Panoptykon Foundation's submission to the consultation on the 'White Paper on Artificial Intelligence - a European Approach to Excellence and Trust'

Warsaw, 10 June 2020

About Panoptykon

Panoptykon Foundation is a Warsaw-based NGO with a mission to protect fundamental rights in the context of growing surveillance and fast-changing information technologies. We believe in 'watching the watchers' and consider data a source of power. Therefore we keep an eye on entities that collect and use personal data in order to influence people (public authorities, intelligence agencies, business corporations). On the legal front we keep track of new legislation, develop alternative regulatory solutions and intervene to protect human rights. In our advocacy we address both policymakers and business lobbies. Through our research and investigations we expose risks related to commercial and public surveillance in order to raise public awareness. We visualize collected data and engage in artistic collaborations in order to reach broader audiences. Since 2010 we have been an active member of European Digital Rights (EDRi).

Contents

1. Scope of the regulatory framework on AI

The working assumption presented in the White Paper is that the EU regulatory framework will apply to products and services that rely on AI, as defined by the HLEG on Artificial Intelligence. In our view the proposed regulatory framework should cover **all AI systems that will be applied to humans and/or *may* affect them**. This approach still excludes mundane and purely internal applications of AI that do not relate to people, e.g. "smart" information management systems, while ensuring that all systems that may impact (groups of) individuals are regulated.

We propose the following definition to limit the scope of AI regulation (and the obligation to conduct human rights impact assessment[1]):

> *all AI applications that may in any way affect humans, in particular their legal situation, their physical or mental condition, or their access to goods and services*

Please note that this definition covers AI applications that may have impact both on individuals and on groups of people (in such case the impact will be societal). It also covers applications of AI regardless of whether the impact is positive or negative; significant or not.

In order to limit the potential of abuse of this definition the **burden of proof should be on the entity wanting to develop or deploy the AI** system to demonstrate that the system does not affect humans in any way.

2. Delineation of scope of 'lawful AI'

The approach presented in the White Paper (based on specific requirements for high-risk applications of AI in pre-defined sectors) is not sufficient to ensure that EU residents will be effectively protected against uses of AI that might interfere with their fundamental rights. It wrongly assumes that risk can always be calculated and mitigated, and ignores the fact that there are uses of AI that are by definition incompatible with the European fundamental rights framework.

Rather than allowing all applications of AI by default, with additional requirements and safeguards only for high-risk applications in pre-defined sectors, the EU should **set explicit legal boundaries for all AI applications**. These legal boundaries should take into account **social and fundamental rights concerns** and provide clarity for developers as to which AI applications may or may *not* be developed and deployed, and for which purposes.

For delineating the scope of 'lawful AI', we recommend a **mixed approach**, which combines:

(i) general rules for inadmissibility of AI applications; and

(ii) specific bans on applications that should be declared incompatible with European law (in particular fundamental rights).

In our opinion general rules are necessary to ensure **flexibility when regulating this dynamic and relatively new sector**. At the same time, based on available research and documented cases

[1] Please refer to point 3 for explanation on the proposed HRIA system.

of abuse, we find enough evidence for European authorities to draw the red line and prevent the most risky experiments in the field of AI.

Keeping in mind these two considerations, we propose the following delineation of the scope of 'lawful AI':

An application of AI should be deemed lawful, unless one of the following conditions[2] is fulfilled, based on the results of mandatory human rights impact assessment[3]:

(1) For AI applications in the public sector: There is no evidence of provable benefits of AI application, regardless whether potential risks can be mitigated (in particular with regard to prediction of individual behaviour, which is the most contested area of AI application[4]);

(2) For AI applications in both the public and private sector: No measures that can mitigate identified risks to human rights have been proposed.

(3) For AI applications in both the public and private sector: The AI system has turned out to be complex to the degree that it does not allow for human review and scrutiny and, therefore, cannot be subjected to standards of transparency and accountability[5].

(4) For AI applications in both the public and private sector: The application violates the essence of fundamental rights and freedoms, in particular human dignity.

In addition, the EU should **proactively ban** applications of AI systems in areas where the fundamental rights and societal implications are too great to risk. Applications identified as such should be named and forbidden in law (or in binding guidelines issued by the relevant supervisory body). The list should not be framed as exhaustive. Technological developments and new evidence should be taken into account in binding decisions (guidelines) of the relevant supervisory bodies and/or in the case law that will interpret general rules mentioned above (thus expanding the list of prohibited AI applications, if necessary).

