CHAI Position Paper on the EU Artificial Intelligence Act

University of California, Berkeley

The Center for Human-Compatible AI (CHAI) is a research group based at UC Berkeley, with academic affiliates at a variety of other universities. CHAI's goal is to develop the conceptual and technical wherewithal to reorient the general thrust of AI research towards provably beneficial systems. CHAI is led by Prof. Stuart Russell.

Recommendations

Make the regulation future-proof and prepare for higher-risk systems

- Update the regulation to address increasingly generalized AI systems that have multiple purposes, such as OpenAI's Generative Pre-trained Transformer 3 and DeepMind's AlphaFold system
- Article 3(13) "reasonably foreseeable misuse" and "interaction with other systems": Explicitly consider the issue of high-risk and societal-scale consequences stemming from the interaction of many low-risk systems
- Article 6 classification rules for high-risk systems: Include recommender systems in the classification rules for high-risk systems
- Article 6 classification rules for high-risk systems: Include the requirements to document perceptual inputs, action outputs, objectives, and the operational environment for high-risk systems
- Article 6 classification rules for high-risk systems: Include the requirements to document time-of-sale properties of systems
- List of high-risk categories in Annex III: Add categories

Protect people from psychological harm

- Article 5 (1) (a) on psychological manipulation: Consider expanding the current definition of "subliminal techniques beyond a person's consciousness"
- Section 3.5 of the Explanatory Memorandum: Include the protection of mental integrity in the Explanatory Memorandum
- Article 53 (AI regulatory sandboxes): Recognize the limit of sandboxes

Clarify key components of the regulation

Article 3 (1) Definition of AI: Delete the mention of "a given set of human-defined objectives"

Article 3 (16): Modify the notion of "recalling" an AI system

Article 3 (33, 34, 35): Redefine the notion of "biometric data" more generally

Article 10 on data governance: Clarify training, validation, and test stipulation

Article 12: Clarify the primary purposes of logging

Article 13 on transparency: Transparency requirements should require internal processing to be made available and understandable

Article 15.3: Clarify the issue of performativity ("feedback loops")

Article 64: Clarify the type of access granted to the datasets used by the provider

Article 5.1(c) on social scoring: Clarify the application of social scoring and reinforce protection from manipulative, exploitative, and social control practices

The EU AI Act is an important step in the right direction toward developing beneficial AI. We particularly welcome the following elements:

- The broad definition of high-risk systems as those potentially impacting health, safety, and fundamental rights, with flexibility to expand
- The recognition and prohibition of "powerful tools for manipulative, exploitative and social control practices"
- The establishment of a permanent Board for oversight and gradual accumulation of expertise
- The prohibition on undeclared "bots" and false-content generation
- The inclusion of post-market monitoring
- The encouragement of codes of conduct along the same lines
- The requirement that a system has "in-built operational constraints that cannot be overridden by the system itself and is responsive to the human operator"
- The inclusion of potentially substantial penalties

We have identified three priorities, detailed in the following pages:

- Make the regulation future-proof and prepare for higher-risk systems
- Protect people from psychological harm
- Clarify key components of the regulation

Recommendations

I. Make the regulation future-proof and prepare for higher-risk systems

1. Update the regulation to address increasingly generalized AI systems that have multiple purposes, such as OpenAI's Generative Pre-trained Transformer 3 and DeepMind's AlphaFold system

Under the current draft, legal requirements depend on the intended purpose of an AI system in a given use case. Regulation is also restricted to AI systems that are directly placed into service or put on the market. However, some of the most impactful AI systems are *precursors* to multiple applications but are not *themselves* directly applied; thus, they may be exempt from regulatory scrutiny.

New systems such as OpenAI's Generative Pre-trained Transformer 3 (GPT-3) and DeepMind's AlphaFold have many possible purposes and could make the regulation obsolete. Future AI systems will be even more generalized than GPT-3 and AlphaFold. Instead of categorizing AI systems by a limited set of stated intended purposes, the regulation should require a complete risk assessment of all of the AI system's intended uses (and foreseeable misuses), incorporating the fact that generalized systems may be further adapted and developed by secondary developers or users.

Therefore, the proposal should require a complete risk assessment of all of an AI system's intended uses (and foreseeable misuses) instead of categorizing AI systems by a limited set of stated intended purposes. One way of achieving this would be to expand title IV ('Transparency obligations for certain AI systems') to include a limited number of additional requirements that apply across all AI applications regardless of their intended purpose. Further, **revising the definitions in Article 3 of the proposed Act to account for plural "use*s*" and "purpose*s*"** will improve the overall scope of the proposal to address these increasingly generalized AI systems.

