be desirable for the security (rather than the safety) of critical infrastructure to be dealt with separately under legislation specifically addressing critical infrastructure than in Annex III. It is also questionable whether submitting AI systems used as safety components for critical infrastructure should be subjected to a general product safety regime including surveillance by market surveillance authorities. It is in particular the comprehensive rights on the part of market surveillance authorities in all Member States under Article 64 to require full access to data and documentation and to the source code of AI systems that seems disturbing from a national security

perspective. Where all data and documentation including the source code need to be fully disclosed to any market surveillance authority in any Member State where the relevant system is placed on the market dangers that the information is leaked to malicious actors, including third state actors, is huge. In any case, point 2 (a) should be understood narrowly and only encompassing safety components in the narrower sense, and not components ensuring the resilience and security of systems.

Illustration 44

AI systems providing heat control for a decentralised heating system, preventing over-heating of the system and damage to property (including a risk of fire), may be included in the market surveillance scheme of the AIA.

However, an AI system designed to detect and defend against external attacks on a power plant should, while it is certainly a high-risk system, not be subjected to the regular product safety regime, but to a special regime.

7.3.3.3 Employment, workers management and access to self-employment

Point 4 of Annex III covers AI systems intended to be used for recruitment or selection of natural persons, notably for advertising vacancies, screening or filtering applications, evaluating candidates in the course of interviews or tests. It also covers AI intended to be used for making decisions on promotion and termination of work-related contractual relationships, for task allocation and for monitoring and evaluating performance and behaviour of persons in such relationships. While it is very important that such AI systems are covered the concrete formulation of Point 4 should be reconsidered. In particular, referring to the evaluation of candidates only where such evaluation occurs in the course of interviews or tests seems to be too narrow. A lot of very sensitive screening and evaluation activities occur outside interviews or tests, e.g. social media harvesting in order to find out details about a candidate's private life.

Illustration 45

Company E applies AI systems throughout their recruitment procedures. In particular, once candidates have been shortlisted, an AI system is used to conduct

extensive social media harvesting and other research about a candidate's private life, collecting, e.g., private photos and posting on social media and analysing them in order to obtain a rather granular personality profile and a recommendation score provided by the recruitment system. The way point 4 (a) of Annex III is currently phrased such an AI system might possibly not be included in the formulation (although this is not entirely clear as one could also say that it is captured by the general formulation upfront).

In order to be on the safe side, point 4 should be amended as follows:

4. Employment, workers management and access to self-employment:

(a) AI systems intended to be used for recruitment or selection of natural persons, notably for advertising vacancies, screening or filtering applications, **or for** evaluating candidates ~~in the course of interviews or tests~~;

(b) AI intended to be used for making decisions on promotion ~~and~~ **or** termination of work-related contractual relationships, for task allocation and for monitoring and evaluating performance and behavior of persons in such relationships.

8 Individual Rights

8.1 Individual rights and product safety legislation

8.1.1 Are there currently any individual rights in the AIA Proposal?

Given that the AIA Proposal is largely modelled on traditional product safety law the vast majority of provisions in the AIA are addressed at those who design and develop AI and place it on the market ('providers') (see above at 2.1). Although national tort law may provide that failure to comply with product safety requirements can lead to a claim for damages on the part of a consumer (on liability see below at 9), and although there may be some cases where there is a direct contractual link between the consumer and the provider and the consumer has contractual rights against the provider, there are no individual rights for consumers against providers that would follow directly from the AIA Proposal. Things are similar with regard to rights against the users of AI systems (who are usually businesses or public authorities): As far as the AIA Proposal includes obligations of users (e.g. Articles 5, 29 and 52) consumers may have individual rights where a 'transmission link' under national tort or contract law gives them such an individual right (e.g. under the doctrine of 'Schutzgesetzverletzung' in Austrian or German law). However, subject to the Court of Justice interpreting the AIA in a more generous way, there are again no individual rights against users of AI systems that follow directly from the AIA.

Illustration 46

Imagine a situation such as in illustration 12 where company V provides an online video game and uses subliminal techniques, making sure that once a gamer shows signs of fatigue and seems to be planning to leave the game that gamer receives an attractive offer to continue, and/or experiences a victory, in order to make users keep playing the game for excessive hours. Where gamer G, as a consequence, suffers physical or psychological harm and needs medical treatment, G will usually have a claim for damages under national tort and/or contract law, and the prohibition under Article 5 (1) (a) AIA may be a decisive argument for granting such a claim. Under certain circumstances, G may even have a right law to require V to stop the unlawful practice and provide

to G a non-infringing online game. However, subject to the Court of Justice taking a different view, G could not sue V before a court on the basis of Article 5 (1) (a) AIA as such.

The same holds true for the requirements following from Title III of the AIA, such as on human oversight.

Illustration 47

An AI system used for cancer diagnosis complies with the requirements under Article 14 AIA by human oversight measures that are partly built into the AI system and partly remain to be implemented by the user. Hospital H fails to provide adequate training to a young doctor who has just started work in the cancer care unit, and who fails to properly consider the instructions for use and to duly monitor the operations of the system, which is why early stage cancer in patient P is overlooked. In this case, P has a claim for damages under national tort and/or contract law, and the obligations which H had under Article 29 AIA may be an important argument in that context. However, again subject to views taken by the Court of Justice, P could not take H to court directly relying on Article 29 AIA as such.

8.1.2 Should the matter remain delegated to the GDPR?

Apart from any 'indirect effect' which the prohibitions and obligations under the AIA may have through national tort law, contract law or other areas of the law, including through national law implementing the PLD, individual rights with regard to AI systems are largely left to the GDPR and other specific legislation, such as non-discrimination law. It is in particular the data subject's rights under Article 22 GDPR as well as Articles 13 (2) (f), 14 (2) (g) and 15 (1) (h) GDPR that apply to the use of AI systems on a more general level.

In the light of the prominence of the debate around Explainable AI ('XAI') it came as a surprise that the AIA Proposal does not contain any sort of information duty on the part of users of high-risk AI systems, or maybe an information right on the part of those affected by decisions taken. As far as a duty to justify and explain does not follow from sectoral legal instruments (such as administrative procedure law or consumer contract law) the only duties that apply on a more horizontal level seem to be the information duties

following from Articles 13 to 15 GDPR. These duties remain unaffected by the AIA, as is clarified by Article 29 (2) AIA. As far as the information duties include 'meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject' they are, in principle, a possible basis for individual rights to obtain information about the parameters on which a decision relies.[95] However, it has to be borne in mind that an enforceable individual right to receive that information exists only in cases covered by Article 22 (1) and (4), i.e. only where a decision is fully automated and made without any meaningful human involvement and where the decision produces legal effects or similarly seriously affects the relevant individual (on details see above at 4.2.1.1).

Illustration 48

M is a young mother who had applied for a place in the nearby kindergarten for her two-year-old son S. She received a rejection letter which only states that, much to the kindergarten company K's regret, S cannot be offered a place this year.

Variant 1:

The decision was made by fully automated means (assuming this was admissible under Article 22 GDPR, e.g. in the light of the high number of applications). Under Article 13 GDPR, M was informed, when she filed her application, about the personal data processed (such as age of S, date application was filed, proximity of residential address, employment of M, and the family income) and the logic involved (e.g. indicating that priority was given to older children, parents who had been on the waiting list for a longer time, who are employed, and who have a low income). K must also, under Article 22 (3) GDPR, give M the opportunity to obtain review of the decision by a human employee, to express her point of view and to contest the decision.

Variant 2:

[95] For details see, e.g., *Hacker/Passoth,* Varieties of AI Explanations under the Law. From the GDPR to the AIA, and beyond, SSRN Papers 2021.

While an algorithm calculates a score from the parameters and recommends to K the children to be accepted, the recommendation is reviewed by K's management, and sometimes (though rarely) the management will not follow the recommendation and take a different decision. In this situation, M has no specific information rights with regard to decision making. Whether she has any information rights under general or specific legal regimes depends on the applicable national law.

Furthermore, the fact that Article 22 GDPR is a misplaced provision (see above at 4.2.1.3) also means that an affected individual's rights to receive an explanation for a decision are not optimally placed in the GDPR. Again, the problem is that the specific problem of 'black box effect' is largely unrelated to traditional privacy concerns and to whether personal data are being processed. The problem becomes particularly visible when the person affected by the decision is not a natural person but a legal person and where the GDPR therefore does not apply to the processing of data relating to that legal person.

Illustration 49

Bank B uses a credit scoring algorithm for assessing the creditworthiness of micro and small enterprises, such as of C GmbH, a one-person private limited company. When C GmbH's application for fresh credit in a difficult situation related to the COVID-19 crisis is rejected upon the recommendation of the credit scoring algorithm, C GmbH does not have a right under the GDPR where the credit scoring algorithm only processed company-related data.

But even within the scope of application of the relevant GDPR provisions it is highly doubtful whether those provisions could ever lead to something close to what has been

discussed as a 'right to an explanation' in the AI debate, given that it seems to be more of a limited right to be informed rather than a full-fledged right to receive an explanation.[96]

Illustration 50

M in illustration 48 receives a rejection letter, after the decision was made by fully automated means. M was informed, when she filed her application, about the personal data processed (such as age of S, date application was filed, proximity of residential address, employment of M, and the family income) and the logic involved (e.g. indicating that priority was given to older children, parents who had been on the waiting list for a longer time, who are employed, and who have a low income). However, M does not understand on what basis her own application was finally rejected.

Leaving the matter entirely to the GDPR therefore comes, again, at the price of many gaps, and fails to live up to the high expectations which the legal community worldwide had with regard to the AIA – after so many years of debate over AI rights and AI ethics one ends up with largely the same provision that had already been the law under Article 15 of the Data Protection Directive[97] from the 1990s. Again, the question arises how the AIA can serve as a regulatory role model worldwide if one needs a very profound knowledge of the whole acquis in order to understand what the rights of persons affected by AI systems really are (see above, e.g., at 5.1.2.4 or 5.1.3.3). It is therefore clearly preferable to have separate rules, that should be as much as possible consistent with what we find in the GDPR, in the AIA.

[96] Wachter, S., Mittelstadt, B., Floridi, L. (2017), Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation, *International Data Privacy Law*, Vol. 7, No. 2, pp. 76–99.

[97] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, OJ L 281, 23.11.1995, p. 31.

8.1.3 Integrating individual rights into the AIA – extend Title III and/or Title IV?

For the reasons stated above, it seems to be advisable to include provisions on individual rights in the AIA itself.[98] These provisions should, as far as possible, be consistent with the existing provisions in the GDPR so as to cause as little potential friction as possible.

One possibility of doing so would be to include the provisions in Article 29, extending the obligations of users of high-risk AI systems, and/or in Articles 13 and 14, enhancing the existing requirements for high-risk AI systems with regard to transparency and human oversight. This would be very much in the logic of the AIA, which seeks to keep truly 'horizontal' provisions to a minimum and to focus on a range of high-risk systems exhaustively listed in Annex III and in NLF legislation. However, it is precisely this restriction which makes integration in Title III questionable as a solution. Article 22 GDPR applies across the board and has its own concept of 'high-risk', namely that a decision has legal effect or similarly significantly affects a natural person.

Illustration 51

Bank B uses an AI system for recommending to branch management, inter alia, the termination of overdraft facilities granted to micro and small business customers where those business customers appear to be in distress and to have difficulties repaying over longer periods. Micro business M, a one-person GmbH run by P, feels it is being treated unfairly and driven into insolvency while the crisis of her business was merely of temporary nature as P had to bother about her divorce and care for her elderly father (who suffered a stroke) at the same time.

This case would not be covered by Article 22 GDPR, as M is a legal person. However, the system is not a high-risk system either. Therefore, the only way (beyond national contract law) to give M certain individual rights would be inclusion of such rights in Title IV of the AIA.

[98] See also *Ebers et al* (fn. 76), p 601.

This is why the author of this study recommends, upon thorough reflection, integration in Title IV, which could be extended to AI systems posing (not just transparency risks, but also) fairness risks. However, at the end of the day, the question whether the relevant provisions should be included in Title III (and therefore restricted to high-risk systems within the meaning of the AIA) or in Title IV (and therefore extended to systems which the GDPR found deserving of specific regulation) is a question of policy choice, and it also depends on further developments with regard to Annex III.

8.2 Title IV with a new focus on individual rights

8.2.1 Scope of a revised Title IV

In defining the scope of the revised Title IV, one should consider that a need for action only exists with regard to decisions that involve a material degree of evaluation or discretion and thus involves a fairness risk for the affected person.

Illustration 52

Library L uses an AI system for recommending the suspension of contracts with users who have repeatedly and manifestly violated borrowing conditions, after analysing and evaluating a broad range of factors and making predictions about a user's likely future behaviour. Library management is free in its decisions, but normally follows recommendations by the system. User U receives a notice that his contract has been suspended for six months (which is possible under the terms of the contract and applicable law), but he believes he is being treated unfairly as he is a working single father of three children and simply did not always have the time to return books during office hours.

This is a decision involving a material degree of evaluation and discretion and should therefore be covered by the revised Title IV.

Illustration 53

Library L uses an AI system for automatically dispatching reminders wherever a user has failed to return borrowed books in time and a grace period of five days has lapsed. Even where such a system qualifies as an AI system and falls within the scope of application of the AIA as such, there is no need for special fairness control as the parameters for dispatching reminders are fully defined and there is sufficient protection by contract law if a reminder is sent without justification.

There should be no difference with regard to the obligations of users whether users employ the AI system themselves or solicit the services of a third party.

Illustration 54

Where Bank B uses an AI system for credit scoring it should be immaterial whether B has bought the AI system from provider P for use in its branch, or whether B solicits the services of credit rating company C, which, in turn, has bought the system from provider P. B should have the same obligations in either case, and P should, in either case, design the AI system in a way that B (as well as C) is able to comply with their obligations as a user.

It needs to be borne in mind that, while Title IV should primarily address the users of AI systems (as does Article 22 GDPR), the obligations of users must be mirrored in corresponding obligations on the part of providers, who must make sure that the relevant AI systems are designed in a way as to enable the users to fulfil their obligations vis-à-vis affected persons.

The introductory provision to the revised Title IV could therefore be formulated as follows:

TITLE IV

~~TRANSPARENCY~~ OBLIGATIONS FOR CERTAIN AI SYSTEMS POSING TRANSPARENCY OR FAIRNESS RISKS

Article 51a
Compliance with the obligations

1. This Title includes obligations for AI systems where one or both of the following conditions are fulfilled:

(a) use of the AI system involves a risk of confusion between AI system and humans, or their operations or activities, where such confusion might harm the legitimate interests of persons exposed to the AI system;

(b) use of the AI system leads to a decision with regard to a person that involves a material degree of evaluation or discretion and thus involves a fairness risk for the affected person.

2. The obligations of users of AI systems under this Title shall apply also to users who do not operate the AI system under their own authority but who solicit the services of another party using the AI system.

3. Providers of AI system whose intended use includes use within the meaning of paragraph 1 shall ensure that AI systems are designed and developed in such a way that users are able to comply with their obligations under this Title.

4. None of the provisions under this Title shall affect any prohibitions or restrictions for AI systems following from Title II or Title IIa or any requirements or obligations set out for high-risk AI systems in Title III of this Regulation.

8.2.2 Revising the current Article 52

If a decision were taken to revise Title IV and give it a new focus on individual rights, Article 52 could largely remain in place, but should equally be revised in several respects. Some aspects that have now been integrated in the introductory provision (above 8.1.3), such as obligations of providers and the relationship to Title III, can be removed from Article 52. Previous paragraph (2) on emotion recognition and biometric categorisation has been integrated in the proposed new Title IIa on restricted AI practices (see above at 6.3.2.3), which is why it can likewise be removed from Article 52 where it seemed to be an alien element anyway. Apart from that, the author recommends adding a new paragraph on transparency obligations with regard to social bots, which are currently not specifically covered by any legislation or upcoming legislation. Recital 70 to the AIA

Proposal anyway seems to suggest that such bots should be covered, but this does not become sufficiently clear from the blackletter.

Article 52 could therefore be phrased as follows:

Article 52
Transparency obligations for certain AI systems

1. **Users of an AI system that interacts with natural persons** ~~Providers~~ shall ensure ~~that AI systems intended to interact with natural persons are designed and developed in such a way~~ that natural persons are informed that they are interacting with an AI system, unless this is obvious from the circumstances and the context of use. ~~This obligation shall not apply to AI systems authorised by law to detect, prevent, investigate and prosecute criminal offences, unless those systems are available for the public to report a criminal offence.~~

2. **Users of an AI system that creates content or engages in [online] activities that are normally engaged in by natural persons ('bot') shall disclose that the content was created, or the [online] activities performed, by an AI system, unless the source of the content or [online] activities cannot reasonably be expected to matter to natural persons exposed thereto** ~~an emotion recognition system or a biometric categorisation system shall inform of the operation of the system the natural persons exposed thereto. This obligation shall not apply to AI systems used for biometric categorisation, which are permitted by law to detect, prevent and investigate criminal offences~~.