We propose explicit bans for at least the following applications of AI systems:

- indiscriminate biometric surveillance and biometric capture and processing in public spaces,
- use of AI to solely determine access to or delivery of essential public services (such as social security, migration control -- including citizen scoring),
- uses of AI which purport to identify, analyse and assess emotion, mood, behaviour, and sensitive identity traits (such as race, disability) in the delivery of essential services,
- prediction of behaviour of individuals for the purposes of law enforcement and criminal justice (such as personal risk scores).

[2] Please note that these are independent criteria, and some of them apply only to public sector applications.

[3] Please see point 3 for the explanation of the human rights impact assessment system and the role of the oversight body in such cases.

[4] See e.g. the work of Solon Barocas, Moritz Hardt and Arvind Narayanan on this issue.

[5] This general rule is based on the recommendation formulated by the Council of Europe.

3. Effective human rights impact assessment regime (HRIA)

Together with other European digital rights organisations, in particular European Digital Rights and Access Now, we call for mandatory, *ex ante* human rights impact assessments for all AI applications that fall into the scope of the regulation (see point 1). The approach presented by the Commission in the White Paper, where a sectoral criterion must be met to qualify an AI application as high-risk, creates a threat of overlooking high-risk AI applications in "low risk" sectors. We argue that the level of risk created by a specific AI application will often depend on the scale and purposes of this particular project, rather than be determined by a sector in which AI is applied. For example, while online advertising may be perceived as a "low-risk" sector in general, there are well-documented cases of harmful or discriminatory advertising (incl. job offers[6]), which should not be left outside the scope of HRIA obligation.

The approach we propose, based on mandatory HRIA for all AI applications that may affect humans, combined with public disclosure obligations, solves this problem, while - at the same time - it does not lead to unnecessary bureaucracy or formalisation.

In fact, in order to effectively serve their function, HRIAs should not be reduced to a burdensome formality. European law should introduce appropriate measures for preventing this from happening. Keeping this risk in mind, we recommend a HRIA system for AI applications that is modelled on the GDPR provisions on data protection impact assessments but with important corrections based on two years of experience with the DPIA. GDPR model should be improved by: (i) introducing a mandatory disclosure scheme, (ii) increasing the role of external reviewers, and (iii) increasing engagement from affected communities and civil society.

Main premises of the "GDPR+" HRIA regime:

- HRIAs should be conducted and documented for **all** AI systems that affect humans, not only those that are known to pose a high risk.
- HRIAs should be an ongoing process and should be periodically reviewed when a 'change in the risks' occurs[7], particularly at the stages of design, development, testing and deployment.
- **Addressees of HRIA obligation and liability regime:**
  - As different actors might be involved in and responsible for different stages of development and deployment of an AI system, HRIAs should be completed by each responsible actor prior to commercialisation (passing on the rights to the system) or deployment.
  - Each entity in the supply chain should be responsible for producing and maintaining appropriate documentation.
  - The liability regime should be designed accordingly (i.e. AI buyers/deployers should not be liable for the failure to comply with regulatory framework on the

[6] See for example: Facebook allowing advertisers to target teens based on psychological vulnerabilities, examples of bias against women and people of colour in job adverts.

[7] What exactly constitutes a 'change in the risks' should be fleshed out by an oversight body (e.g. an equivalent of the European Data Protection Board) with a (non-exhaustive) list of examples, as proposed by Reuben Binns.

part of AI sellers/developers, if in the process of negotiating the contract they behaved with due diligence).

- **Minimum requirements regarding the scope and content of HRIAs:**
  - As far as the design of the system is concerned, HRIAs should include:
    - normative explanations of how the system was built (including steps that have been taken to ensure that the outcomes are unbiased and fair);
    - key technical parameters, such as: loss function; formal fairness metrics (applied in the training process); presentation of performance measurements; (in the case of accuracy and the other performance metrics) results of cross-validation (training/ testing splits) and any external validation carried out; presentation of confusion matrix (i.e. the table that provides the range of performance metrics) and ROC curve (receiver operating characteristics) / AUC (area under the curve);
    - an assessment of the impact of these design decisions on (groups of) individuals.
  - Performance of the AI system should be tested on real (not only training) data prior to deployment. Results of these tests should be part of the HRIA documentation.
  - HRIA should not be limited to an evaluation of the models or algorithms behind the AI system, but should include an evaluation of how decision-makers might collect or influence the inputs and interpret the outputs of such a system[8].
  - HRIA should not only look at impact on individuals but also evaluate collective, societal, institutional and governance implications the system poses[9].
  - HRIA should also outline adequate steps and measures to mitigate negative implications.
- **Mandatory disclosure scheme:**
  - AI developers need to ensure that **full documentation, covering all stages of the AI development process**[10] is available to the regulators.
  - Regardless of the level of impact, HRIAs, including conclusions from the external review process, should be **made available to the public** in an easily accessible and machine-readable format. Intellectual property rights and trade secrets cannot be used as an excuse for not disclosing the HRIA. The document may be redacted but it should still offer meaningful insight into the HRIA process.

