Issue Brief: The EU Artificial Intelligence Act

Comments on Articles 3 & 5 of the Draft Proposal

March 1, 2022

Authored by:

Mikołaj Barczentewicz (Senior Scholar, International Center for Law & Economics)

ICLE | 1104 NW 15th Avenue Suite 300 | Portland, OR 97209 | 503.770.0076

icle@laweconcenter.org | @laweconcenter | www.laweconcenter.org

I. INTRODUCTION

European Union (EU) legislators are considering legislation— the Artificial Intelligence Act (AIA), the original draft of which was published by the European Commission in April 2021[1]—that aims to ensure the safety of AI systems in uses designated as "high risk". As originally drafted, however, the AIA's scope was not at all limited to AI; it would instead cover virtually all software. EU governments seem to have realized this problem and are trying to fix the proposal, while some pressure groups have pushed to move the draft in the opposite direction.

The AIA proposal is currently under consideration by specialized committees of the European Parliament. The parliamentary stage began with a long disagreement among the various committees regarding who should have decisive influence over the Parliament's position on the bill. With that disagreement now resolved, discussions on the legislation's merits are ongoing.

The purpose of this brief is to inform debate on the proposal's fundamental features: its scope and the key provisions setting out prohibited AI practices (related to so-called "subliminal techniques" and "social scoring").

II. GENERAL DEFINITION OF AN 'AI SYSTEM'

One of the key definitions determining the AIA's effective scope is that of "artificial intelligence system" ("AI system"), in Article 3(1). In the original AIA proposal, the Commission's definition was based, in part, on a recommendation from the Organisation for Economic Co-operation and Development (OECD).[2] According to the OECD:

> An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy.

The Commission's proposal modified the OECD recommendation to define an AI system as:

---

[1] *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, EUROPEAN COMMISSION, (Apr. 21, 2021), *available at* https://perma.cc/RWT9-9D97.

[2] *Recommendation of the Council on Artificial Intelligence*, OECD LEGAL INSTRUMENTS, (May 21, 2019), https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

> [S]oftware that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with[3]

Setting aside the "techniques and approaches listed in Annex I" for a moment, the Commission's general definition is overly broad to such a significant degree that it would cover nearly all software.[4] The most meaningful change in the Commission's version was dropping any reference to "autonomy". The OECD's general definition is not appreciably narrower, however, as its reference to "*varying levels* of autonomy" (emphasis mine) would render even systems with only very limited autonomy within the scope of the definition.

The general definition should be much more aligned with people's intuitive understanding of what an AI system is. This would help to avoid the outcome of the AIA having significant unexpected effects on EU businesses and citizens, thereby offending the basic principles of the rule of law. A partial solution in that direction would be to reinstate the "autonomy" element from the OECD definition, but with a qualification that the level of autonomy must be "significant" (instead of "varying").

In the compromise text that it promulgated in November 2021, the European Council's Presidency proposed a modified definition of an AI system as one that:[5]

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

This definition departs from the OECD approach and seems intended to align more closely with the prevailing public understanding of AI. On reflection, however, this definition is also not appreciably narrower than the Commission's original proposal. The core of the Council's Presidency

---

[3] European Commission, *supra* at Note 1.

[4] Mikołaj Barczentewicz and Benjamin Mueller, *More Than Meets The AI: The Hidden Costs of a European Software Law*, CENTER FOR DATA INNOVATION, (Dec. 1, 2021), *available at* https://www2.datainnovation.org/2021-more-than-meets-the-ai.pdf; *see also* Mikolaj Barczentewicz, *EU's Compromise AI Legislation Remains Fundamentally Flawed*, TRUTH ON THE MARKET, (Feb. 8, 2022), https://truthonthemarket.com/2022/02/08/eus-compromise-ai-legislation-remains-fundamentally-flawed.

[5] *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*, COUNCIL OF THE EUROPEAN UNION, (Nov. 29, 2021), *available at* https://data.consilium.europa.eu/doc/document/ST-14278-2021-INIT/en/pdf; *see also*, Barczentewicz, *supra*, note 4.

definition is in Article(3)(1)(ii), according to which an AI system is one that: "infers how to achieve a given set of human-defined objectives using learning, reasoning or modelling implemented with the techniques and approaches listed in Annex I". A broad interpretation of the terms "reasoning" and "modelling" could be read to cover all algorithms and much of applied statistics (statistical modelling)—i.e., applications clearly beyond the scope of the widespread understanding of what constitutes AI.

These three attempts to define an AI system demonstrate the difficulty of providing a definition that manages to avoid being overly broad without limiting the scope to an exclusive list of covered technologies. In fact, the Commission and the Council do also adopt this latter approach, with an identical list of "techniques and approaches" enumerated in Annex I of both proposals. That list, however, is itself overly broad:

> (a) Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning;
>
> (b) Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems;
>
> (c) Statistical approaches, Bayesian estimation, search and optimization methods.