3. Users of an AI system that generates or manipulates image, audio or video content that appreciably resembles existing persons, objects, places or other entities or events and would falsely appear to a person to be authentic or truthful ('deep fake'), shall disclose that the content has been artificially generated or manipulated.

3a. **Paragraphs 1, 2 and 3** ~~However, the first subparagraph~~ shall not apply where the use is authorised by law to detect, prevent, investigate and prosecute criminal offences or it is necessary for the exercise of the right to freedom of expression and the right to freedom of the arts and sciences guaranteed in the Charter of Fundamental Rights of the EU, and subject to appropriate safeguards for the rights and freedoms of third parties.

4. Paragraphs 1, 2 and 3 shall not **be read as legitimising the use of AI systems referred to beyond what is permitted by other law** ~~affect the requirements and obligations set out in Title III of this Regulation~~.

8.2.3 Introducing a right to a 'fair' or 'reasonable' decision?

One of the most difficult questions with regard to AI law and regulation is whether, and if so to what extent, there can be a right to a fair or reasonable decision, and what that would mean. While there seems to be a trend to reduce the term 'fairness' to absence of discrimination on the basis of a protected attribute[99] this study uses the term in a broader sense, closer to its original meaning in everyday language.

8.2.3.1 What is different with AI as compared with human decision-making?

The first question that needs to be addressed in this context is whether, and if so why, fairness standards should be higher when AI is involved than when decisions are made by humans. The main arguments that can be put forward in favour of a higher standard are (i) the fact that AI can be designed and its operations analysed and controlled, while humans cannot be designed and their thoughts can be controlled only to a very limited extent (and fundamental rights concerns against such control would prevail over the desire for fairness), and (ii) AI systems work at scale, and a flaw in one AI system may easily affect the majority of relevant systems on the market[100].

Illustration 55

Variant 1:

Recruitment officer R discards applicant A, despite that applicant's excellent qualifications, because her face and movements remind him of his divorced wife. While this may be less than ideal and certainly cause unjustified detriment to A, it would be impossible and definitely disproportionate (and possibly a fundamental rights violation in itself) to subject all recruitment officers and their decisions to psychological scrutiny, plus it is highly unlikely that other recruitment officers in other companies will think in a similar way.

Variant 2:

[99] See, e.g., the easily understandable but obviously simplistic description at https://www.ibm.com/docs/en/cloud-paks/cp-data/2.5.0?topic=openscale-fairness-metrics-overview.
[100] *Gerards/Xenidis*, Algorithmic Discrimination in Europe: Challenges and Opportunities for Gender Equality and Non-Discrimination Law, Special Report for the European Commission, 2021, p. 46.

R is chief developer of recruitment software in the company with by far the greatest share on the market for that kind of software. The software that has been trained with labelled data sets under R's supervision systematically discards all applicants who display a faint resemblance with R's divorced wife. As about 70 % of all companies use that software (or other software developed on the basis of that software), and as A happens to have a strong resemblance with R's divorced wife, A has real difficulties finding a job. In this case, the dimension of harm caused is huge, and scrutiny of the software would, at least to a certain extent (e.g. by running test cycles with the help of other software), be possible and would not in itself raise fundamental rights concerns.

Even if there are strong arguments in favour of having much stricter fairness tests for AI than for human decision-making this does not automatically mean that the AIA should include an individual 'right to a fair decision'. If we take individual rights seriously and assume there will be proper mechanisms for the enforcement of such rights, we need to make sure such rights are not overreaching.[101] However, it is extremely difficult to define what 'fairness' means.

8.2.3.2 Spotlight on non-discrimination: different fairness metrics

There is broad consensus that AI systems should not discriminate, and that a user who deploys AI systems that have discriminatory effects may be in violation of anti-discrimination law in a similar manner as if that user had used human employees. However, the uncertainty starts already with defining what 'non-discrimination' in algorithmic decision-making actually means.[102] In particular, there is no generally accepted view as to which out of several (and partly mutually exclusive) fairness metrics is to be used for what system.[103]

[101] See also *Ebers et al*, (fn. 76) p. 600.
[102] *Wachter/Mittelstadt/Russell*, Why Fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI, Computer Law & Security Review, 2021, p. 6; *Mittelstadt*, From Individual to Group Privacy in Big Data Analytics, Philosophy & Technology, 2017, p. 475.
[103] *Dunkelau/Leuschel*, Fairness-Aware Machine Learning, 2019, p. 23 f

In the first place, we have to realise that 'fairness' depends on the concrete intended purpose, and that important choices are made already when defining that purpose.

Illustration 56

Provider P is developing recruitment software, analysing application documents and other data about past candidates as well as their performance in the relevant positions. Assume that it is possible to formulate a binary classification question (e.g. 'qualified: yes/no') and to determine the 'true' value at least for past cases. In this situation, a predictive model would have to consider, for future cases, the relationship between 'true positives' (tp), 'true negatives' (tn), 'false positives' (fp) and 'false negatives' (fn).

In the first place, P would have to consider the intended purpose of the system, i.e. how it is to be used by employers like company C. It makes a big difference whether C will use the system for taking some burden off its HR staff by filtering out obviously unfit candidates (in which case, e.g., C may accept a higher fp rate as remaining fp will likely be filtered out by HR staff anyway) or whether C will use the system for ranking the top 5 candidates after a hearing (in which case an fp among the top 5 might be a big problem).

Beyond defining the task and adjusting the system model, a choice needs to be made as to the appropriate fairness metrics. While, arguably, the best solution always lies in the combination of different metrics, it is normally not possible to satisfy them all to the same extent, so even combining different fairness metrics does not alleviate a system designer from making choices.[104] A major choice to be made is whether (or: to what extent) to focus on individual fairness (IF) or on group fairness (GF).

Illustration 57

P in illustration 56 would have to consider whether it is more important to treat similar individuals similarly or to ensure equal treatment (whatever that means)

[104] See also *Ebers et al* (fn. 76) p. 596; *Dunkelau/Leuschel* (fn. 103), p. 23 f.

between different groups of applicants characterised by a protected attribute such as gender or ethnic origin. If the focus is on individual fairness P will optimise the ratio for each individual applicant assessed by the system. However, in the light of criticism voiced against IF (inter alia: that assessment of 'similarity' between individuals will in itself be biased and that it will perpetuate gaps between privileged and unprivileged groups) and the much higher degree of public awareness and sensitivity, P may prefer to strive primarily for GF, sacrificing a degree of accuracy in the individual case (paradoxically accepting unequal treatment of some individuals on grounds of gender or ethnic origin).

Even if most fairness checks nowadays rely, at least to a large extent, on notions of GF, it is not clear what GF really means[105], as there are still many different fairness metrics to choose from[106], such as Unawareness, Demographic Parity, Proportional Parity, Accuracy Parity, Equalized Odds, Predictive Rate Parity, etc.

Illustration 58

A very minimalistic method for P in illustration 57 to strive towards group fairness would be Unawareness, i.e. deleting data referring directly to protected attributes from training data sets and making sure that if, e.g., 'sex' is removed from input data about an individual applicant the result does not change. The reason why this method is not sufficient is that there can be many highly correlated features (e.g. career breaks, which are typical for women with children) that are proxies of the sensitive attribute (e.g. sex).

P satisfies Demographic Parity in its crudest form if acceptance rates (tp + fp) of applicants from two relevant groups are equal. So, if there is a pool of 1,000 applicants, 750 of which are male and 250 female, and the task is to create a shortlist of 100 applicants, this would mean 50 male and 50 female applicants. As this would obviously ignore the difference in the distribution of applications, Proportional Parity is normally used instead ([tp + fp] / [tp + fp + tn + tn]). This criterion would be fully met if the shortlist includes 75 male and 25 female

[105] See also *Binns*, On the Apparent Conflict Between Individual and Group Fairness, 2020.
[106] See also *Kleinberg/Mullainathan/Raghavan*, Inherent Trade-Offs in the Fair Determination of Risk Scores, ACM SIGMETRICS Performance Evaluation Review, p. 40.

applicants. The downside of this method is that correlations between the protected attribute and a desired attribute are ruled out entirely (e.g. that only the toughest women made it to that point anyway) and random choices could be incentivised (e.g. to pick, out of laziness, just any 25 female applicants out of the pool of 250 and not the most qualified ones).

This is why P may prefer Accuracy Parity, which is satisfied if the probability of a true prediction ([tp + tn] / [tp + fp + tn + fn]) between the two groups is equal. As this would still allow undesirable trading (e.g. the system is more accurate with male applicants, which is why, in order to satisfy the criterion, the system deliberately makes some unjustifiable choices among male applicants) the approach works better if focused on a particular value (e.g. tp). For instance, Equalized Odds is satisfied if the system accepts an equal proportion of individuals from the true qualified fraction of each group (tp / [tp + fn]). So, if among the 750 male applicants 33 % were qualified, but among the 250 female applicants 50 % were qualified, the system would shortlist 67 male and 33 female applicants. This would come closest to what we would call 'equal opportunities', but of course it would not necessarily help reduce the gap that exists in society.

There exists a host of further fairness metrics, most of which share the strengths and weaknesses of the approaches described, but are better adapted to a particular task and context (for instance, Predictive Rate Parity (tp / [tp + fp]) may display better the employer's view that the score should reflect the applicant's true capability).

8.2.3.3 Fairness of parameters for decision-making

Far beyond what is commonly discussed under the heading of 'algorithmic fairness', there are huge difficulties defining 'fairness' already with regard to the parameters used as a basis for decision-making, and their relative weight.[107]

[107] See *Dunkelau/Leuschel* (fn. 103), p. 13 ff; *Narayanan*, 21 fairness definitions and their politics, 2018.

Illustration 59

Company C uses recruitment software for shortlisting applicants for a vacant leadership position in upper management.

Variant 1: Absolutely prohibited parameters (to be overridden)

Assume that male applicants automatically receive a higher score by the system (e.g. because the training data of successful applicants from the past were almost exclusively data relating to men), i.e. the system would already fail under an Unawareness test (see illustration 58). Less beneficial treatment of female applicants as such in an employment context (beyond genuine occupational requirements) absolutely prohibited as direct discrimination under EU anti-discrimination law. This would not change even if there were statistical and/or scientific evidence, for instance, that a high testosterone level is beneficial for, performing in this position, i.e. the system would, in such a case, have to override even the statistical and/or scientific evidence.

Variant 2: Relatively prohibited parameters (calling for justification)

The score awarded by the system depends, to a large extent, on a particular BigFive personality trait profile (e.g. high extraversion and low accommodation). Looking for a certain personality profile of applicants would, as such, not be prohibited. However, if there were evidence that this profile highly correlates with gender (e.g. applicants with this profile are, statistically, 85% male and 15% female) this could amount to indirect discrimination. and would be in need of objective justification. Methods to detect the existence of such problematic correlations have been discussed in the context of illustration 58, but the more legal question whether there is objective justification can be answered only after the problematic parameter has been identified, which may require 'post-hoc' explanations of a 'black box' model (see below at 8.2.5.1).

Variant 3: Entirely arbitrary inferences

As a result of some undetected bias in the training data, individuals whose first name starts with 'B' and who hold a degree from a university with a name exceeding 20 letters are downgraded by the system as probably failing in the

position. From all we currently know and can imagine, there will never be any scientific explanation for precisely this correlation (while there may of course be explanations why graduates from a particular university tend to fail), and there would probably be different statistical results with a much larger data set, so the inference is entirely arbitrary and coincidental. Arbitrary decisions are generally considered as 'unfair', but they are difficult to detect, except with the help of rather granular 'post-hoc' explanation methods (see below at 8.2.5.1).

Variant 4: Statistically sound inferences without scientific explanation

According to training data, individuals using particular formatting options in their application documents statistically perform much better in positions of the relevant kind, which is why such applicants automatically receive a higher score by the system. There is, as yet, no plausible explanation for the patterns detected by the AI, while it is theoretically possible that a causal explanation exists. In any case, the question arises whether mere statistics should ultimately decide about an individual's career.

Variant 5: Statistically sound inferences with weak scientific justification

According to training data, individuals wearing glasses statistically perform better in positions of the relevant kind, which is why applicants with glasses automatically receive a higher score than applicants without this kind of visual aids. There are hypothetical scientific explanations for this correlation (e.g. these individuals may have damaged their eyesight by working hard, or subordinates might more readily accept their authority because glasses are associated with seniority and high education). Still, the question remains whether it is 'fair' that what ultimately amounts to a health condition, or to the very personal decision whether to wear glasses or contact lenses, should influence an applicant's chances of being hired, and more generally whether such statistics should ultimately decide about an individual's career.

Variant 6: Failure to take into account individual circumstances

Among the main parameters relied upon by the system is the number of previous positions held by an applicant and the time spent in those positions (filtering out notorious 'job-hoppers'). There is strong statistical as well as scientific justification

for this parameter. However, the system does not include in its calculation very individual explanations and justifications (e.g. applicant A used to be married to a diplomat and had to follow her husband to changing destinations, but she is now divorced and looking for stability – as A is always filtered out by recruitment software from the beginning she never gets a chance to explain this to human HR staff).

In the above illustration, the system deficiency described in Variant 1 clearly leads to unlawful decisions, and that described in Variant 2 may lead to unlawful decisions in the absence of objective justification. In the light of the amount of public attention that has been attracted by potential discriminatory effects of AI[108], it is surprising that the AIA Proposal does not address this in any specific way and does not even clearly refer to the lawfulness of decisions made. While it is certainly possible to read this into a range of different provisions in Articles 9 to 17 with regard to providers[109] and Article 29 with regard to users, the absence of even the term 'discrimination' in the blackletter comes as a surprise (although ample use has been made of the term in Recitals and the Explanatory Memorandum).

Completely arbitrary inferences, as described in Variant 3, are not strictly prohibited by the law, except to the extent that the general right to equality (e.g. following from Article 14 ECHR) in conjunction with other fundamental rights, and/or other public or private law (e.g. contractual or pre-contractual duties of care), prohibits arbitrary decisions. Arbitrary inferences are not specifically addressed in the AIA Proposal either, but only indirectly through terms such as 'limitations of performance' or 'level of accuracy' (see e.g. Articles 13 and 15), i.e. there is arguably a general obligation to minimise such deficiencies through appropriate data governance and testing (although such arbitrary inferences can never be fully excluded with some machine learning techniques).

What is much more tricky is the situation in Variant 4 where there is mere statistical evidence of a correlation without a proven causal link, and in Variant 5, where there is statistical evidence of a correlation and some (albeit largely hypothetical) scientific

[108] See, e.g., *Wachter/Mittelstadt/Russell* (fn. 102), *passim*; *Gerards/Xenidis* (fn. 100), *passim*.
[109] See, e.g., Article 9 on risk management in general; Article 10 (2) (f) and (5) on bias monitoring, detection and correction; Article 13 (1) requiring a degree of transparency that ensures the user can comply with its own obligations (which in turn, according to Article 29 (2), include obligations under non-discrimination law as a subset of 'Union or national law'); or Article 17 (1) (a) addressing a 'strategy for regulatory compliance'.

evidence of an (albeit rather weak) causal link with a desired attribute. Predictive analytics systems heavily rely on such statistical probabilities, and statistics is the very basis of much of what we perceive as 'AI'. Data governance (in particular ensuring that training data sets are sufficiently large and sufficiently diverse) and proper system design and testing should help avoid that a particular statistical correlation is given disproportionate weight (e.g. that a top 2 % candidate is discarded exclusively due to the fact that she prefers contact lenses over glasses), which is again something that can be subsumed under 'bias', 'accuracy' and similar general terms.[110] However, for want of an open debate in society to what extent and under what conditions we are prepared to accept statistics[111] as a basis for algorithmic decision-making it is largely impossible to say whether the decisions described in Variants 4 and 5 are 'fair' or not.

Finally, also the situation in Variant 6 poses a particular challenge with regard to the use of AI. In the light of limitations in the availability of training data (e.g. it is improbable that there will be many data on freshly divorced ex-wives of diplomats in the data sets used for training recruitment AI) and the multitude of possible individual constellations, it is difficult to tackle this, except again by way of reference to very general notions of 'risk management' etc.