2. Article 3(13) "reasonably foreseeable misuse" and "interaction with other systems": Explicitly consider the issue of high-risk and societal-scale consequences stemming from the interaction of many low-risk systems

High-risk and societal-scale consequences can arise from the interaction between many low-risk systems. One example of this is the 2010 Flash Crash: over the course of 15 minutes in 2010, high-speed trading algorithms sold many stocks at a low price, then immediately noticed that those

stocks were being sold at a low price. This led to the systems selling more at a lower price, causing stock values to spiral downwards. The resulting market instability caused a "flash crash" with a temporary loss of approximately USD $1T in market value. Many AI systems are not designed with safety as a primary goal. Often, the development of these AI systems fails to account for how the safety of their actions is affected by human behavior or the behavior of other AI systems. **While we admit this is a very difficult topic for legislation, we recommend considering the issue of high-risk consequences from the interaction of many low-risk systems.** Systems may not be directly harming individuals in obvious ways but may do so collectively, for example, by making users more vulnerable to extreme views.

Any assessment procedure should ensure that **AI providers consider the impact of their applications on society at large, for example, by including possible societal impact in the requirements for technical documentation (Annex IV).**

3. Article 6 classification rules for high-risk systems: Include recommender systems in the classification rules for high-risk systems

The Article 6 classification rules for high-risk systems generally seem to omit "recommender systems." But these present serious risks, e.g., by producing information flows that might affect stability by inducing ethnic conflict or that might influence operations by foreign actors. Young women are disproportionately affected by addictive content provided by those information systems. Since the introduction of widespread social media in 2009, women aged 15–19 have experienced a 70% increase in suicide rate, and women aged 10–14 have experienced a 151% increase.[1] On a separate issue, a United Nations report found that Facebook's content played a "determining role" in the 2018 Myanmar genocide of the Rohingya minority group.[2] **We recommend including recommender systems in the classification rules for high-risk systems.**

4. Article 6 classification rules for high-risk systems: Include the requirements to document perceptual inputs, action outputs, objectives, and the operational environment for high-risk systems

While there is ample discussion of the need to document datasets in the main body of the regulation, the key issue of documenting objectives is mentioned only briefly in Annex IV.2(b). Generally speaking, all AI systems can be characterized according to perceptual inputs, action outputs, objectives, and the environment in which the systems operate. (See Artificial Intelligence: A Modern Approach Ch. 2 and the OECD classification system.) **Therefore, the regulation should require documentation of all these aspects.** We are particularly concerned with the scope of action, i.e., the

[1] Centers for Disease Control and Prevention, *Increase in Suicide Mortality in the United States, 1999–2018*, April 2020.
[2] United Nations Human Rights Council, *Report of Independent International Fact-Finding Mission on Myanmar*, August 2018.

range of possible effects the system can have on its environment, including physical, psychological, social, and economic effects on persons therein, and whether the objectives of the system are aligned with the interests of the affected persons.

5. Article 6 classification rules for high-risk systems: Include the requirements to document time-of-sale properties of systems

While the regulation mentions systems that learn after being sold, not much is said about how to ensure they continue to behave in a way that is consistent with the requirements after the sale. **There needs to be more specificity about what must be documented about the properties of the system at the point when it leaves the provider;** for example, an analysis showing that the system will continue to satisfy requirements with high probability considering possible future data from which the system will learn.

6. List of high-risk categories in Annex III: Add categories

The list of categories in Annex III seems incomplete. We suggest adding:

- **Healthcare applications other than medical devices.** These could be personal healthcare apps, fitness devices (Fitbit, etc.), hospital triage and in-patient/out-patient decisioning, etc. Racial and gender bias is a serious concern here.
- **Systems that control access to news media and related content.**
- **Deliberately addictive immersive video games.**

II. Protect people from psychological harm

1. Article 5 (1) (a) on psychological manipulation: Consider expanding the current definition of "subliminal techniques beyond a person's consciousness"

The current definition of "subliminal techniques beyond a person's consciousness" in the text seems too restrictive. While this is an important point, we recognize it is hard to formalize. The canonical example for "subliminal" would be subliminal images in videos that impact behavior without registering at a conscious level. **Other techniques that may be covered by "subliminal" include:**

1. the addictive techniques developed by neuroscientists working on video games using carefully calibrated adrenaline stimulation frequencies;
2. the "dark patterns" used by marketers to, for example, extract valuable personal data as part of a spurious "sign-up" process;

3. the manipulative consequences of reinforcement learning (RL) algorithms in social media, which learn to send sequences of content items that modify a person's psychological state and tendencies so that they become more reliable consumers of certain types of (often extreme) content.

In the future, policymakers may consider prohibiting the use of reinforcement learning in public-facing social media algorithms that interact with natural persons sequentially by recommending or communicating content in order to maximize an internal metric related to clickthrough or engagement, when health, safety, and well-being are not sufficiently taken into account.