Of these three sets, only (a) is limited to those technologies commonly understood as AI. Combined with either of the general definitions discussed above, the letters (b) and (c) would cover nearly all software and all programmatic applications of statistics. The clear solution to bring the AIA's scope within the public's common understanding of AI is to delete letters (b) and (c) from the list. Barring that change, the AIA would, in practice, serve as an "all-software" law for which no appropriate impact assessment has been conducted and whose breadth would (unreasonably) surprise those to whom it will apply.

III. PROHIBITED AI PRACTICES (ARTICLE 5)

The provisions prohibiting certain AI practices (Article 5) also may be drafted more broadly than intended. This is particularly striking in the AIA's prohibitions of "subliminal techniques" and "social scoring".

A. Subliminal Techniques (Article 5(1)(a))

The Commission's original proposal aimed to prohibit:

> [T]he 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;

This phrasing did not change significantly in the Council's Presidency proposal:

> [T]he placing on the market, putting into service or use of an AI system that deploys subliminal techniques beyond a person's consciousness with the objective to or the effect of materially distorting a person's behaviour in a manner that causes or is reasonably likely to cause that person or another person physical or psychological harm;

It is unclear, to start, whether this provision is needed at all. There are already similar prohibitions in the Unfair Commercial Practice,[6] as noted in the Recital 16 from the Presidency compromise text.

Moreover, the prohibition refers to a concept ("subliminal techniques") that has much more limited application in the scientific literature than it does in the popular imagination. The draft text of both versions of Article 5(1)(a) suggest the drafters may be conflating science and science fiction. Popular belief in the power of subliminal techniques dates to a 1957 hoax by a publicity-seeking market researcher.[7] While a small number of studies[8] in the years since have suggested that subliminal techniques may have some effects on human behaviour, this has been found only in very limited circumstances. The scientific consensus has falsified the existence of the kinds of "mind control"—e.g., through subliminal advertising—that appears to have motivated drafters of the proposed prohibition.

A legal prohibition on commercial firms merely having the "objective" (intent) of "materially distorting a person's behaviour" will create significant uncertainty as to the lawfulness of practices that, given the current state of scientific knowledge, appear unlikely to have much effect on human

---

[6] *Directive 2005/29/EC of the European Parliament and of the Council*, EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION, (Nov. 27, 2019), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32005L0029.

[7] Johan C. Karremans, Wolfgang Stroebe, & Jasper Claus, *Beyond Vicary's fantasies: The impact of subliminal priming and brand choice*, 42 J. EXP. SOC. PSYCHOL. 792-798, (November 2006) https://www.sciencedirect.com/science/article/abs/pii/S0022103105001496?via%3Dihub.

[8] Sara Garofalo, Laura Sagliano, Francesca Starita, Luigi Trojano & Giuseppe di Pellegrino, *Subliminal determinants of cue-guided choice*, 10 SCI. REP. 11926, (Jul. 17, 2020) https://www.nature.com/articles/s41598-020-68926-y.

behaviour. At the very least, this prohibition should be limited only to practices proven to have the effect of materially distorting a person's behaviour in ways that cause demonstrable harm.

If interpreted broadly, "subliminal techniques beyond a person's consciousness with the objective to or the effect of materially distorting a person's behaviour" could refer to any aspects of, e.g., digital advertising or social-media recommendation systems that a user does not "consciously" notice, but that may nonetheless affect their behaviour. The difficulty arises primarily in defining what is meant by "subliminal". It could variously be defined as something a user doesn't notice in a given instance; something that is not easily noticeable; or something that is not known to the user but is potentially knowable by her.

A user may not notice that several instances of the same ad has been displayed in succession; in such a case, is it falsifiable that the ad might nonetheless have affected the user's behaviour if the user ultimately buys the advertised product? Does all advertising (or recommendation systems) "distort" human behaviour? The AIA's text is mute on these distinctions, as it does not define "distortion", just as it does not define "subliminal techniques".

Also troubling is the extremely low threshold set by the standard of "reasonable likelihood to cause psychological harm". Many things in the world are "reasonably likely" to cause at least some amount of subjective psychological harm. People can find mere disagreement with their peers to be offensive and, therefore, harmful. At the very least, it should be expressly clear that the law's scope covers only significant harms that can be detected and quantified according to some objective standard. Better still would be to follow well-established models of regulation and expressly limit prohibitions to causes of physical or economic harm. The law is not a good instrument to address psychological harms, especially merely subjective ones.

Even if there is a legislative gap that (a) would address, which is doubtful, the current proposal is so poorly drafted that lawmakers should undertake a fundamental reassessment of what this provision intends to prohibit, and tailor adequate legislative language accordingly.