8.2.3.4 Should an individual right to a fair decision be introduced in the AIA?

On balance, and given the difficulties of defining fairness in a legislative act on AI and of affording individuals a right to a fair decision, the author supports the decision made by the drafters not to include more specific fairness provisions in the AIA itself,[112] and in particular not an individual (and enforceable) right to a fair or reasonable decision. As far as decision-making is incompatible with European non-discrimination law, such non-discrimination law applies in any case, irrespective of whether discrimination occurs through human or algorithmic decision making, and irrespective of whether the user had any intention to discriminate.[113] Obliging the user of AI systems to demonstrate in each and every case that all parameters used, and their relative weight, was 'fair' or

[110] See also European Union Agency for Fundamental Rights (FRA), Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights, 2019; German Data Ethics Commission, Opinion of the Data Ethics Commission, 2019, p. 8.
[111] See also *Wachter/Mittelstadt/Russell* (fn .102), p. 4.
[112] *Contra* BEUC (fn. 68), p. 21.
[113] Opinion of the Data Ethics Commission (fn. 39), p.167.

'reasonable'[114] might easily result in overreaching results and ignore the user's own rights and liberties to organise their business and take business decisions as they deem appropriate. It may therefore be preferable to focus on a right to appropriate scrutiny and a right to receive an explanation.

8.2.4 A right to appropriate scrutiny

What seems more promising as a means to promote fairness in algorithmic decision-making is a right to appropriate scrutiny of individual decision-making. As has been demonstrated above (at 8.1.2), Article 22 GDPR provides a degree of protection, but it is nevertheless advisable to introduce a specific rule in the AIA. As to scope, the AIA provisions would be narrower than Article 22 GDPR (in conjunction with Articles 13 to 15 GDPR, see above at 8.1.2) by only covering decisions involving a degree of evaluation and discretion (in line with the specific focus on AI), but also broader than Article 22 GDPR in at least two respects, i.e. by covering also recommender systems and by protecting affected persons (including legal persons) irrespective of any processing of personal data (see above at 8.2.1).

8.2.4.1 Towards an individual right akin to Article 22 GDPR in the AIA

In order to ensure consistency with Article 22 GDPR and for stakeholders to be able to build on some decades of case law in the field of data protection, the same criteria for determining whether a decision has significant impact on a person as apply under Article 22 GDPR should also apply under the AIA, i.e. reference should be made to whether or not a decision creates legal effects concerning a person or similarly significantly affects that person (see above at 4.2.1.1). It should be mentioned, though, that the author of this study believes that this criterion must remain focussed on the significance of the decision as such and be analysed without regard to 'discrimination' and similar aspects, for otherwise things get circular.

[114] See *Wachter*/*Mittelstadt,* A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI, Colombia Law Review, 2019, p. 81 ff.

Illustration 60

Platform operator P uses an AI system making suggestions to customers of the type "Customers who bought this item were also interested in …". Customer C who just bought a calculator receives a suggestion to take a look also at a leather case for this calculator. This low-risk decision is definitely not a decision that should fall under any enhanced fairness requirements provided for by the AIA. Contrary to what seems to be the view of WP29 with regard to Article 22 GDPR,[115] this does not change even if the low-risk decision turned out to be discriminatory (e.g. making this suggestion more often to female customers than to male customers). Discrimination must be stopped, not confirmed by a human or explained, and it is non-discrimination law that deals with such matters.

The AIA should take over the solution adopted under Article 22 GDPR that there should normally be meaningful ex-ante control but that in certain cases, such as conclusion of mass transactions, appropriate other safeguards, in particular ex-post control, could be sufficient. However, it is suggested that the AIA provision should be a little more specific than the GDPR when it comes to the details of human scrutiny. In particular, there should be a clarification that the relevant person must have the abilities, training and decision-making authority, sufficient information with regard to the individual case, and that adequate safeguards against automation bias must be in place. Also the requirements of ex-post control could be a bit more explicit than in the GDPR.

Considering the difference in scope, the provision in the AIA should not be a prohibition in the form of an 'all-or-nothing' rule (as is Article 22 (1) GDPR) with a set of rather counter-intuitive exceptions, but should instead focus on the level of scrutiny required. Also, it should not be strictly (another) human intervention that is required as other types of verification, including by the same system or by an independent AI system, can be sufficient, depending on the nature and significance of the decision and the role played by the AI in the decision-making process.

[115] Article 29 WP, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, (2017), p. 10.

Illustration 61

There have been 1,304 applications for a vacant position in upper management within company C. Applicants are evaluated and ranked by recruitment software. As a first step, obviously unsuitable candidates are discarded, as a second step the 100 most promising applications selected, and as a third step a shortlist of 20 candidates prepared to be scrutinised by the HR department, which will then decide which applicants to invite to a hearing. All candidates invited to the hearing will be ranked by the recruitment software, and the score achieved by a candidate will play a role in the ultimate decision.

Concerning the first step, no further measures of verification beyond general human oversight required under Articles 14, 29 of the AIA Proposal should be necessary, i.e. it should not be required that the (possibly over 1,000) applications of obviously unsuitable candidates are dealt with individually by humans because it is highly improbable that an applicant discarded in this first step would ultimately be selected. Verification duties would increase with every step, and with regard to the ultimate step, i.e. the granular ranking of hearing candidates, it would be necessary to have human HR staff with all the information, decision-making power and appropriate training to resist automation bias.

A possible formulation of such a provision in the AIA could read as follows:

Article 52a
Scrutiny of individual decision-making

1. No decision which produces legal effects concerning a person, or which similarly significantly affects that person, is taken by the user on the basis of the output from an AI system unless the appropriateness and fairness of this decision has been verified by means that are appropriate to the nature and the significance of the decision and the role of the AI system in the decision-making process.

2. Unless otherwise specified by Union or Member State law, verification within the meaning of paragraph 1 may, in particular, consist in meaningful scrutiny, before the decision is taken, by a natural person who is equipped with the appropriate

(a) abilities, training and decision-making authority;

(b) information with regard to the individual case; and

(c) safeguards against automation bias.

3. The user may replace ex-ante verification within the meaning of paragraphs 1 and 2 by equivalent other measures where the affected person has given explicit consent or where ex-ante verification is impossible or would cause unreasonable effort and is not strictly necessary for safeguarding the affected person's rights and freedoms and legitimate interests. Unless otherwise specified by Union or Member State law, such equivalent other measures may, in particular, consist in the right to

(a) obtain human intervention that satisfies the requirements under paragraph 2;

(b) provide additional information and express his or her point of view; and

(c) contest the decision with a meaningful chance of having it revised.

8.2.4.2 Relationship with the proposed new Article 5c

The proposed new Article 52a bears strong resemblance with the proposed new Article 5c on decisions based on biometric techniques (see 6.4 above), which begs the question whether Article 5c should be integrated in Title IV. However, the proposed Article 5c addresses a slightly different problem than is addressed by Article 52a, because Article 5c is mostly about situations where the AI system is used to establish facts, whereas Article 52a is about situations where the AI system is used to take decisions involving a material degree of evaluation or discretion. This does not exclude that, in an individual case, both provisions can be applicable at the same time.

Illustration 62

An incident on High Street is captured by video surveillance facilities. With the help of biometric techniques, S is identified by an AI system as being the individual displayed on video recordings. Verification of the fact that the individual is indeed S would be addressed by Article 5c. However, where a second AI system is used to analyse the situation and to recommend whether S's conduct should be qualified as battery or in some different way under applicable criminal law, this would be a decision to which Article 52a applies.

8.2.5 A right to receive an explanation

Another promising tool to promote fairness in algorithmic decision-making is an individual right to receive an explanation.[116] As has already been pointed out above (8.1.3) it strikes as odd after so many years of debate about explainable AI ('XAI') that the AIA Proposal does not even mention the concept, at least not in the relationship between user and affected person (while Article 13 provides for a degree of transparency in the relationship between provider and user).

8.2.5.1 Approaches to explainability

There are largely two different approaches to ensure the explainability of AI systems: 'ante-hoc XAI' or 'white box' models, and 'post-hoc' explanation of 'black-box' models.

'Ante-hoc XAI' or 'white box' models ensure transparency from the outset because they are inherently interpretable. The idea in all of them is to directly quantify the computation and parameters and keep them at an interpretable level. This includes, for example, regressions or decision trees and random forests. Generative Additive Models (GAMs) allow to identify the weighting of each input variable. In hybrid models, rule-based methods are combined with machine learning methods, restricting black-box elements to selected subtasks.

'Post-hoc XAI' or post explanation of 'black box' models uses a range of different methods that provide either for parallel logging during training or simulate or test the entire model to quantify it. 'Local Interpretable Model-Agnostic Explanations' (LIME)[117] rely on the idea to make a given model ('local') understandable ('interpretable') for a human without knowledge about a specific model ('model-agnostic'). For example, linear classifiers may be applied to neural network results to make them interpretable. The 'Counterfactual Method' relies on the fact that the output of a model is the direct result of the input. Targeted input elements are manipulated until one can observe a change in the output; repeating this method systematically, one can work out which subtleties in the input explain the output. 'Layer-wise Relevance Propagation' (LRP)[118], by contrast, tries to

---

[116] For a similar opinion see vzbv (fn. 68), p. 21; BEUC (fn. 68), p. 20.

[117] *Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin,* "Why Should I Trust You?" Explaining the Predictions of Any Classifier (2016), arXiv:1602.04938v1

[118] *Alexander Binder*, *Sebastian Bach*, *Gregoire Montavon*, *Klaus-Robert Müller*, and *Wojciech Samek*, Layer-wise Relevance Propagation for Deep Neural Network Architectures, in: Kim K., Joukov N. (eds) Information

ensure the explainability by a backward distribution. To do this, a neural network back-propagates the output to the weighted nodes from the layer before, allowing to identify the most important node-edge combinations and thus to mark the greatest influence of certain parts of the input. 'Partial Dependency Plot' (PDP) shows what effect features have on the output of the model, e.g. whether the relationship between target and feature is linear, monotonic, or complex. Besides the methods mentioned there is a whole range of other approaches for explainable AI, for example 'Individual Conditional Expectation' (ICE), 'Accumulated Local Effects' (ALE), 'Feature Interaction', etc.[119]

The boundaries between 'white box' and 'black box' AI become almost blurred with approaches of Rationalization,[120] where black box systems are equipped with a deeper computational layer that logs why an action triggers and makes this information explainable to humans.

8.2.5.2 Towards an individual right in the AIA

An explanation of individual decision-making could significantly enhance fairness of algorithmic decision-making for a number of reasons, not least by enabling users of AI systems to assess better the quality of the decisions they make and improve the quality where necessary, and by empowering those affected by AI and enabling them to contest a decision with a realistic chance of having it revised. In order to ensure consistency with Article 22 GDPR, reference should again be made to whether or not a decision creates legal effects concerning a person or similarly significantly affects that person.

What seems important is that the explanation covers not only the role of the AI system in the decision-making process, but also the logic involved, the main parameters of decision-making, and their relative weight. It should also include the input data relating to the affected person and each of the main parameters on the basis of which the decision was made (such 'main parameter' information rights exist, e.g., in the P2B Regulation and in the CRD, see above at 7.3.2.1). For such information to be meaningful and for reasonably enabling the affected person to contest the decision the information must include an

---

Science and Applications (ICISA) 2016. Lecture Notes in Electrical Engineering, vol 376; https://doi.org/10.1007/978-981-10-0557-2_87.

[119] For an overview, see, e.g., *Kilian Semmelmann*, Was ist Explainable AI (XAI)? Deifinitionen und Beispiele, https://datadrivencompany.de/explainable-ai/.

[120] *Upol Ehsan*, *Brent Harrison*, *Larry Chan*, *Mark O. Riedl*, Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations (2017), arXiv:1702.07826v2.

easily understandable description of inferences drawn from input data (this would normally mean a description in natural language, but could mean other forms of description where the affected person also uses AI).

Illustration 63

Applicant A in Variant 6 of illustration 59 receives a rejection letter. The explanation to be provided would include the role of the AI system in the hiring process (e.g. filtering out unfit candidates, or creating a shortlist of most promising candidates) and the main parameters for decision-making and their relative weight (e.g. professional qualifications, years of experience in similar positions, frequency of change of employer, personality traits, …).

For some of the main parameters, the relationship with individual input data will be obvious and easily assessable by the applicant (e.g. A will know that she has changed her employer quite frequently). Still, it may help all parties involved in detecting errors and improving the quality of the system if the data used as a basis for calculation are disclosed (e.g. that the system calculated that A spent an average of 12.5 months with an employer). For parameters that are not so obvious (e.g. personality traits) it would not be sufficient to disclose the individual input data (e.g. the replies A gave in a questionnaire) but also an understandable description of inferences drawn from the input data (e.g. A's BigFive scores for extraversion and accommodation).

The explanation should be provided at the time when the decision is communicated to the affected person. However, the user may provide the explanation only upon the affected person's request, where providing the explanation immediately would cause unreasonable effort and is not strictly necessary for safeguarding the affected person's rights and freedoms and legitimate interests. In this case, the user would have to notify the affected person of the existence of the right, and of how it can be exercised.

Illustration 64

Strictly speaking, and for the explanation to be fully effective to safeguard A's legitimate interests in illustration 63, A would have to receive the explanation in

time before C takes any further steps in the hiring process. However, informing applicants at each stage of a hiring process might harm C's legitimate interests (e.g. making it difficult for C to revert to applications previously discarded and causing administrative burden and delay if applicants challenge interim decisions). This is why the AIA should content itself with requiring the explanation at the point in time when the decision is communicated to the affected person, leaving the determination of this point in time to otherwise applicable law.

As this means that A will likely receive the explanation only at a point in time when the vacancy has already been filled it is not even necessary to send A an explanation automatically (in fact, many applicants will prefer not to be confronted with their own deficiencies). Instead, C could give A notice that explanation can be obtained upon request.

Even if this means that A will receive the explanation only at a point in time when the vacancy has already been filled it will help A understand why her application has been discarded and enable her to phrase her next application differently and to make company C and/or provider P aware of the problem.

In a number of cases, a right to receive an explanation will not be appropriate. This will include cases where the AI system had only minor influence within the decision-making process.

Illustration 65

If the AI system was used by C only for establishing certain facts, such as for discarding obviously incomplete applications (e.g. without a CV), this would not fall under the scope of revised Title IV at all as such 'decisions' do not involve any evaluation or discretion. But even in a less clear-cut case, such as where the AI system only adds a colour code (red/yellow/green) to an application, but all applications are assessed individually by HR staff nevertheless and the colour code does not play a significant role in the whole process, an applicant's right to receive an explanation might seem to be overreaching.

Obviously no explanation should be required for AI systems that are authorised by law to detect, prevent, investigate and prosecute criminal offences or other unlawful behaviour, because if the parameters and their relative weight becomes known this will facilitate circumvention and help foster crime and other unlawful activities. As it will be difficult to list exhaustively all case where explanations would be inappropriate, there should be a possibility for Union or Member State law to provide for exceptions from, or restrictions to, the duty to provide explanations, provided such law lays down appropriate other safeguards for the affected person's rights and freedoms and legitimate interests. A person should also be able to give explicit consent not to receive an explanation (which must be 'free' and satisfy other established criteria for consent within the meaning of the GDPR).

In essence, a possible formulation in the AIA could read as follows:

Article 52b
Explanation of individual decision-making

1. A decision which is taken by the user on the basis of the output from an AI system and which produces legal effects concerning a person, or which similarly significantly affects that person, shall be accompanied by a meaningful explanation of

(a) the role of the AI system in the decision-making process;

(b) the logic involved, the main parameters of decision-making, and their relative weight; and

(c) the input data relating to the affected person and each of the main parameters on the basis of which the decision was made.

For information on input data under point (c) to be meaningful it must include an easily understandable description of inferences drawn from other data if it is the inference that relates to a main parameter.

2. Paragraph 1 shall not apply to the use of AI systems

(a) that have only minor influence within the decision-making process;

(b) that are authorised by law to detect, prevent, investigate and prosecute criminal offences or other unlawful behaviour;

(c) for which exceptions from, or restrictions to, the obligation under paragraph 1 follow from Union or Member State law, which lays down appropriate other safeguards for the affected person's rights and freedoms and legitimate interests; or

(d) where the affected person has given explicit consent not to receive an explanation.

3. The explanation within the meaning of paragraph 1 shall be provided at the time when the decision is communicated to the affected person. However, the user may provide the explanation only at a later point upon the affected person's request, where providing the explanation immediately is not strictly necessary for safeguarding the affected person's rights and freedoms and legitimate interests, in which case the user shall inform the affected person of the right under this Article and how it can be exercised.