    Publication of HRIAs enables public debate on the quality of impact assessment and makes it possible for civil society or investigative journalists to identify/flag potential abuses. We don’t see another way to ensure that potentially dangerous uses of AI are not qualified as ‘low-impact’ or ‘low-risk’ by the AI

[8] Based on a recommendation formulated by the Council of Europe (Recommendation 1).
[9] *Ibid.*
[10] See for example ‘AI Blindspot’ discovery process for explanation of the development process.

developer/deployer and therefore escape any form of external review or public scrutiny.

- **When high impact is established:**
  - Independent **external review** of AI systems should be performed by a specialised review body (we support the direction of "prior conformity assessment" process outlined in the White Paper[11]). The report from the review should be annexed to the HRIA.
  - After an external review is completed, the **oversight body should be notified and provided with the HRIA and the review report**, as well as any supporting documentation. The role of the oversight body could be modelled on Article 36 of the GDPR.
- In the **public sector**, apart from the requirement to conduct and publish HRIAs as well as additional requirements for high-impact applications (as discussed in the bullet point above), there should be a **mandatory consultation process** involving groups that are likely to be affected by the proposed system and relevant civil society groups. The public entity responsible for deploying this system should publicly **respond to concerns** raised during the consultation. The oversight body should be provided with full documentation of the process, including submissions from other stakeholders and responses of the public entity.

The chart below summarises requirements for AI developers/deployers, which depend on the level of impact established in the HRIA process (high/low):

Conclusion of HRIA	**Public disclosure of HRIA**	**Mandatory external review**	**Notification to oversight body**	**Obligation to keep records and documentation**	**Obligation to provide documentation to oversight body**
high-impact	YES	YES	YES	YES	YES
low-impact	YES	NO	NO	YES	NO (unless requested during ex-post control)

[11] Inspiration taken from cybersecurity regulation: "Conformity assessment is a procedure for evaluating whether specified requirements relating to an ICT product, ICT service or ICT process have been fulfilled. That procedure is carried out by an independent third party that is not the manufacturer or provider of the ICT products, ICT services or ICT processes that are being assessed." "the conformity assessment bodies shall be accredited by national accreditation bodies appointed pursuant to Regulation (EC) No 765/2008."

In summary, we propose that enforcement of AI regulation be **modelled on the enforcement of the DPIA, with an additional requirement for external review**, which should address both the shortcomings of self-assessment and the risk of overloading public oversight bodies. Internal and external reviews should be combined with strong administrative measures (such as high financial penalties and decision to halt the deployment) if an entity fails to implement adequate measures to mitigate risks identified during HRIA. In cases when there are no measures that can effectively mitigate identified risks, deployment of an AI system should be deemed illegal (in line with the criteria specified in point 2 above).

The European Commission should propose a template for HRIAs or issue guidelines and recommendations, in order to ensure a common and objective understanding of what constitutes a risk for human rights and how to assess its likelihood and severity. Digital Innovation Hubs described in the White Paper may also be used to help SMEs in this process.

We also encourage the European Commission to create a support system (e.g. grants) for civil society organisations to investigate publicly available HRIAs in order to identify and flag potential abuses, therefore contributing to building an ecosystem of trust and excellence in the field of AI in a similar way that fact-checking organisations currently support the EU's fight against disinformation.

4. General explanation of an AI system (public-facing)

Regardless of conducting and publishing the results of HRIAs (redacted, if necessary) as well as providing full documentation to the regulator (whenever requested), deployers of all AI systems that may in any way affect individuals should be obliged to publish **general, public-facing explanations** on how these systems work.