B. Social Scoring (Article 5(1)(c))

The Commission's original proposal also would prohibit:

> (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

The Council's Presidency proposal broadened the scope of this prohibition by amending the initial part to read:

> (c) the placing on the market, putting into service or use of AI systems for the evaluation or classification of natural persons…

There is significant risk that the Council's proposal to extend this prohibition to private actors would be interpreted to prohibit certain business practices that are beneficial to customers. A key example is the use of credit and risk scoring for pricing and underwriting in the lending and insurance markets. The ability to use more individual factors for risk scoring allows businesses to better assess risk and thus offer more attractive financial products to customers with lower risk. Some of the individual factors and features used in risk-scoring models may indeed be "unrelated to the contexts in which the data was originally generated or collected".

The question then shifts to whether such beneficial practices constitute "detrimental or unfavourable treatment of certain natural persons or whole groups thereof". The nature of risk scoring is that it allows more *accurate* pricing of risk, which means that prices will be higher for those with higher risk and lower for those with lower risk. One could argue that those with higher risk are treated unfavourably by systems that offer them higher prices. It is not clear, however that such "disadvantage" is inherently unfair in a way that requires legislative intervention. This "disadvantage" merely means that those with higher risk are not able to shift the cost of their risk to others with lower risk. If there are worthy social reasons to make credit or insurance more accessible to groups of people with higher risk, this is better addressed through public subsidies to such people, not by forcing lenders or insurers to be blind to risk.

Moreover, the prohibition would target detrimental treatment of people "unjustified or disproportionate to their social behaviour or its gravity". Implementing this limitation in risk-scoring systems will make it hard, if not impossible, to use state-of-the-art AI techniques that may be highly accurate in predicting risk, but without a transparent demonstration of which factors were relied on and to what extent. A lender or insurer using such techniques may find it difficult (or even impossible) to prove that they have complied with this provision. Moreover, even for systems whose risk-weighting terms are readily transparent, some factors (or combination of factors) that are highly correlated with risk may be unintuitive, e.g., due to the complexities of causation in the real world.

Thus, the use of such factors could be deemed "unjustified or disproportionate" due to the human bias towards intuitive ("common sense") explanations.

Given that social scoring, as understood in Article 5(1)(c), is, by definition, based on data related to specific individuals, it is within the scope of the General Data Protection Regulation (GDPR). Hence, some concerns about, e.g., risk scoring by private actors are already alleviated by the GDPR and especially by application of the principles of lawfulness, fairness, and accountability. Indeed, the final sentence of Recital 17, added by the Council, seems to address some of the issues raised here:

> This prohibition should not affect lawful evaluation practices of natural persons done for one or more specific purpose in compliance with the law.

This sentence, however, does not improve matters in the many situations where "evaluation practices" (like risk scoring) are done today "in compliance with the law" chiefly because they are not prohibited by the law. It seems that the drafters forgot that, unlike public authorities, private actors can lawfully do anything that they are not legally prohibited from doing. Hence, this sentence has one of two potential meanings. Either whatever is "in compliance with the law" today (including compliance due to not being prohibited) will not be caught by the prohibition—an interpretation that is extremely hard to square with Article 5(1)(c)—or the sentence only applies to practices "in compliance with the law" in the narrow sense of being expressly permitted or mandated by the law. That latter set is likely much smaller than what the Council drafters had in mind when adding this sentence to Recital 17. Even on the narrower interpretation, there remain serious questions of how such an exemption can be read into Article 5(1)(c).

Applying the prohibition on "social scoring" to private actors is more likely to stifle innovation and deny customers access to valuable services than to bring any appreciable benefit that could not be provided more fairly and effectively in other ways. The concerns identified above may be less applicable in the case of the use of "social scoring" by public authorities, where considerations of transparency and equality may carry more weight than those of efficiency, cost, and innovation. However, even with respect to their use by public authorities, the costs of a prohibition should be rigorously evaluated, with an adequate assessment of lost opportunities for innovation in the provision of public services.

IV. CONCLUSION

This policy brief focused on two fundamental issues with the AIA proposal. It argued that the basic definitions employed in the AIA turn the regulation into a law governing all software, which is far from how the AIA has been presented to the public and to legislators. This disconnect threatens to undermine the democratic legitimacy of this legislative process and render it unlikely

that the full effects of AIA are adequately considered. The solution proposed here is to limit application of the AIA to technology widely considered to be AI (i.e., to machine learning), thus removing the risk of legislating by surprise.

Moreover, this policy brief considered two poorly drafted and inadequately justified general prohibitions of AI practices involving "subliminal techniques" and "social scoring". The burden is on the European Commission to show that such prohibitions are needed, and this burden has not been met. Even if prohibition of some such practices is justified, much more care needs to be put into its drafting.

There are other concerns about the AIA proposal that this policy brief did not examine, leaving them for future work. They include, i.e., general questions surrounding the adequacy of the official impact assessment provided for the AIA; the scope of the "high risk" designation; and the details of compliance obligations to be imposed on "high risk" AI systems.