9 Liability

9.1 The current and future law of liability for AI

9.1.1 The relationship between safety and liability

In principle, product safety law and product liability law (or other types of liability for products) both serve as a response to risks[121] and relate to each other like a system of communicating tubes. While product safety law takes an ex ante-perspective and seeks to avoid risks from being created and harm from being caused, product liability law takes an ex post-perspective and seeks to provide compensation for risks that have materialised and harm that has been caused. Of course, liability may also serve as an incentive for taking precautionary measures in order to avoid liability in the first place, so there is a certain ex ante aspect even to liability.[122]

Approaches as to the right balance between safety and liability differ. Europe has always put the stress on safety, for various reasons, including that death, personal injury, and (other) fundamental rights infringements cannot simply be reduced to a monetary figure and that a purely economic approach often fails to take into account the real cost of accidents, e.g. the economic harm caused by a general lack of trust on the part of consumers and other collective harm and social concerns.[123]

9.1.2 The status quo of liability law

The stress on safety does not mean that Europe has turned a blind eye on liability. There exists a rather complicated 'safety net' of different liability regimes.[124] Most of these

---

[121] For details *Wendehorst/Duller* (fn. 12), p. 33 ff.

[122] Article 10:101 Principles of European Tort Law (PETL): '*Damages also serve the aim of pre-venting harm*'. However, this effect should not be mistaken for punitive damages, which are not recognized in most European legal systems.

[123] See Commission, 'Communication from the Commission on the precautionary principle' COM(2000) 1 final.

[124] Expert Group on Liability and New Technologies – New Technologies Formation, Liability for AI and other digital technologies, European Commission (2019); *Karner/Geistfeld/Koch*, Comparative law study on civil liability for artificial intelligence (2021).

liability regimes exist at national level. They include fault liability, vicarious liability, strict liability and mixed or other forms of liability. It is only liability for defective products within the meaning of the Product Liability Directive (PLD)[125] that has undergone Europe-wide harmonisation. Apart from that, there exist some special liability regimes at EU level, notably liability under Article 82 GDPR for infringement of data protection law. In addition, a number of EU Directives in the field of consumer protection[126] and non-discrimination law[127] oblige Member States to provide in their national legislation for effective remedies, including compensation for harm caused.

Some of the liability regimes mentioned (notably the liability regime established by the PLD) specifically provide for liability on the part of the producer or other economic operators involved with the putting into circulation of products. Other liability regimes, explicitly or implicitly, provide for liability on the part of the users of products, and many liability regimes (such as general fault liability regimes under national law) are simply neutral and may lead to liability on the part of a variety of different players.

9.1.3 Making liability law fit for AI – the current debate

National laws may or may not in principle be fit for coping with the challenges posed by AI systems (this is difficult to tell due to the high level of generality of national tort laws and of the fact that, as yet, hardly any case law exists). In any case, relying on national law to evolve and adapt to new technologies would mean decades of uncertainty, quite the opposite of a level playing field, and victims in particular jurisdictions put at a massive disadvantage. This is why action needs to be taken at EU level, ensuring that victims who have suffered harm from the operation of AI systems receive adequate compensation irrespective of the (otherwise) applicable law.

[125] Council Directive 85/374/EEC of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products, OJ L 1985/210, 29.
[126] E.g. Article 11a UCPD (see above fn. 29).
[127] E.g. Article 8 (2) of the Gender Equal Access to Goods and Services Directive (see above fn.32).

Preparatory work has been underway for some time, notably with the 2019 report of the Expert Group on Liability and New Technologies,[128] the 2020 Commission report[129] and finally the 2020 legislative resolution by the European Parliament,[130] which includes a full proposal for a draft Regulation (for details see below 9.2). Legislative proposals by the Commission that had initially been announced for late 2021 or early 2022[131] have again been postponed, and only mid October 2021 a public consultation has been launched, which is open until 10 January 2022.[132]

The main point of debate is at the moment to what extent liability for AI systems should be addressed within the context of a revision of the PLD[133] and/or whether there should be an entirely new legal regime of AI liability (as has been proposed, e.g., by the European Parliament). The answer to that question is, in the first place, relevant for defining the main addressees of liability, i.e. whether liability for the specific risks posed by AI should primarily be on the providers of AI systems or on the users of AI systems. The answer to the question is, however, also relevant for identifying the DGs and Committees within the European Commission and the European Parliament that would primarily be in charge of the file.

[128] Expert Group on Liability and New Technologies – New Technologies Formation, Liability for AI and other digital technologies, European Commission (2019).
[129] Report from the Commission to the European Parliament, the Council and the European Economic and Social Committee, Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics, COM(2020) 64 final.
[130] European Parliament Resolution of 20 October 2020 with Recommendations to the Commission on a Civil Liability Regime for Artificial Intelligence (2020/2014(INL)).
[131] Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions. New Coordinated Plan on AI 2021 Review' COM (2021) 205 final, p. 33; in other places, even late 2021 is mentioned as a possibility, cf. European Commission, A European Approach to Artificial Intelligence.
[132] European Commission, Civil liability – adapting liability rules to the digital age and artificial intelligence.
[133] See, e.g., BEUC, Product liability 2.0. How to make EU rules fit for consumers in the digital age (2020); Twigg-Flesner, Guiding Principles for Updating the Product Liability Directive for the Digital Age, ELI Innovation Paper, (2021).

9.2 The EP Proposal for a Regulation on AI Liability

9.2.1 Cornerstones of the EP Proposal

The cornerstone of the EP Proposal for a Regulation of AI liability is a strict liability regime for the operators of 'high-risk' AI systems to be enumeratively listed in an Annex, accompanied by a regime of rather strict fault liability for the operators of other AI systems.

9.2.1.1 Strict operator liability for certain high-risk AI systems

According to Article 4 of the EP Proposal, operators of AI systems shall be strictly liable for any harm or damage that was caused by a physical or virtual activity, device or process driven by an AI system. The EP Proposal ultimately adopted the division into 'frontend operators' (i.e. the person deploying the AI system) and 'backend operators' (i.e. the person that continuously controls safety-relevant features of the AI system, such as by providing updates or cloud services) that had been developed by the author of this study and included in the 2019 report of the Expert Group on Liability and New Technologies.[134] According to the final version of the EP Proposal, not only the frontend operator, but also the backend operator may become strictly liable.[135] However, the backend operator's liability is covered only if it is not already covered by the PLD. [136] The only defence available to the operator is force majeure.[137] For the applications subject to strict liability, mandatory insurance is being proposed.[138]

'High-risk' AI systems for the purpose of the proposed Regulation are to be exhaustively listed in an Annex to that Regulation. Interestingly, the final version of the Proposal was published with the Annex left blank. The Annex attached to the first published draft from April 2020 had met with heavy resistance due to its many inconsistencies, and it may have proved too difficult to agree on a better version. Also, it seemed opportune to wait for the list of 'high-risk' AI applications that would be attached to the AIA. In any case, given the rapid technological developments and the required technical expertise, the idea is that

[134] NTF Expert Group (fn.128) Key Findings nos 10 and 11.
[135] See Article 3 (d) to (f) of the EP Proposal for a Regulation (fn. 130).
[136] See Article 3(e).
[137] See Article 4(3).
[138] Cf EP Proposal for a Regulation (fn. 130) Article 4(4).

the Commission should review the Annex without undue delay, but at least every six months, and if necessary, amend it through a delegated act.[139]

Now that the AIA Proposal is on the table the question arises whether the list of 'high-risk' AI systems in the AI Liability Regulation can be identical with the list of 'high-risk' AI systems provided by Article 6(1) and Annex II of the AIA, in conjunction with relevant product safety legislation. However, as tempting as it may be to simply refer to the AIA, it would lead to overreaching and inappropriate results.

Illustration 66

A small robot vacuum cleaner or a toy robot vehicle for children would be qualified as 'high-risk' under Article 6 AIA in conjunction with NLF product safety legislation (for details see above at 7.1.2). However, it would be exaggerated to impose strict liability for harm caused by small toy robots or robot vacuum cleaners, in particular if that strict liability is imposed on operators. Those machines hardly ever cause significant physical harm by themselves, and if they do, it is usually because it was improper for the (frontend) operator to deploy them in the particular situation, such as where the operator of a retirement home uses an unsupervised cleaning robot in places and at times when elderly residents might stumble over it. If all operators of small vacuum cleaner robots (e.g., including the millions of businesses that use it for cleaning their office space during the night, or even consumers) had to face strict liability and to take out corresponding insurance this would be extremely inefficient and benefit no one but the insurance industry.

9.2.1.2 Enhanced Fault Liability for Other AI Systems

The EP Proposal not only includes a strict liability regime for 'high-risk' applications, but also a harmonised regime of rather strict fault liability for all other AI systems. Article 8 provides for fault-based liability for 'any harm or damage that was caused by a physical or virtual activity, device or process driven by the AI-system', and fault is presumed, i.e. it is

[139] EP Resolution on AI liability (fn. 130), Recommendation to the Commission no 16.

for the operator to show that the harm or damage was caused without his or her fault.[140] In doing so, the operator may rely on either of the following grounds: The first ground is that the AI-system was activated without his or her knowledge while all reasonable and necessary measures to avoid such activation outside of the operator's control were taken. The second ground is that due diligence was observed by performing all the following actions: selecting a suitable AI-system for the right task and skills, putting the AI-system duly into operation, monitoring the activities and maintaining the operational reliability by regularly installing all available updates. It looks as if these two grounds are the only grounds by means of which an operator can exonerate themselves, but Recital 18 also allows for a different interpretation, i.e. that the two options listed in Article 8(2) should facilitate exoneration by establishing 'counter-presumptions'.

The proposed fault liability regime is problematic not only because of the lack of clarity in drafting, but also because Article 8(2)(b) might be unreasonably strict. More importantly, in the absence of any restriction to professional operators, even consumers would face this type of enhanced liability for any kind of AI device, from a smart lawnmower to a smart kitchen stove. Recital 18 mentions a need to adapt the standard of care for consumers, but this is not reflected in the blackletter. This could mean burdening consumers with obligations to ensure that irregularities are properly reported to the provider and that updates are properly installed, irrespective of their digital skills, and possibly confronting them with liability risks they would hardly ever have had to bear under national legal systems.

Illustration 67

50-year-old consumer K owns a smart kitchen stove, which was recommended to him by a retailer. K is not entirely happy with the stove, which tends to switch off a bit too early or a bit too late, and K is also overwhelmed by blinking signs on the display (which signal that a software update would be available). One day, while K is in the garden with friends, the stove fails to switch off for hours while the food is already burnt, causing a fire and severe damage to both the kitchen and the premises (owned by landlord L). In this case, Article 8 of the EP Proposal, even if

[140] In fact, the drafting is not very clear with regard to this point. Recital 17 seems to underline that fault is always presumed and that the operator needs to exonerate themselves. However, Recital 19 also refers to proof of fault by the victim.

read in the light of Recital 18, would arguably mean that K is liable because he failed to report irregularities to the provider as well as to install all available updates. It is not even clear whether it would make a difference if K could prove that installing the update would not have prevented the damage.

9.2.1.3 Types of harm covered

Article 2(1) of the Proposal declares the proposed Regulation to apply where an AI system has caused 'harm or damage to the life, health, physical integrity of a natural person, to the property of a natural or legal person or has caused significant immaterial harm resulting in a verifiable economic loss'. Article 3(i) provides for a corresponding definition of 'harm or damage'.

While life, health, physical integrity and property were clearly to be expected in such a legislative framework, the inclusion of 'significant immaterial harm resulting in a verifiable economic loss' came as a surprise. If immaterial harm or the economic consequences resulting from it – such as loss of earnings due to stress and anxiety that do not qualify as a recognised illness – is compensated through a strict liability regime whose only threshold is causation,[141] the situations where compensation is due are potentially endless and difficult to cover by way of insurance. This is so because there is no general duty not to cause significant immaterial harm of any kind to others, unless it is caused by way of non-compliant conduct (such as by infringing the law or by intentionally acting in a way that is incompatible with public policy).

Illustration 68

Company C uses advanced recruitment software for creating a shortlist of the top 20 applicants for a position. Applicant A is not shortlisted, which is why A is very sad and obviously suffers economic loss by not receiving the job offer. If the recruitment software were listed as a 'high-risk' AI system also within the meaning of the EP Proposal and C were therefore strictly liable for any harm caused by the operation of the system this would mean that A would receive full compensation even if the recommendation made by the recruitment software

[141] Proposal for a Regulation (fn. 130) Article 4(1).

was absolutely well-founded and if there was no discrimination or other objectionable element involved.

While some passages of the report seem to choose somewhat more cautious formulations, calling upon the Commission to conduct further research,[142] Recital 16 explains very firmly that 'significant immaterial harm' should be understood as meaning harm as a result of which the affected person suffers considerable detriment, an objective and demonstrable impairment of his or her personal interests and an economic loss calculated having regard, for example, to annual average figures of past revenues and other relevant circumstances.

9.3 Towards three pillars of future AI liability law

At the end of the day, liability for damage caused by AI systems may have to rest on three different pillars.

9.3.1 Strict operator liability for 'high-physical-risk' devices

As far as death, personal injury or property damage caused by a 'high-risk' product that includes AI for safety-relevant functions is concerned, strict liability seems to be a proper response. However, as has been demonstrated (above at 9.2.1.1), not every product that qualifies as a 'high-risk' product under the AIA fulfils the requirements that should be met for justifying strict liability (and the accompanying burden of insurance). The justification for imposing strict liability – in particular that the relevant product or activity leads to significant and/or frequent harm despite the absence of any fault or any identifiable defect, mal-performance or non-compliance – does not coincide with the justification for imposing particular precautionary measures against unsafe products. While the AI systems for which strict liability is justified will most likely be a subset of the AI systems for which enhanced safety measures are justified, by far not all AI systems of the latter type should be included in a strict liability regime, e.g. when they are normally safe except when clearly defective.

[142] EP Resolution on AI liability (fn. 130) Recommendation to the Commission no 19.

I should also be borne in mind that strict liability for physical risks caused by AI-driven devices might create significant inconsistencies if not accompanied by strict liability for the same type of devices where those devices are not AI-driven but steered by humans or by technology other than AI. A victim run over by a vehicle does not care that much whether the vehicle was AI driven or not. So if strict liability is found to be appropriate for a particular type of device of a certain minimum weight running at a certain minimum speed in public spaces (or other spaces where they typically get into contact with persons involved with the operation), this will normally be the case irrespective of whether the device is human-driven or AI-driven. For instance, large cleaning machines, lawnmowers or delivery vehicles in public spaces might generally have to be included in strict liability regimes even where, in the relevant jurisdiction, this is so far not the case. So a strict liability regime should, at the end of the day, not be restricted to AI systems.

9.3.2 Vicarious operator liability

Maybe even more important for AI liability would be a type of vicarious liability. Liability for acts or omissions of other persons, such as human auxiliaries, differ vastly across the legal systems of the Member States. A new European scheme of vicarious liability might restrict itself to ensuring that a principal that employs AI for a sophisticated task faces the same liability as a principal that employs a human auxiliary.[143] However, the EU legislator could also go one step further and introduce a fully harmonised concept of vicarious liability that does not suffer from the outset from the shortcomings we see in existing national concepts.

By and large, this new European scheme of vicarious liability could provide that a business or public authority is liable for damage caused by its human auxiliaries acting within the scope of their functions, or any AI employed by the business or public authority, where these auxiliaries or AI fail to perform – for whatever reason – at the standard that could reasonably be expected from them.[144] This comes close to strict liability insofar as it requires neither fault nor a defect (or general lack of reliability in the case of human auxiliaries), but some output that does not meet the standards of conduct to be expected

[143] *Wendehorst/Duller*, Safety and Liability (fn. 12) 92.

[144] This would amount to a combination between Article 6:102 (Liability for auxiliaries) and Article 4:202 (Enterprise Liability) of the Principles of European Tort Law (PETL) prepared by the European Group on Tort Law <http://egtl.org/PETLEnglish.html>.

from a business or public authority in the fulfilment of their functions. What this level of quality is depends on the task to be fulfilled.

Illustration 69

Bank B uses a credit scoring algorithm for assessing the creditworthiness of customers seeking credit, such as C. It is B's duty to provide proper assessment along the lines of any criteria prescribed by the law or stated by the business. It should be immaterial for purposes of liability whether B uses the services of a human employee or whether B deploys an AI system for the task.

However, it also transpires that vicarious AI liability can only go as far as the operator of the AI would itself be liable, under national law, for violation of the same standard of conduct. This means that there must exist some statutory or contractual duty of care on the part of the operator.