The purpose behind producing a general explanation is to summarise, in the most accessible form and non-expert language, **key information on AI systems for all interested stakeholders**. Because of the risks specific to AI systems, people need to be informed that they are dealing with an AI system and understand - on a general level - objectives, logic and risks involved in using this system. This transparency and accountability mechanism should work in a similar way to nutrition labels for processed food or information leaflets for pharmaceutical products. The design of general explanations for AI systems should be developed in collaboration between oversight bodies, UX designers, AI deployers and civil society groups.

General explanation of the system and its functioning should include the following information[12]:

- who it was developed by, who deploys the system and contact details for human review (if applicable);
- what is the problem it tries to solve and expected use case/s;

[12] This approach to explaining AI-assisted outcomes/decisions is based on the following resources:
ICO: Explaining decisions made with AI , Part 1
Margaret Mitchell et al: Model Cards for Model Reporting

- translation of the system workings (i.e. its input and output variables and key parameters) and what role these factors play in reasoning about the real-world problem that the system is trying to address or solve;
- the generation and class of the algorithm;
- categories of data that were used to train the model and their sources;
- normative explanations of how the system was built, including steps that have been taken to ensure that the outcomes are unbiased and fair and to maximise its accuracy, reliability, security and robustness;
- indication of who (i.e. which groups or categories of people) may be affected in a negative way, based on test results and documented errors produced by the model.

5. Individual explanation (user-facing)

The GDPR (in Article 22) creates a legal obligation to explain "the logic behind" an automated decision, if it was based on profiling, did not engage human oversight and produced legally binding or significant effects for an individual. On a number of occasions we have argued that the narrow scope of this provision does not ensure transparency and accountability of AI-driven decisions towards affected individuals. We see the upcoming horizontal AI regulation as a chance to fill this gap.

It is also important to account for cases in which AI outcomes affect a category or a group of individuals in a way that is experienced by an individual but cannot be "singled out" as a decision or outcome relating to a particular individual. For example, behavioural advertising is targeted at certain segment(s) of customers, potentially affecting each and every individual in this group, but it is nearly impossible for a specific individual to prove that she or he has been targeted based on specific criteria.

In terms of scope, an obligation to provide an individual with an explanation of AI-assisted decision (or simply an outcome of the algorithmic data processing) should apply:

(i) whenever algorithmic data processing (or an AI-assisted decision) in any way affects individual or group of individuals, particularly in terms of their legal situation, access to goods and services, mental or physical health and wellbeing (the burden of proof that the outcome of the system had no impact on humans should be on the system deployer);

(ii) regardless of whether:

- the impact was significant;
- human oversight was involved;
- decision/outcome was based on processing personal data (as defined by the GDPR) or statistical models only (so called "big" data).

In terms of content, individual explanation of an AI-assisted decision (or an outcome of algorithmic data processing) should contain at least[13]:

- the **reasoning behind** a particular algorithmically-generated outcome in plain, easily understandable, and everyday language (incl. clarification of **how a statistical result has been applied to the individual concerned** to show how the reasoning behind the decision takes into account the specific circumstances, background and personal qualities of affected individuals);
- enumeration of the **input data used for a specific decision, and the sources of that data**;
- explanation of **output data**, particularly **if the decision recipient has been placed in a category which may not be clear to them**;
- explanation of **why the outcome is correct** (the confidence a system has in an outcome), legal and fair, i.e. that the decision is based on proportional and necessary data processing, using pertinent categories of data and relevant profiling mechanisms;
- explanation of **how formal fairness criteria were applied** to this particular decision or output;
- explanation of **how others similar to the individual were treated** (i.e. whether they received the same decision outcome as the individual) and the characteristics of similarly classified individuals;
- if there was a meaningful human involvement in the decision-making process, explanation of **how and why a human judgement** (assisted by an AI output) **was reached**;
- instruction of **how to request a human review** of an AI-enabled decision or object to the use of AI, including details on who to contact, and what the next steps will be (e.g. how long it will take, what the human reviewer will take into account, how they will present their own decision and explanation).

Whenever possible, individuals should also be offered a chance to test **counterfactual explanations**[14]. This is an interactive process, which enables an individual to test different possible decision outcomes and see how a change in their behaviour (within the range of possible behaviours, as defined by the system provider) would have brought about a different outcome.

[13] Based on the ICO and the Alan Turing Institute's Guidelines.

[14] Counterfactual explanations describe the smallest change to the world that can be made to obtain a desirable outcome, or to arrive at the closest possible world, without needing to explain the internal logic of the system.