2. Section 3.5 of the Explanatory Memorandum: Include the protection of mental integrity in the Explanatory Memorandum

Article 3 of the European Union Charter of Fundamental Rights, which deals with the right to protection of **mental integrity**, among other topics, should be mentioned prominently in Section 3.5 of the Explanatory Memorandum**. We strongly recommend Article 3 of the European Union Charter of Fundamental Rights and its mention of mental integrity in the Explanatory Memorandum.**

3. Article 53 (AI regulatory sandboxes): Recognize the limit of sandboxes

Unless there are real humans inside the sandbox, it is hard to see how this could reveal (say) potential psychological harms. **We suggest that Article 13 recognizes the limits of sandboxes for unknown risks and takes a prudent approach in dealing with psychological harms.**

III. Clarify key components of the regulation

1. Article 3 (1) Definition of AI: Delete the mention of "a given set of human-defined objectives"

In the definition of AI, the mention of "for a given set of human-defined objectives" seems to require the objectives to be explicitly present in the AI system or for the system to accept objectives as input. That leaves out systems with implicit objectives. For example, in a self-driving car, the objective to avoid collisions with pedestrians is not explicitly present anywhere in the system, although the human-supplied destination is explicit. In the "new model" for AI proposed in *Human Compatible* by Stuart Russell[3], AI systems will operate with explicit uncertainty about what the human objectives are.

[3] Russell, S., 2019. *Human Compatible*. 1st ed. Penguin Books.

It would be unfortunate if these were considered not to be AI systems. **We recommend striking that part of the definition, and we think that the sentence works well without it.**

We agree that the Annex I AI techniques should be expandable as new methods arise. **We suggest including "genetic programming" and natural language processing (NLP) methods other than deep learning, such as parsing and information extraction.**

2. Article 3 (16): Modify the notion of "recalling" an AI system

The concept of "recalling" in the sense of returning the AI system to the provider is not adapted to a software system. **We suggest ensuring the provision of a way to immediately and remotely disable the software systems, with appropriate warnings, safety measures, user sign-off, etc.**

3. Article 3 (33, 34, 35): Redefine the notion of "biometric data" more generally

There seems to be a problem with "biometric" being defined too restrictively in Article 3 (33) (i.e., data sufficient for identification), and that definition is inappropriate for (34) and (35). Emotion recognition and categorization of persons do not require this type of biometric data. W**e suggest defining "biometric data" generally** (i.e., any measurement of a person's physical body or physical behavior) **and "identifying biometric data" more strictly.**

4. Article 10 on data governance: Clarify training, validation, and test stipulation

The stipulation for training, validating, and testing AI systems is both too prescriptive and too weak for some cases. **We suggest instead requiring "appropriate and rigorous statistical methodology."** The mention of "possible biases" is also quite vague. The mention that datasets should be "representative, free of errors and complete" is unclear as the definition of representativity needs to be refined, being free of errors is almost impossible for large datasets, and having a "complete" dataset may be impossible.

5. Article 12: Clarify the primary purposes of logging

The primary purposes of logging include the ability to interrogate the system's decisions and the ability to reproduce failures for forensic and diagnostic purposes. For some types of systems, e.g., a self-driving car, this is a very stringent requirement, but arguably a necessary one. It involves careful checkpointing, data recording, version control, etc.

6. Article 13 on transparency: Transparency requirements should require internal processing to be made available and understandable

Article 13 mentions that high-risk AI systems should be "sufficiently transparent to enable users to interpret the system's output and use it appropriately." This is an unusual meaning of "transparent," as it says nothing about internal processing being available and/or understandable, only that the output itself is understandable. For example, if the machine says "Guilty! Sentence is death!"—that is perfectly clear and operationalizable. **This transparency requirement should be more stringent and require internal processing to be made available and understandable.**

7. Article 15.3: Clarify the issue of performativity("feedback loops")

We appreciate that the regulation mentions the issue of performativity ("feedback loops"). **We suggest clarifying the issue of performativity.** It is often not "outputs used as inputs," but outputs indirectly influencing future inputs. For example, once people understand the acceptance criteria for automated credit decisioning, new applicants may look statistically different from those on which the system was trained and may include a much higher percentage of fraudulent applications ("adverse selection"). This was partially responsible for the Great Recession of 2008–9. For a good reference on a new mathematical theory of such learning systems, see Perdomo et al., 2020, "Performative prediction."

8. Article 64: Clarify the type of access granted to the datasets used by the provider

Article 64 mentions "full access to the training, validation and testing datasets used by the provider." What access would be granted, if any, for systems not based solely on statistical learning? These non-ML systems are mostly neglected in this regulation. **We suggest clarifying the type of access granted to the datasets used by the provider.**

9. Article 5.1(c) on social scoring: Clarify the application of social scoring and reinforce protection from manipulative, exploitative, and social control practices

Article 5.1(c) on social scoring needs clarification on its application. It is unclear whether this article would apply to using previous criminal records in determining the length of custodial sentences, or to other uses. While it is highly desirable to orient AI towards non-social-control practices, **this article needs to be clarified to protect citizens from "powerful tools for manipulative, exploitative and social control practices."** Other provisions should also be given to protect people from this kind of practice done by non-state actors such as corporations, political parties, and foreign governments.