Such duty could also follow from the AIA. It is in particular the engagement in prohibited AI practices that should lead to liability, irrespective of whether the operator was acting intentionally or negligently with regard to the fact that, e.g., the AI was exploiting age-specific vulnerabilities. With an associated liability scheme in mind, it becomes even more apparent, though, that the very 'pointillistic' style of Title II of the AIA Proposal is a problem and that, if fundamental rights protection is taken seriously, it would have been necessary to have a more complete list of blacklisted AI practices (above at 5.1 and 5.2) plus ideally a general clause to cover unforeseen cases (above at 5.3).

Whether vicarious liability for AI would be included in a separate legal instrument on AI liability or in the AIA itself would be of secondary importance. If included in the AIA itself, the provision could be phrased as follows:

Article 72a
Vicarious liability for AI systems

1. A user of an AI system shall be liable for harm caused by any lack of accuracy or other shortcoming in the operation of the system to the same extent as that user would be liable for the acts or omissions of a human employee mandated with the same task as the AI system.

2. Where a human employee would not have been able to fulfil the task fulfilled by the AI system (such as where the task requires computing capabilities exceeding

those of humans) the point of reference for determining the required level of performance is available comparable technology which the user could be expected to use.

9.3.3 Defect Liability for AI

Meanwhile, it is widely accepted that the PLD must in any case be adapted to the challenges of digital ecosystems at large.[145] While the majority of changes may become necessary due to developments not strictly associated with AI, it may be advisable to include also specific rules on AI in order to take into account the AI-specific problems a victim may have in proving a defect.

9.3.3.1 Lack of 'safety' in a traditional sense

The debate about a reform of the PLD has so far been focused entirely on physical risks. Because of the difficulties a victim has to show that an AI system was defective, no defect of the AI should have to be established by the victim for AI-specific harm caused by AI-driven products. Rather, it should be sufficient for the victim to prove that the harm was caused by an incident that might have something specifically to do with the AI (e.g. the cleaning robot making a sudden move in the direction of the victim) as contrasted with other incidents (e.g. the victim stumbling over the powered-off cleaning robot).[146]

9.3.3.2 Lack of 'fundamental rights safety'

However, defect liability should be made fully operational also for fundamental rights risks that come with AI. In order to achieve this, the first step must be to formulate an equivalent to the established concept of 'safety' in traditional product safety legislation. As far as traditional safety risks are concerned, it is possible for Article 6(1) of the PLD simply to state: 'A product is defective when it does not provide the safety which a person is entitled to expect, taking all circumstances into account,…'. In a similar vein, Article 2(b) of the General Product Safety Directive (GPSD) defined as a 'safe product' any product

[145] Among the plethora of pleas made in this direction, see only C Twigg-Flesner in European Law Institute (ELI) (ed), *Guiding Principles for Updating the Product Liability Directive for the Digital Age* (2021) <https://europeanlawinstitute.eu/fileadmin/user_upload/p_eli/Publications/ELI_Guiding_Principles_for_Updating_the_PLD_for_the_Digital_Age.pdf>.

[146] *Wendehorst/Duller*, Safety and Liability (fn. 12)6, 93.

which, ‘under normal or reasonably foreseeable conditions of use…does not present any risk or only the minimum risks compatible with the product's use, considered to be acceptable and consistent with a high level of protection for the safety and health of persons…’. However, due to the ‘fuzzy’ nature of the risks there does not currently exist any general definition of ‘fundamental rights safety’ of an AI system that might be made operational for purposes of liability law.

So among the requirements listed in Chapter 2 of Title III those requirements should be identified which constitute ‘AI-specific safety’ (which would, by and large, be the requirements listed in Articles 13 through 15 of the draft AIA). Those requirements should – at least for the purposes of product liability – be clearly separated from the requirements that are about managing safety (mostly Article 9), ensuring safety (selected aspects of which are listed in Article 10) or documenting safety (Articles 11 and 12).

Illustration 70

Bank B uses a credit scoring algorithm provided by provider P for assessing the creditworthiness of customers seeking credit, such as C. Due to significant bias (the system qualifies all customers holding more than three mobile phone contracts as not creditworthy) C is offered credit only at an interest rate of 6 % p.a. while the normal interest rate would have been 2 % p.a. (C holds four mobile phone contracts, one for himself and three for his children), causing financial harm to C of EUR 20,000.

While B should in any case be liable under vicarious liability (see 9.3.2), also P should be liable to C under a revised product liability regime that includes liability where a lack of ‘fundamental rights safety’ has resulted in material loss. As it would be unduly onerous for C having to prove that he would have been offered cheaper credit if the system had worked properly the burden of proof with regard to causation should shift to the defendant once a prima facie case has been established that a system falls short of the requirements in Articles 13 to 15.

Illustration 71

Company C uses recruitment software provided by provider P for shortlisting applicants for a vacant position in upper management. The software turns out to have a significant gender bias, systematically grading down female applicants as compared with male applicants. Female applicant F can demonstrate that she is as qualified as male applicant M, and that M ended up in the middle of the shortlist, while F was discarded. However, at the end of the day, M does not get the positon either, because T is much better qualified than both M and F.

In this situation, the system would lack the 'fundamental rights safety' required under the AIA (leading, inter alia, to very poor accuracy within the meaning of Article 15), and the burden of proof with regard to causation should shift to the defendant. However, both C and P can rebut the presumption as F would not have been hired in any case. So F can only exercise rights against C under applicable non-discrimination law, but has no rights against P.

Illustration 72

University U uses an AI system for the assessment of participants in admission exams. Applicant A is discarded by the system (because of her poor performance), which is why A loses a full year as well as income of estimated EUR 30,000. Angry and frustrated, A makes investigations and finds out that the technical documentation for the system does not fully comply with the requirements of Article 11 AIA with Annex IV, which is why A claims damages for the harm suffered.

In this situation, there would likewise be no causation with regard to A's loss. But unlike in illustration 71 there would not have been a problem with the 'fundamental rightssafety' of the system as such, so there should not be any presumption of causation and the burden of proof should not shift to U.

Whether this scheme would then be included in a revised product liability regime or in the AIA itself would be of secondary importance. If included in the AIA itself, the provision could be phrased as follows (partly borrowing from similar wording in Article 82 GDPR):

Article 72b
Right to compensation and liability

1. Where non-compliance of a party with any obligations following from Titles II, IIa, III or IV of this Regulation has resulted in an increased risk for the safety or fundamental rights of a person, and where that person has suffered economic or non-economic harm *[Opt: material or non-material damage]* because the risk has materialised, the person shall have the right to receive compensation from the party who failed to comply with its obligations.

2. Where a high-risk AI system fails to comply with the requirements set out in Articles 13 to 15 and the harm suffered is of a kind typically resulting from such non-compliance there shall, for the purposes of liability under paragraph 1, be a presumption that the non-compliance has caused the harm.

3. A party who has failed to comply with its obligations shall be exempt from liability under paragraph 1 if it proves that it is not in any way responsible for the non-compliance.

4. Where more than one party has failed to comply with their obligations and is liable under paragraph 1, each party shall be held liable for the entire damage in order to ensure effective compensation of the affected person. Where a party has paid full compensation for the damage suffered that party shall be entitled to claim back from the other liable parties that part of the compensation corresponding to their part of responsibility for the damage.

10 Enforcement

The relevant rules on enforcement of the AIA are put down in Title VIII, Chapter 3 of the AIA Proposal.

10.1 Enforcement of compliance with the AIA

10.1.1 Enforcement modelled on product safety and market surveillance law

According to Article 63, and in line with the product safety law approach (above at 2.1), enforcement of the AIA is generally subject to the regime of the Market Surveillance Regulation (MSR).[147] Operators identified in Title III Chapter 3 of the AIA (i.e. providers, distributors, users) are to be understood as economic operators within the meaning of the MSR, and any AI system within the meaning of the AIA Proposal is to be understood as a 'product' within the meaning of the MSR. Furthermore, Article 65 (1) clarifies that the notion of 'risk' for purposes of applying the MSR does not only cover risks to the health and safety of persons, but also fundamental rights risks (see above at 2.3 and at 7.2).

For high-risk AI systems to which NLF legislation listed in Section A of Annex II applies, market surveillance authorities in charge of supervising compliance with the AIA are the authority responsible for market surveillance activities designated under the legal acts listed. For AI systems placed on the market, put into service or used by financial institutions regulated by Union legislation on financial services, the market surveillance authority in charge shall be the relevant authority responsible for the financial supervision of those institutions under that legislation. For biometric identification systems and other AI systems used for law enforcement purposes or for immigration or asylum matters, Member States shall designate as market surveillance authorities for the purposes of the AIA either the competent data protection supervisory authorities under the LED or GDPR or the national competent authorities supervising the activities of the law enforcement,

[147] Regulation (EU) 2019/1020 of the European Parliament and of the Council of 20 June 2019 on market surveillance and compliance of products and amending Directive 2004/42/EC and Regulations (EC) No 765/2008 and (EU) No 305/2011, OJ L 169, 25.6.2019, p. 1–44

immigration or asylum authorities putting into service or using those systems. For Union institutions, agencies and bodies, the European Data Protection Supervisor shall act as their market surveillance authority.

Enforcement procedures at national as well as at Union level are dealt with in Articles 65 and 66 AIA Proposal.

10.1.2 Absence of individual or collective redress mechanisms?

Consumer organisations have voiced criticism as to the absence of individual or collective redress mechanisms of consumers in the AIA.[148]

As far as the absence of individual redress is concerned this is simply the consequence of the absence of individual rights, as far as such individual rights are not created by way of 'transmission links' under national law, such as where the requirements under the AIA are implied in a contract with regard to AI, or where damages are sought based on the doctrine of Schutzgesetzverletzung or similar doctrines (see above at 8.1.1).

As far as collective redress is concerned, the AIA Proposal did not come with an explicit proposal to add the AIA to Annex I of the Representative Actions Directive (RAD)[149]. However, it is questionable whether such an explicit proposal is necessary. Annex I (8) RAD refers to 'Directive 2001/95/EC of the European Parliament and of the Council of 3 December 2001 on general product safety (OJ L 11, 15.1.2002, p. 4): Articles 3 and 5', i.e. to the General Product Safety Directive (GPSD). This Directive will soon be replaced by the General Product Safety Regulation (GPSR),[150] and references to the GPSD will be replaced by references to the GPSR. While the AIA Proposal, in Article 65 (1), only refers to the MSR, and not to the GPSD or future GPSR, it is not far-fetched to assume that the integration of AI systems into product safety law extends to the application of the future GPSR and therefore to the RAD. However, this should definitely be clarified.

---

[148] See vzbv (fn. 68) p. 27; BEUC (fn. 68), p. 23 f.

[149] Directive (EU) 2020/1828 of the European Parliament and of the Council of 25 November 2020 on representative actions for the protection of the collective interests of consumers and repealing Directive 2009/22/EC, OJ L 409, 4.12.2020, p. 1–27.

[150] Proposal for a Regulation of the European Parliament and of the Council on general product safety, amending Regulation (EU) No 1025/2012 of the European Parliament and of the Council, and repealing Council Directive 87/357/EEC and Directive 2001/95/EC of the European Parliament and of the Council, COM(2021) 346 final.

Should the author's suggestions to include in the AIA provisions on individual rights (see above at 4) and/or on liability (above at 9) be adopted, there would automatically be (a need for) private enforcement, and it would be close to inevitable to explicitly list the relevant rights in Annex I to the RAD.

10.2 Combating risks beyond compliance with the AIA

Given that the provisions in the AIA have a limited scope (with large parts of the AIA being restricted to high-risk AI systems) and may not serve to prevent all sorts of possible risks presented by AI systems, it is important to have mechanisms in place to ensure safety and compliance beyond the AIA.

10.2.1 Enforcement of compliance with other law

Article 67 AIA deals with AI systems that comply with the requirements under the AIA but nevertheless present a risk, e.g., because they lead to prohibited discrimination. Where a market surveillance authority in charge finds that although an AI system is in compliance with the AIA it presents a risk to the health or safety of persons, to the compliance with obligations under Union or national law intended to protect fundamental rights or to other aspects of public interest protection, it shall require the relevant operator to take all appropriate measures to ensure that the AI system concerned, when placed on the market or put into service, no longer presents that risk, to withdraw the AI system from the market or to recall it within a reasonable period, commensurate with the nature of the risk, as it may prescribe. The provider or other relevant operators shall ensure that corrective action is taken, and the Member State whose national authority has detected the lack of safety or compliance shall immediately inform the Commission and the other Member States. The Commission can then, by way of a decision addressed at Member States, take appropriate measures.

10.2.2 Management of systemic risks: borrowing from the DSA

What seems to be clearly missing, however, is more specific provisions on systemic risks posed by certain AI systems. While Articles 9 (risk-management systems), 17 (quality management systems) and 61 (post-market monitoring system) are rather broad with regard to risks covered, none of them seems to address specifically risks that are not

inherent in the AI system as such, but that result only from, or are significantly enhanced by, widespread use of the AI system and its interaction with other systems.

Illustration 73

Any deficiencies in recruitment software provided by provider P may cause harm to individuals, but certain deficiencies may be unavoidable even where all requirements under the AIA are met, and they may ultimately be acceptable in the light of the existing deficiencies in human decision-making. However, this may change once 70 % of bigger companies use that software (or other software developed on the basis of that software), because any minor deficiencies in the system (such as its tendency to score down applicants with certain traits that do not correlate with a protected characteristic under non-discrimination law, see above illustration 55) will suddenly have huge impact on the labour market, drastically reduce diversity in companies, and may (e.g. if taken up by service providers advising applicants) have a significant impact on people's behaviour. Such effects may only become visible many years after the AI system has been introduced, when effects are already largely irreversible, or they may not be detected at all.

It is therefore recommended to insert, in Title VIII, on post-market monitoring, information sharing and market surveillance, a new Chapter 2a, which would be modelled on Articles 25 ff. of the proposed DSA and could be phrased along the lines of the following:

CHAPTER 2A

ADDITIONAL OBLIGATIONS FOR VERY LARGE PROVIDERS TO MANAGE SYSTEMIC RISKS

Article 62a
Very large providers

1. This Chapter shall apply to providers of high-risk AI systems listed in Annex III for which both of the following conditions are fulfilled:

(a) the provider has a share of […] percent or above in the market for AI systems of the relevant type, considering the AI system's core functionalities, with regard to the whole Union, or a share of […] percent or above in the relevant market in at least three Member States; and

(b) […] percent or above of decision-making of the relevant kind listed in Annex III significantly relies on the use of that type of AI system.

When calculating the share within the meaning of point (a), AI systems that are not placed on the market or put into service under the provider's own name or trademark, but that use the provider's AI system as a basis or component in a way that significantly influences any systemic risks presented by those AI systems, shall be included.

2. The Commission shall adopt delegated acts in accordance with Articles 73 and 74, after consulting the Board, to lay down a specific methodology for calculating the market share referred to in paragraph 1. In those delegated acts, the Commission may also define different percentages than referred to in paragraph 1 for particular high-risk AI systems where there is reason to believe that systemic risks resulting from that type of AI system are significantly higher or lower than for other AI systems listed in Annex III.

3. The Board shall verify, at least once a year, whether the market shares of providers whose AI systems are used in the Union is equal to or higher than the shares referred to in paragraphs 1 and 2. On the basis of that verification, it shall adopt a decision designating the provider as a very large provider for the purposes of this Regulation, or terminating that designation, and communicate that decision, without undue delay, to the provider concerned and to the Commission.

4. The Commission shall ensure that the list of designated very large providers is published in the Official Journal of the European Union and keep that list updated. The obligations of this Chapter shall apply, or cease to apply, to the very large providers concerned from four months after that publication.

Article 62b
Systemic risk assessment

1. As part of the quality management system referred to in Article 17 and post-market monitoring system referred to in Article 61, very large providers shall identify, analyse and assess, at least once a year, any significant systemic risks stemming from the functioning and use made of the AI systems provided by them in the Union.

2. This risk assessment shall be specific to the AI systems they provide and shall, in any case, include the following systemic risks:

(a) any negative effects for the exercise of fundamental rights, for example respect for private and family life, data protection, the prohibition of discrimination, the rights of the child and access to an effective remedy and a fair trial, as enshrined in Articles 7, 8, 21, 24 and 47 of the Charter respectively;

(b) any negative effects for democracy, the rule of law, the functioning of state institutions, the stability of societies and economies, protection of the

environment and the combat against climate change, and other important public interests;

(c) any risks resulting from uniformity of decision-making, including for the emergence of new disadvantaged groups, the reduction of diversity in affected groups (e.g. recruited individuals), and a steering function for human behaviour as affected individuals adapt their behaviour to the parameters relied on by the AI system;

(d) any risks resulting from a reduction in human skills and competences, including for the ability to detect and correct errors and to act independently of the AI system where the system is unavailable;

(e) risks of intentional manipulation of their AI system, including by means of targeted inauthentic behaviour of affected persons, malicious interferene by third parites, or hybrid warfare, with an actual or foreseeable negative effect on important public or private interests.

Article 62c
Mitigation of systemic risks

1. Very large providers shall put in place reasonable, proportionate and effective mitigation measures, tailored to the specific systemic risks identified pursuant to Article 62b. Such measures may include, where applicable:

(a) adapting AI systems, their decision-making processes, their features or functioning, or the instructions and specificaitons accompanying them;

(b) reinforcing the internal processes or supervision of any of their activities in particular as regards detection of systemic risk;

(c) …

2. The Board, in cooperation with the Commission, shall publish comprehensive reports, once a year, which shall include the following:

(a) identification and assessment of the most prominent and recurrent systemic risks reported by very large providers or identified through other information sources;

(b) best practices for very large providers to mitigate the systemic risks identified.

3. The Commission, in cooperation with the Board, may issue general guidelines on the application of paragraph 1 in relation to specific risks, in particular to present best practices and recommend possible measures, having due regard to the possible consequences of the measures on fundamental rights enshrined in the Charter of all parties involved. When preparing those guidelines the Commission shall organise public consultations.

Article 62d
Independent audit

1. Very large providers shall be subject, at their own expense and at least once a year, to audits to assess compliance with the following:

(a) the obligations set out in Chapter 3 of Title III;

(b) any commitments undertaken pursuant to the codes of conduct referred to in Article 69.

2. Audits performed pursuant to paragraph 1 shall be performed by organisations which:

(a) are independent from the very large providers concerned;

(b) have proven expertise in the area of risk management, technical competence and capabilities;

(c) have proven objectivity and professional ethics, based in particular on adherence to codes of practice or appropriate standards.

3. The organisations that perform the audits shall establish an audit report for each audit. The report shall be in writing and include at least the following:

(a) the name, address and the point of contact of the very large provider subject to the audit and the period covered;

(b) the name and address of the organisation performing the audit;

(c) a description of the specific elements audited, and the methodology applied;

(d) a description of the main findings drawn from the audit;

(e) an audit opinion on whether the very large provider subject to the audit complied with the obligations and with the commitments referred to in paragraph 1, either positive, positive with comments or negative;

(f) where the audit opinion is not positive, operational recommendations on specific measures to achieve compliance.

4. Very large providers receiving an audit report that is not positive shall take due account of any operational recommendations addressed to them with a view to take the necessary measures to implement them. They shall, within one month from receiving those recommendations, adopt an audit implementation report setting out those measures. Where they do not implement the operational recommendations, they shall justify in the audit implementation report the reasons for not doing so and set out any alternative measures they may have taken to address any instances of non-compliance identified.

Article 62e

Transparency reporting obligations for very large providers

1. Very large providers shall make publicly available and transmit to the Board and the Commission, at least once a year and within 30 days following the adoption of the audit implementing report provided for in Article 62d(4):

(a) a report setting out the results of the risk assessment pursuant to Article 62b;

(b) the related risk mitigation measures identified and implemented pursuant to Article 62c;

(c) the audit report provided for in Article 62d(3);

(d) the audit implementation report provided for in Article 62d(4).

3. Where a very large provider considers that the publication of information pursuant to paragraph 2 may result in the disclosure of confidential information of that provider or of the users of the AI system, may cause significant vulnerabilities for the security of its AI system, may undermine public security or may harm users or affected individuals, the provider may remove such information from the reports. In that case, that provider shall transmit the complete reports to the Board and the Commission, accompanied by a statement of the reasons for removing the information from the public reports.

Article 62f

Data access and scrutiny by vetted researchers

1. Upon a reasoned request from the Commisison, very large providers shall, within a reasonable period, as specified in the request, provide access to data to vetted researchers who meet the requirements in paragraphs 3 of this Article, for the sole purpose of conducting research that contributes to the detection, identification and understanding of systemic risks in the Union as set out in Article 62b(1), including as regards the adequacy, efficiency and impacts of the risk mitigation measures pursuant to Article 62c. In making a request, the Commission shall take due account of the rights and interests of the providers and users of the AI system concerned, including the protection of personal data, the protection of confidential information, in particular trade secrets, and maintaining the security of their AI systems.

2. Very large providers shall facilitate and provide access to data pursuant to paragraph 1 through appropriate interfaces specified in the request, including online databases or application programming interfaces.

3. Upon a duly substantiated application from researchers, the Commission shall award them the status of vetted researchers and issue data access requests pursuant to paragraph 1, where the researchers demonstrate that they meet all of the following conditions:

(a) they are affiliated to a research organisation as defined in Article 2 (1) of Directive (EU) 2019/790 of the European Parliament and of the Council;

(b) they are independent from commercial interests;

(c) they are in a capacity to preserve the specific data security and confidentiality requirements corresponding to each request and to protect personal data, and they describe in their request the appropriate technical and organisational measures they put in place to this end;

(d) the application submitted by the researchers justifies the necessity and proportionality for the purpose of their research of the data requested and the timeframes within which they request access to the data, and they demonstrate the contribution of the expected research results to the purposes laid down in paragraph 1;

(e) the planned research activities will be carried out for the purposes laid down in paragraph 1;

(f) they carry their activities according to the procedures laid down in delegated acts referred to in paragraph 7;

(g) they have not already filed the same application with the Commission.

4. The Commission shall issue a decision terminating the access if it determines, following an investigation either on its own initiative or on the basis information received from third parties, that the vetted researcher no longer meets the conditions set out in paragraph 3. Before terminating the access, the Commission shall allow the vetted researcher to react to the findings of its investigation and its intention to terminate the access.

5. Upon completion of the research envisaged in paragraph 1, the vetted researchers shall make their research results available to the Commission free of charge. The Commission may make the research results publicly available, taking due account of the rights and interests of the providers and users of the AI system concerned, including the protection of personal data, the protection of confidential information, in particular trade secrets, and maintaining the security of their service.

6. The Commission shall, after consulting the Board, adopt delegated acts laying down the technical conditions under which providers of very large providers are to share data pursuant to paragraphs 1 and 2 and the purposes for which the data may be used. Those delegated acts shall lay

down the specific conditions and relevant objective indicators, as well as procedures under which such sharing of data with vetted researchers can take place in compliance with Regulation (EU) 2016/679, taking into account the rights and interests of the providers and users of the AI system concerned, including the protection of confidential information, in particular trade secrets, and maintaining the security of their AI system.

10.3 Confidentiality and essential public interests

According to Article 64, the market surveillance authorities in the Member States shall be granted full access to the training, validation and testing datasets used by the provider, including through application programming interfaces ('API') or other appropriate technical means and tools enabling remote access. Where necessary to assess the conformity of the high-risk AI system with the AIA requirements, a market surveillance authority shall even be granted access to the source code of the AI system. Further documentation may be requested and accessed.

Even though Article 64 (6) clarifies that any information and documentation obtained by the national public authorities or bodies shall be treated in compliance with the confidentiality obligations set out in Article 70 it is clear that, in particular with regard to AI systems used for law enforcement or similar purposes, or used to control critical infrastructure, it is next to impossible to shield extremely sensitive information from malicious third parties, including organised crime and/or third country intelligence services.

Illustration 74

Provider P provides an AI system used as safety component in the management and operation of power grids. Among the features of the AI system is the early detection of irregularities that could indicate a beginning attack on the power grid system that is essential for power supply in several Member States. If P has to provide all data and give insight into the source code to, potentially, national authorities in 27 Member States (assuming that P has some minimum contacts

with many Member States that would theoretically justify a request) the probability that an official in one of these authorities will be prepared to leak this information to malicious third parties, such as terrorist organisations or third countries collecting information for purposes of hybrid warfare, is very high.

Illustration 75

Provider P provides an AI system to be used by law enforcement authorities for predicting the occurrence or reoccurence of actual or potential drug dealing offences, based on profiling. Again, if P has to provide all data and give insight into the source code to, potentially, national authorities in 27 Member States, the probability that an official in one of these authorities will leak information to organised crime, enabling organised crime to avoid detection, is rather high.

This is why it is recommended to enhance the confidentiality requirements and to insert a clause protecting essential public interests, e.g.

Article 70a
Exceptions for AI systems with enhanced confidentiality requirements

1. A provider of a high-risk AI system that is confronted with a request by a competent national authority for information, documentation, access to data, disclosure of the source code or a similar measure under this Regulation may refuse to comply with the request if that provider can demonstrate that the relevant materials would, if disclosed to unauthorised parties, jeopardise public and national security interests.

2. A provider relying on paragraph 1 shall immediately notify the Commission of the refusal to comply with the request and the reasons of the refusal. The Commission shall, upon having investigated the matter, issue a decision addressed at the relevant national authority and the provider. In that decision, the Commission may provide that only Commission staff holding the appropriate level of security clearance shall be allowed to access the relevant materials, and impose further restrictions and safeguards as appropriate.

ANNEX: Consolidated Recommended Amendments to the AIA Proposal

Recitals

It is recommended to clarify in the Recitals that the notion of 'fundamental rights risks' may include economic risks and risks for society at large. 'Fundamental rights' are often understood as specifically meaning individual rights listed in the Charter (see 7.2.2 and 7.2.3), which might give rise to the misunderstanding that risks such as fraud or the undermining of democratic elections are not covered. From a consumer perspective, the inclusion of economic risks and societal risks is definitely of key importance.

[…] Whereas: […]

(32) As regards stand-alone AI systems, meaning high-risk AI systems other than those that are safety components of products, or which are themselves products, it is appropriate to classify them as high-risk if, in the light of their intended purpose, they pose a high risk of harm to the health and safety or the fundamental rights of persons, taking into account both the severity of the possible harm and its probability of occurrence and they are used in a number of specifically pre-defined areas specified in the Regulation. The identification of those systems is based on the same methodology and criteria envisaged also for any future amendments of the list of high-risk AI systems. **The notion of fundamental rights risks is understood very broadly and not restricted to risks for rights explicitly mentioned under a separate Article in the Charter as long as there is a sufficient link between the risk and enjoyment of rights under the Charter. Notably, the notion of fundamental rights risk may, depending on the context, include mere economic risks where those risks are sufficiently severe, for instance because they affect access to essential goods or services (such as energy supply, credit or insurance) or because they operate on a large scale and may significantly affect the general standard of living of a natural person (such as large scale personalised pricing). The notion of fundamental rights risks also includes risks for society at large, such as for democratic institutions and a fair and open discourse.**

[…]

Definitions

It is important to detach the definitions of 'emotion recognition system' and 'biometric categorisation system' from the definition of 'biometric data', which has been copied from the GDPR and requires that the data allow or confirm the unique identification of a natural person. Emotion recognition and biometric categorisation, however, do not (necessarily) rely on personal data that allow or confirm the unique identification of a particular individual. It is therefore recommended to introduce a separate definition of 'biometrics-based data' (see 6.3.1).

It is also important to modify the notion of 'real-time' in the context of remote biometric identification because the pivotal point is not so much the duration of delay between capturing of live templates and identification but rather whether identification occurs on a large scale over a period of time (see 6.2.1.3). Where this is not the case and identification is just limited to a particular past incident, such as a crime captured by a video camera, we may not need the same strict regulation as for real-time remote identification.

Further suggestions as to concrete formulations have been made as stated above, but these are of lower priority.

Article 3
Definitions

For the purpose of this Regulation, the following definitions apply:

[…]

(33) 'biometric data' means personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic data;

(33a) 'biometrics-based data' means personal data resulting from specific technical processing relating to physical, physiological or behavioural signals or characteristics of a natural person, such as facial expressions, movements, pulse frequency, voice, keystrokes or gait, which may or may not allow or confirm the unique identification of a natural person;

(34) 'emotion recognition system' means an AI system for the purpose of identifying or inferring emotions**, thoughts** or intentions of natural persons on the basis of their ~~biometric~~**biometrics-based** data;

(35) 'biometric categorisation system' means an AI system for the purpose of assigning natural persons to specific categories such as sex, age, hair colour, eye colour, tattoos, ethnic origin**, health, mental ability, personality traits** or sexual or political orientation, on the basis of their ~~biometric~~**biometrics-based** data;

(36) ‘remote biometric identification system’ means an AI system for the purpose of identifying natural persons at a distance through the comparison of a person’s biometric data with the biometric data contained in a reference database, and without **the conscious cooperation of the persons to be identified** ~~prior knowledge of the user of the AI system whether the person will be present and can be identified~~ ;

(37) ‘‘real-time’ remote biometric identification system’ means a remote biometric identification system whereby the capturing of biometric data, the comparison and the identification ~~all~~ occur **on a continuous or large-scale basis over a period of time and without limitation to a particular past incident (such as a crime recorded by a video camera);** ~~without a significant delay. This comprises not only instant identification, but also limited short delays in order to avoid circumvention.~~

[…]

List of prohibited AI practices in Title II

It is recommended to broaden the scope of the three existing per se-prohibitions – manipulation by subliminal techniques, exploitation of vulnerabilities and social scoring – in several ways (see 5.1.1.3, 5.1.2.4 and 5.1.3.3). In particular, it is recommended to replace ‘physical or psychological harm’ by ‘material and unjustified harm’, both with the aim of including economic harm and of avoiding overreaching effects (see above the modifications in points (a) and (b)).

It is likewise recommended to extend the prohibition of exploitation of vulnerabilities from group-specific vulnerabilities to individual vulnerabilities, e.g. very individual personality traits discovered with the help of data analytics (see above point (b)), and to remove the restriction to public authorities in the prohibition of social scoring with a view to extending it to social scoring conducted by private actors, e.g. by a provider of a gatekeeper platform service (see above the modification in point (c)).

In terms of AI practices missing in the list of prohibited practices, it is recommended to add total surveillance (see 5.2.1) and violation of mental privacy and integrity (see 5.2.2) (see above new points (ba) and (bb)).

In any case, it is recommended to clarify that Article 5 needs to be seen in the context of a host of prohibitions following from other law, which apply irrespective of whether AI is involved or not (see above paragraph 1a), and to allow for flexibility by empowering the Commission to update the list of prohibited AI practices by way of delegated acts (see above paragraph 1b)).

TITLE II

PROHIBITED ARTIFICIAL INTELLIGENCE PRACTICES

Article 5

1. The following artificial intelligence practices shall be prohibited:

(a) the placing on the market, putting into service or use of an AI system that deploys subliminal techniques beyond a person's consciousness in order to materially distort a person's behaviour in a manner that causes or is likely to cause that person or another person ~~physical or psychological~~ **material and unjustified** harm;

(b) the placing on the market, putting into service or use of an AI system that exploits any of the vulnerabilities of

(i) a specific group of persons due to their age, physical or mental disability **or social or economic situation; or**

(ii) an individual whose vulnerabilities are characteristic of that individual's known or predicted personality or social or economic situation

~~in order to materially distort the behaviour of a person pertaining to that group~~ in a manner that causes or is likely to cause that person or another person ~~physical or psychological~~ **material and unjustified** harm;

(ba) the putting into service or use of an AI system for the comprehensive surveillance of natural persons in their private or work life to an extent or in a manner that causes or is likely to cause those persons or another person material and unjustified harm;

(bb) the placing on the market, putting into service or use of an AI system for the specific technical processing of brain data in order to read or manipulate a person's thoughts against that person's will or in a manner that causes or is likely to cause that person or another person material and unjustified harm.

(c) the placing on the market, putting into service or use of AI systems ~~by public authorities or on their behalf~~ for the evaluation or classification ~~of~~ the trustworthiness of natural persons over a certain period of time based on their social behaviour or known or predicted personal or personality characteristics, with the social score leading to either or both of the following:

(i) detrimental or unfavourable treatment of certain natural persons or whole groups thereof in social contexts which are unrelated to the contexts in which the data was originally generated or collected;

(ii) detrimental or unfavourable treatment of certain natural persons or whole groups thereof that is unjustified or disproportionate to their social behaviour or its gravity;

(d) *[to be moved to new Article 5a]*

1a. In addition to the prohibited AI practices referred to in paragraph (1), AI practices referred to in Annex Ia shall also be considered prohibited. The Commission is empowered to adopt delegated acts in accordance with Article 73 to update the list in Annex Ia on the basis of a similar threat to fundamental rights and European values as posed by the practices listed in paragraph (1).

1b. Paragraphs (1) and (1a) are without prejudice to prohibitions that apply where an artificial intelligence practice violates other laws, including data protection law, non-discrimination law, consumer protection law, and competition law.

[Paragraphs 2 to 4 to be moved to new Article 5a]

Biometric techniques as 'restricted AI practices'

With regard to real-time remote biometric identification an entirely new regulatory approach has been suggested (see 6.2.2). As the provisions on real-time remote biometric identification do not resemble the per se-prohibitions in Article 5, but rather stipulate conditions for the use of these techniques, they should be moved to a separate Title IIa on 'Restricted AI practices'. As to the structure, the close interplay with Article 9 GDPR would become much clearer if the new provision were structured in a similar way and if explicit reference to the several justifications in Article 9 GDPR were made (see above paragraph 1). There should be a clarification that the new provisions do not in any way derogate basic principles of other laws, notably of the GDPR, such as that data must only be stored as far as strictly necessary to achieve the relevant law enforcement purpose (see above paragraph 5).

Given that real-time remote identification achieved with the help of other than biometric techniques (e.g. with the help of mobile phone signals) may be almost as problematic it could be an option to remove the restriction to biometric identification and include also other techniques of mass identification.

TITLE IIA

RESTRICTED ARTIFICIAL INTELLIGENCE PRACTICES

Article 5a
'Real-time' remote biometric [Opt.: or other] identification

1. **AI systems may be used for 'real time' remote biometric identification *[Opt.: or other 'real time' remote identification]* in publicly accessible spaces only when such surveillance is limited to what is strictly necessary for:**

(a) **the use for a specific purpose to which the persons identified have given their explicit consent within the meaning of Article 9 (2)(a) of Regulation (EU) 2016/679;**

(b) **the use for purposes and under conditions referred to in Article 9 (2)(b) and (j) of Regulation (EU) 2016/679;**

(c) **the use for migration, asylum or border control management;**

(d) the use ~~of 'real-time' remote biometric identification systems in publicly accessible spaces~~ for the purpose of law enforcement, ~~unless and~~ in as far as such use is strictly necessary for one of the following objectives:

(i) the targeted search for specific potential victims of crime, including missing children;

(ii) the prevention of a specific, substantial and imminent threat **to public security, in particular** to the life or physical safety of natural persons**,** or of a terrorist attack;

(iii) the detection, localisation, identification or prosecution of a perpetrator or suspect of a criminal offence referred to in Article 2(2) of Council Framework Decision 2002/584/JHA and punishable in the Member State concerned by a custodial sentence or a detention order for a maximum period of at least three years, as determined by the law of that Member State.

2. The use of 'real-time' remote [biometric] identification systems in publicly accessible spaces for the purpose~~s of law enforcement for any of the objectives~~ referred to in paragraph 1 point**s c) and** d) shall take into account the following elements:

(a) the nature of the situation giving rise to the possible use, in particular the seriousness, probability and scale of the harm caused in the absence of the use of the system;

(b) the consequences of the use of the system for the rights and freedoms of all persons concerned, in particular the seriousness, probability and scale of those consequences.

In addition, the use of 'real-time' remote [biometric] identification systems in publicly accessible spaces ~~for the purpose of law enforcement~~ for any of the objectives referred to in paragraph 1 point**s c) and** d) shall comply with necessary and proportionate safeguards and conditions in relation to the use, in particular as regards the temporal, geographic and personal limitations.

3. As regards paragraphs 1, point**s (c) and** (d) and 2, each individual use ~~for the purpose of law enforcement~~ of a 'real-time' remote [biometric] identification system in publicly accessible spaces shall be subject to a prior authorisation granted by a judicial authority or by an independent administrative authority of

the Member State in which the use is to take place, issued upon a reasoned request and in accordance with the detailed rules of national law referred to in paragraph 4. However, in a duly justified situation of urgency, the use of the system may be commenced without an authorisation and the authorisation may be requested only during or after the use.

The competent judicial or administrative authority shall only grant the authorisation where it is satisfied, based on objective evidence or clear indications presented to it, that the use of the 'real-time' remote biometric identification system at issue is necessary for and proportionate to achieving one of the objectives specified in paragraph 1, point**s (c) and** (d), as identified in the request. In deciding on the request, the competent judicial or administrative authority shall take into account the elements referred to in paragraph 2.

4. A Member State may decide to provide for the possibility to fully or partially authorise the use of 'real-time' remote [biometric] identification systems in publicly accessible spaces ~~for the purpose of law enforcement~~ within the limits and under the conditions listed in paragraphs 1, point**s (c) and** (d), 2 and 3. That Member State shall lay down in its national law the necessary detailed rules for the request, issuance and exercise of, as well as supervision relating to, the authorisations referred to in paragraph 3. Those rules shall also specify in respect of which of the objectives listed in paragraph 1, point**s (c) and** (d), including which of the criminal offences referred to in point **(d)** (iii) thereof, the competent authorities may be authorised to use those systems for the purpose of law enforcement.

5. Further requirements or restrictions following from other Titles of this Act or from other laws, in particular data protection law and non-discrimination law, remain unaffected. In any case, only such personal data may be collected through remote biometric identification as are strictly necessary to achieve the purpose stated in paragraph (1), and must be erased as soon as they are no longer necessary in relation to this purpose.

As emotion recognition systems and biometric categorisation systems pose a significant threat to fundamental rights, and as they are currently not covered by Article 9 GDPR (but only by Article 6 GDPR), it is recommended to establish for these techniques a regulatory regime similar to that of Article 9 GDPR (see 6.3.2.3). This regime could then also include biometric identification that does not qualify as real-time remote biometric identification. It has been demonstrated in some detail why an 'Article 9 regime' would not be overreaching, at least not if a de minimis-exception is added (see paragraph 2 above). If such a provision is introduced it might be advisable to integrate the provision on transparency obligations which is currently to be found in Article 52(2) AIA Proposal (see paragraph 3 above). There should also be a clarification that further requirements or

restrictions following from other Titles of the Act or from other law remain unaffected (see paragraph 4 above).

Article 5b
Other use of biometric techniques

1. Biometric identification systems not covered by Article 5a, emotion recognition systems and biometric categorisation systems may be used only when such use is limited to what is strictly necessary for:

(a) the use for a specific purpose to which the affected persons have given their explicit consent within the meaning of Article 9 (2)(a) of Regulation (EU) 2016/679;

(b) the use for purposes and under conditions referred to in Article 9 (2)(b), (c), (g), (h), (i) and (j) of Regulation (EU) 2016/679;

(c) the use for the purpose of law enforcement, migration, asylum or border control management in as far as purposes are proportionate to the aim pursued, respect the essence of the fundamental rights and interests affected and provide for suitable and specific measures to safeguard them.

2. Emotion recognition systems and biometric categorisation systems may also be used where processing of the personal data of the natural person concerned is otherwise based on a legal ground under Regulation (EU) 2016/679 and the data are used exclusively for triggering a reaction that can, by its very nature, not have a negative impact on that natural person's legitimate interests and fundamental rights, and the data are erased or fully anonymised instantaneously without leaving any trace to the identifiable natural person.

3. Users of AI systems within the meaning of paragraph (1) shall inform of the operation of the system the natural persons exposed thereto unless this is inconsistent with the purpose within the meaning of paragraph (1) for which the system is used.

4. Further requirements or restrictions following from other Titles of this Act or from other laws, in particular data protection law, non-discrimination law and consumer protection law, remain unaffected.

Although decision making will be addressed in more detail only in Part II of this Study, it is recommended to include, in the specific context of biometric techniques, a special rule on decision making based on biometric techniques. This rule would be without prejudice to Article 22 GDPR, but as the latter applies only to fully automated decisions without meaningful human intervention there is a conspicuous gap which should be filled. The proposed Article 5c combines elements of Article 22 GDPR and Article 14 (5) AIA Proposal but modifies the latter as it is problematic in several respects (see 6.4).

Article 5c
Decisions based on biometric techniques

1. No action or decision which produces legal effects concerning the person exposed to biometric identification, emotion recognition or biometric categorisation, or which similarly significantly affects that person, is taken by the user on the basis of the output from the system unless this has been verified by means that are independent from the system and that provide a degree of reliability and accuracy appropriate to the significance of the action or decision. In particular, emotion recognition systems and biometric categorisation systems must, as such, not be used as legal evidence that the natural person concerned has in fact had the emotions, thoughts or intentions recognised by the system or belongs in fact to the category assigned by the system

2. Further requirements or restrictions following from other Titles of this Act or from other laws remain unaffected.

List of high-risk AI systems in Annex III

It is recommended to extend the list in Annex III in several respects. First of all, Point 1 should be extended to biometric techniques in general and cover also emotion recognition systems where those systems are to be used for preparing decisions that may have legal effects or similarly significantly affect him or her (see 7.3.3.1). Minor amendments have also been suggested with regard to Point 4 in order to capture, e.g., social media harvesting in the employment context.

With regard to consumer interests, it is of utmost importance to add, in Point 5, a number of applications that imply a comparable fundamental rights risk as credit scoring does. These applications include individual risk assessment in the insurance context, customer rating according to complaint history and similar factors, and personalised pricing (see 7.3.2.2 and 7.3.2.3). With regard to the exception for small scale providers there should be a clarification that it includes only small scale providers who are at the same time the 'providers' (within the meaning of the AIA) of the relevant AI systems.

What is missing entirely in Annex III is AI systems intended for use by consumers. The AIA as it currently stands seems to assume that systems intended for consumers are covered by Article 6 (1) in conjunction with NLF product safety legislation. However, this is not necessarily the case as NLF product safety legislation fails to cover a number of high-risk AI systems, or may not subject them to third-party conformity assessment (see 7.3.1). This is why it is recommended to insert a new area, which could be titled 'Use by vulnerable

groups or in situations that imply vulnerability to fundamental rights risks' and that would include, for the time being, virtual assistants used for making important decisions (e.g. a shopping assistant, be it provided as a standalone digital service or embedded in devices such as a home assistant device or a smart fridge) and particular AI systems specifically intended for children (see above Point 5a).

ANNEX III
HIGH-RISK AI SYSTEMS REFERRED TO IN ARTICLE 6(2)

High-risk AI systems pursuant to Article 6(2) are the AI systems listed in any of the following areas:

1. Biometric **techniques** ~~identification and categorisation of natural persons~~:
   (a) AI systems intended to be used for the 'real-time' and 'post' remote biometric identification of natural persons;
   (b) **AI systems intended to be used for emotion recognition of natural persons where that recognition may lead to a decision that produces legal effects for the relevant natural person or similarly significantly affect him or her;**
2. Management and operation of critical infrastructure:
   (a) AI systems intended to be used as safety components in the management and operation of road traffic and the supply of water, gas, heating and electricity.
3. Education and vocational training:
   (a) AI systems intended to be used for the purpose of determining access or assigning natural persons to educational and vocational training institutions;
   (b) AI systems intended to be used for the purpose of assessing students in educational and vocational training institutions and for assessing participants in tests commonly required for admission to educational institutions.
4. Employment, workers management and access to self-employment:
   (a) AI systems intended to be used for recruitment or selection of natural persons, notably for advertising vacancies, screening or filtering applications, **or for** evaluating candidates ~~in the course of interviews or tests~~;
   (b) AI intended to be used for making decisions on promotion and termination of work-related contractual relationships, for task allocation and for monitoring and evaluating performance and behavior of persons in such relationships.
5. Access to and enjoyment of essential private services and public services and benefits**, including access to products**:
   (a) AI systems intended to be used by public authorities or on behalf of public authorities to evaluate the eligibility of natural persons for public assistance

benefits and services, as well as to grant, reduce, revoke, or reclaim such benefits and services;

(b) AI systems intended to be used

(i) to evaluate the creditworthiness of natural persons or establish their credit score~~,~~

(ii) to evaluate the behaviour of natural persons such as with regard to complaints or the exercise of statutory or contractual rights in order to draw conclusions for their future access to private or public services,

(iii) for making individual risk assessments of natural persons in the context of access to essential private and public services, including insurance contracts, or

(iv) for personalised pricing within the meaning of Article 6 (1) (ea) of Directive 2011/83/EU,

with the exception of AI systems put into service by small scale providers **of AI systems** for their own use;

(c) AI systems intended to be used to dispatch, or to establish priority in the dispatching of emergency first response services, including by firefighters and medical aid.

5a. Use by vulnerable groups or in situations that imply vulnerability to fundamental rights risks

(a) AI systems intended to be used by children in a way that may seriously affect a child's personal development, such as by educating the child in a broad range of areas not limited to areas which parents or guardians can reasonably foresee at the time of the purchase;

(b) AI systems, such as virtual assistants, intended to be used by natural persons for taking decisions with regard to their private lives that have legal effects or similarly significantly affect the natural persons;

[…]

10.3.1 Title IV with a new focus on individual rights

In this study, it is recommended not to leave individual rights entirely to the GDPR, first, because it is not likely that the GDPR will be changed in the near future and the gaps filled (e.g. with regard to AI systems merely recommending action to a human) and, second, because AI-related individual rights are rather misplaced in the GDPR. The reason for the latter is that these individual rights are not focussed on the processing of input data relating specifically to the affected data subject but on the output data, which may have

been generated with the help of (training etc.) data relating to very different data subjects, or with the help of non-personal data. This is why it is suggested to give Title IV of the AIA a new focus on individual rights in the context of AI systems that present either a transparency or a fairness risk (for details see 8.2.1), and also to rephrase the existing Article 52 on transparency obligations in the light of this new focus and clarify its application to social bots that merely generate content (see 8.2.2).

~~TRANSPARENCY~~ OBLIGATIONS FOR CERTAIN AI SYSTEMS POSING TRANSPARENCY OR FAIRNESS RISKS

Article 51a
Compliance with the obligations

1. This Title includes obligations for AI systems where one or both of the following conditions are fulfilled:

(a) use of the AI system involves a risk of confusion between AI system and humans, or their operations or activities, where such confusion might harm the legitimate interests of persons exposed to the AI system;

(b) use of the AI system leads to a decision with regard to a person that involves a material degree of evaluation or discretion and thus involves a fairness risk for the affected person.

2. The obligations of users of AI systems under this Title shall apply also to users who do not operate the AI system under their own authority but who solicit the services of another party using the AI system.

3. Providers of AI system whose intended use includes use within the meaning of paragraph 1 shall ensure that AI systems are designed and developed in such a way that users are able to comply with their obligations under this Title.

4. None of the provisions under this Title shall affect any prohibitions or restrictions for AI systems following from Title II or Title IIa or any requirements or obligations set out for high-risk AI systems in Title III of this Regulation.

Article 52
Transparency obligations for certain AI systems

1. **Users of an AI system that interacts with natural persons** ~~Providers~~ shall ensure ~~that AI systems intended to interact with natural persons are designed and developed in such a way~~ that natural persons are informed that they are interacting with an AI system, unless this is obvious from the circumstances and the context of use. ~~This obligation shall not apply to AI systems authorised by law to detect, prevent, investigate and prosecute criminal offences, unless those systems are available for the public to report a criminal offence.~~

2. **Users of an AI system that creates content or engages in [online] activities that are normally engaged in by natural persons ('bot') shall disclose that the content was created, or the [online] activities performed, by an AI system, unless the source of the content or [online] activities cannot reasonably be expected to matter to natural persons exposed thereto** ~~an emotion recognition system or a biometric categorisation system shall inform of the operation of the system the natural persons exposed thereto. This obligation shall not apply to AI systems used for biometric categorisation, which are permitted by law to detect, prevent and investigate criminal offences~~.

3. Users of an AI system that generates or manipulates image, audio or video content that appreciably resembles existing persons, objects, places or other entities or events and would falsely appear to a person to be authentic or truthful ('deep fake'), shall disclose that the content has been artificially generated or manipulated.

3a. Paragraphs 1, 2 and 3 ~~However, the first subparagraph~~ shall not apply where the use is authorised by law to detect, prevent, investigate and prosecute criminal offences or it is necessary for the exercise of the right to freedom of expression and the right to freedom of the arts and sciences guaranteed in the Charter of Fundamental Rights of the EU, and subject to appropriate safeguards for the rights and freedoms of third parties.

4. Paragraphs 1, 2 and 3 shall not **be read as legitimising the use of AI systems referred to beyond what is permitted by other law** ~~affect the requirements and obligations set out in Title III of this Regulation~~.

A central part of the revised Title IV should be two additional provisions that mirror and adapt Article 22 GDPR (right to scrutiny of individual decision-making, see 8.2.4) as well as the respective information duties in Articles 13 to 15 GDPR (right to explanation of individual decision-making, see 8.2.5) in a way tailored to the specific situation of AI-driven decision-making. Major benefits for affected persons would include that these individual rights do not only apply for fully automated decisions, but also to decisions recommended to humans, and that the right to receive an explanation would be much more explicit and include, in particular, the main parameters of decision-making and their relative weight as well as an easily understandable explanation of inferences drawn if the inference itself is a main parameter.

Article 52a
Scrutiny of individual decision-making

1. No decision which produces legal effects concerning a person, or which similarly significantly affects that person, is taken by the user on the basis of the output

from an AI system unless the appropriateness and fairness of this decision has been verified by means that are appropriate to the nature and the significance of the decision and the role of the AI system in the decision-making process.

2. Unless otherwise specified by Union or Member State law, verification within the meaning of paragraph 1 may, in particular, consist in meaningful scrutiny, before the decision is taken, by a natural person who is equipped with the appropriate

(a) abilities, training and decision-making authority;

(b) information with regard to the individual case; and

(c) safeguards against automation bias.

3. The user may replace ex-ante verification within the meaning of paragraphs 1 and 2 by equivalent other measures where the affected person has given explicit consent or where ex-ante verification is impossible or would cause unreasonable effort and is not strictly necessary for safeguarding the affected person's rights and freedoms and legitimate interests. Unless otherwise specified by Union or Member State law, such equivalent other measures may, in particular, consist in the right to

(a) obtain human intervention that satisfies the requirements under paragraph 2;

(b) provide additional information and express his or her point of view; and

(c) contest the decision with a meaningful chance of having it revised.

Article 52b
Explanation of individual decision-making

1. A decision which is taken by the user on the basis of the output from an AI system and which produces legal effects concerning a person, or which similarly significantly affects that person, shall be accompanied by a meaningful explanation of

(a) the role of the AI system in the decision-making process;

(b) the logic involved, the main parameters of decision-making, and their relative weight; and

(c) the input data relating to the affected person and each of the main parameters on the basis of which the decision was made.

For information on input data under point (c) to be meaningful it must include an easily understandable description of inferences drawn from other data if it is the inference that relates to a main parameter.

2. Paragraph 1 shall not apply to the use of AI systems

(a) that have only minor influence within the decision-making process;

(b) that are authorised by law to detect, prevent, investigate and prosecute criminal offences or other unlawful behaviour;

(c) for which exceptions from, or restrictions to, the obligation under paragraph 1 follow from Union or Member State law, which lays down appropriate other safeguards for the affected person's rights and freedoms and legitimate interests; or

(d) where the affected person has given explicit consent not to receive an explanation.

3. The explanation within the meaning of paragraph 1 shall be provided at the time when the decision is communicated to the affected person. However, the user may provide the explanation only at a later point upon the affected person's request, where providing the explanation immediately is not strictly necessary for safeguarding the affected person's rights and freedoms and legitimate interests, in which case the user shall inform the affected person of the right under this Article and how it can be exercised.

10.3.2 Liability

Liability for AI systems should not primarily be dealt with in the AIA itself, but be largely a matter for product liability law, national tort law and/or a new EU regime of AI liability. However, as these liability regimes are not well suited to address harm caused by fundamental rights risks, it is advisable to insert two provisions in the AIA itself, one on vicarious liability (see 9.3.2) and one on liability for lack of 'fundamental rights safety' (see 9.3.3). The former would help overcome existing uncertainties with regard to vicarious liability under national law (such as §§ 1313a, 1315 ABGB or §§ 278, 831 BGB), the latter would increase legal certainty (including concerning compensation of non-economic loss) with regard to doctrines such as Schutzgesetzverletzung (cf. § 1311 ABGB or § 823 (2) BGB).

Article 72a
Vicarious liability for AI systems

1. A user of an AI system shall be liable for harm caused by any lack of accuracy or other shortcoming in the operation of the system to the same extent as that user would be liable for the acts or omissions of a human employee mandated with the same task as the AI system.

2. Where a human employee would not have been able to fulfil the task fulfilled by the AI system (such as where the task requires computing capabilities exceeding those of humans) the point of reference for determining the required level of performance is available comparable technology which the user could be expected to use.

Article 72b
Right to compensation and liability

1. Where non-compliance of a party with any obligations following from Titles II, IIa, III or IV of this Regulation has resulted in an increased risk for the safety or fundamental rights of a person, and where that person has suffered economic or non-economic harm *[Opt: material or non-material damage]* because the risk has materialised, the person shall have the right to receive compensation from the party who failed to comply with its obligations.

2. Where a high-risk AI system fails to comply with the requirements set out in Articles 13 to 15 and the harm suffered is of a kind typically resulting from such non-compliance there shall, for the purposes of liability under paragraph 1, be a presumption that the non-compliance has caused the harm.

3. A party who has failed to comply with its obligations shall be exempt from liability under paragraph 1 if it proves that it is not in any way responsible for the non-compliance.

4. Where more than one party has failed to comply with their obligations and is liable under paragraph 1, each party shall be held liable for the entire damage in order to ensure effective compensation of the affected person. Where a party has paid full compensation for the damage suffered that party shall be entitled to claim back from the other liable parties that part of the compensation corresponding to their part of responsibility for the damage.

10.3.3 Enforcement

In addition to including the AIA, or the relevant provisions thereof, in the list of legal instruments in Annex I to the Representative Actions Directive (RAD) (see 10.1.2), it is recommended to include a new enforcement mechanism with regard to systemic risks (for details see 10.2.2). Systemic risks may arise, in particular, where a high-risk AI system that complies with the AIA has, in the light of its significant market coverage, the potential of changing our societies and economies, causing characteristic features and smaller deficiencies (that may be acceptable in an AI system when seen in isolation) to become a systemic risk. For example, bias in a system that is dominant on the relevant market could cause new disadvantaged groups to emerge that can no longer be captured by non-discrimination law as it currently exists, or widespread use of an AI system could have detrimental effects on human skills and competences. The new enforcement mechanism suggested has been inspired by Articles 25 ff DSA, and it includes data access for vetted researchers.

CHAPTER 2A

ADDITIONAL OBLIGATIONS FOR VERY LARGE PROVIDERS TO MANAGE SYSTEMIC RISKS

Article 62a
Very large providers

1. **This Chapter shall apply to providers of high-risk AI systems listed in Annex III for which both of the following conditions are fulfilled:**

(a) **the provider has a share of […] percent or above in the market for AI systems of the relevant type, considering the AI system's core functionalities, with regard to the whole Union, or a share of […] percent or above in the relevant market in at least three Member States; and**

(b) **[…] percent or above of decision-making of the relevant kind listed in Annex III significantly relies on the use of that type of AI system.**

When calculating the share within the meaning of point (a), AI systems that are not placed on the market or put into service under the provider's own name or trademark, but that use the provider's AI system as a basis or component in a way that significantly influences any systemic risks presented by those AI systems, shall be included.

2. **The Commission shall adopt delegated acts in accordance with Articles 73 and 74, after consulting the Board, to lay down a specific methodology for calculating the market share referred to in paragraph 1. In those delegated acts, the Commission may also define different percentages than referred to in paragraph 1 for particular high-risk AI systems where there is reason to believe that systemic risks resulting from that type of AI system are significantly higher or lower than for other AI systems listed in Annex III.**

3. **The Board shall verify, at least once a year, whether the market shares of providers whose AI systems are used in the Union is equal to or higher than the shares referred to in paragraphs 1 and 2. On the basis of that verification, it shall adopt a decision designating the provider as a very large provider for the purposes of this Regulation, or terminating that designation, and communicate that decision, without undue delay, to the provider concerned and to the Commission.**

4. **The Commission shall ensure that the list of designated very large providers is published in the Official Journal of the European Union and keep that list updated. The obligations of this Chapter shall apply, or cease to apply, to the very large providers concerned from four months after that publication.**

Article 62b
Systemic risk assessment

1. As part of the quality management system referred to in Article 17 and post-market monitoring system referred to in Article 61, very large providers shall identify, analyse and assess, at least once a year, any significant systemic risks stemming from the functioning and use made of the AI systems provided by them in the Union.

2. This risk assessment shall be specific to the AI systems they provide and shall, in any case, include the following systemic risks:

(a) any negative effects for the exercise of fundamental rights, for example respect for private and family life, data protection, the prohibition of discrimination, the rights of the child and access to an effective remedy and a fair trial, as enshrined in Articles 7, 8, 21, 24 and 47 of the Charter respectively;

(b) any negative effects for democracy, the rule of law, the functioning of state institutions, the stability of societies and economies, protection of the environment and the combat against climate change, and other important public interests;

(c) any risks resulting from uniformity of decision-making, including for the emergence of new disadvantaged groups, the reduction of diversity in affected groups (e.g. recruited individuals), and a steering function for human behaviour as affected individuals adapt their behaviour to the parameters relied on by the AI system;

(d) any risks resulting from a reduction in human skills and competences, including for the ability to detect and correct errors and to act independently of the AI system where the system is unavailable;

(e) risks of intentional manipulation of their AI system, including by means of targeted inauthentic behaviour of affected persons, malicious interferene by third parites, or hybrid warfare, with an actual or foreseeable negative effect on important public or private interests.

Article 62c
Mitigation of systemic risks

1. Very large providers shall put in place reasonable, proportionate and effective mitigation measures, tailored to the specific systemic risks identified pursuant to Article 62b. Such measures may include, where applicable:

(a) adapting AI systems, their decision-making processes, their features or functioning, or the instructions and specificaitons accompanying them;

(b) reinforcing the internal processes or supervision of any of their activities in particular as regards detection of systemic risk;

(c) …

2. **The Board, in cooperation with the Commission, shall publish comprehensive reports, once a year, which shall include the following:**

(a) identification and assessment of the most prominent and recurrent systemic risks reported by very large providers or identified through other information sources;

(b) best practices for very large providers to mitigate the systemic risks identified.

3. **The Commission, in cooperation with the Board, may issue general guidelines on the application of paragraph 1 in relation to specific risks, in particular to present best practices and recommend possible measures, having due regard to the possible consequences of the measures on fundamental rights enshrined in the Charter of all parties involved. When preparing those guidelines the Commission shall organise public consultations.**

Article 62d
Independent audit

1. **Very large providers shall be subject, at their own expense and at least once a year, to audits to assess compliance with the following:**

(a) the obligations set out in Chapter 3 of Title III;

(b) any commitments undertaken pursuant to the codes of conduct referred to in Article 69.

2. **Audits performed pursuant to paragraph 1 shall be performed by organisations which:**

(a) are independent from the very large providers concerned;

(b) have proven expertise in the area of risk management, technical competence and capabilities;

(c) have proven objectivity and professional ethics, based in particular on adherence to codes of practice or appropriate standards.

3. **The organisations that perform the audits shall establish an audit report for each audit. The report shall be in writing and include at least the following:**

(a) the name, address and the point of contact of the very large provider subject to the audit and the period covered;

(b) the name and address of the organisation performing the audit;

(c) a description of the specific elements audited, and the methodology applied;

(d) a description of the main findings drawn from the audit;

(e) an audit opinion on whether the very large provider subject to the audit complied with the obligations and with the commitments referred to in paragraph 1, either positive, positive with comments or negative;

(f) where the audit opinion is not positive, operational recommendations on specific measures to achieve compliance.

4. Very large providers receiving an audit report that is not positive shall take due account of any operational recommendations addressed to them with a view to take the necessary measures to implement them. They shall, within one month from receiving those recommendations, adopt an audit implementation report setting out those measures. Where they do not implement the operational recommendations, they shall justify in the audit implementation report the reasons for not doing so and set out any alternative measures they may have taken to address any instances of non-compliance identified.

Article 62e
Transparency reporting obligations for very large providers

1. Very large providers shall make publicly available and transmit to the Board and the Commission, at least once a year and within 30 days following the adoption of the audit implementing report provided for in Article 62d(4):

(a) a report setting out the results of the risk assessment pursuant to Article 62b;

(b) the related risk mitigation measures identified and implemented pursuant to Article 62c;

(c) the audit report provided for in Article 62d(3);

(d) the audit implementation report provided for in Article 62d(4).

3. Where a very large provider considers that the publication of information pursuant to paragraph 2 may result in the disclosure of confidential information of that provider or of the users of the AI system, may cause significant vulnerabilities for the security of its AI system, may undermine public security or may harm users or affected individuals, the provider may remove such information from the reports. In that case, that provider shall transmit the complete reports to the Board and the Commission, accompanied by a statement of the reasons for removing the information from the public reports.

Article 62f
Data access and scrutiny by vetted researchers

1. Upon a reasoned request from the Commisison, very large providers shall, within a reasonable period, as specified in the request, provide access to data to vetted researchers who meet the requirements in paragraphs 3 of this Article, for the sole purpose of conducting research that contributes to the detection, identification and understanding of systemic risks in the Union as set out in Article 62b(1), including as regards the adequacy, efficiency and impacts of the risk mitigation measures pursuant to Article 62c. In making a request, the Commission shall take due account of the rights and interests of the providers and users of the AI system concerned, including the protection of personal data,

the protection of confidential information, in particular trade secrets, and maintaining the security of their AI systems.

2. Very large providers shall facilitate and provide access to data pursuant to paragraph 1 through appropriate interfaces specified in the request, including online databases or application programming interfaces.

3. Upon a duly substantiated application from researchers, the Commission shall award them the status of vetted researchers and issue data access requests pursuant to paragraph 1, where the researchers demonstrate that they meet all of the following conditions:

(a) they are affiliated to a research organisation as defined in Article 2 (1) of Directive (EU) 2019/790 of the European Parliament and of the Council;

(b) they are independent from commercial interests;

(c) they are in a capacity to preserve the specific data security and confidentiality requirements corresponding to each request and to protect personal data, and they describe in their request the appropriate technical and organisational measures they put in place to this end;

(d) the application submitted by the researchers justifies the necessity and proportionality for the purpose of their research of the data requested and the timeframes within which they request access to the data, and they demonstrate the contribution of the expected research results to the purposes laid down in paragraph 1;

(e) the planned research activities will be carried out for the purposes laid down in paragraph 1;

(f) they carry their activities according to the procedures laid down in delegated acts referred to in paragraph 7;

(g) they have not already filed the same application with the Commission.

4. The Commission shall issue a decision terminating the access if it determines, following an investigation either on its own initiative or on the basis information received from third parties, that the vetted researcher no longer meets the conditions set out in paragraph 3. Before terminating the access, the Commission shall allow the vetted researcher to react to the findings of its investigation and its intention to terminate the access.

5. Upon completion of the research envisaged in paragraph 1, the vetted researchers shall make their research results available to the Commission free of charge. The Commission may make the research results publicly available, taking due account of the rights and interests of the providers and users of the AI system concerned, including the protection of personal data, the protection of confidential information, in particular trade secrets, and maintaining the security of their service.

6. The Commission shall, after consulting the Board, adopt delegated acts laying down the technical conditions under which providers of very large providers are to share data pursuant to paragraphs 1 and 2 and the purposes for which the data may be used. Those delegated acts shall lay down the specific conditions and relevant objective indicators, as well as procedures under which such sharing of data with vetted researchers can take place in compliance with Regulation (EU) 2016/679, taking into account the rights and interests of the providers and users of the AI system concerned, including the protection of confidential information, in particular trade secrets, and maintaining the security of their AI system.

In addition to a new enforcement mechanism for systemic risks it is suggested to insert a provision that avoids threats to public and national security interests (see 10.3) which could result if national authorities in all 27 Member States had full access to all relevant data and the source code of, e.g., AI systems that are safety components in critical infrastructure (such as AI systems used to detect attacks on power grids within the Union).

Article 70a
Exceptions for AI systems with enhanced confidentiality requirements

1. A provider of a high-risk AI system that is confronted with a request by a competent national authority for information, documentation, access to data, disclosure of the source code or a similar measure under this Regulation may refuse to comply with the request if that provider can demonstrate that the relevant materials would, if disclosed to unauthorised parties, jeopardise public and national security interests.

2. A provider relying on paragraph 1 shall immediately notify the Commission of the refusal to comply with the request and the reasons of the refusal. The Commission shall, upon having investigated the matter, issue a decision addressed at the relevant national authority and the provider. In that decision, the Commission may provide that only Commission staff holding the appropriate level of security clearance shall be allowed to access the relevant materials, and impose further restrictions and safeguards as appropriate.

Federal Ministry of
Social Affairs, Health, Care
and Consumer Protection
Stubenring 1, 1010 Vienna, Austria
+43 1 711 00-0
sozialministerium.at