Consultation on the White Paper on AI – a European approach
Submission by Haydn Belfield, José Hernández-Orallo, Seán Ó hÉigeartaigh, Matthijs M. Maas, Alexa Hagerty, Jess Whittlestone
June, 2020

Executive Summary

We are a group of academic researchers on AI, with positions at the Universitat Politecnica de Valencia, the University of Copenhagen, the University of Cambridge, and the Leverhulme Centre for the Future of Intelligence – a leading international centre in AI ethics. We have published dozens of technical papers on machine learning and artificial intelligence and reports and white papers on the ethics and governance of artificial intelligence.

Our submission mainly focuses on the "Ecosystem of Trust: Regulatory Framework for AI", and in particular the mandatory conformity assessments for high-risk AI applications carried out by independent testing centres. **Our key recommendation is to keep this proposed structure and not water it down.**

In this submission we 1) support the Commission's proposed structure, defend this approach on technical, policy, practical, and ethical grounds, and offer some considerations for future extensions; and 2) offer some specific recommendations for the mandatory requirements.

There are several important concerns about current AI systems, such as safety, security, impact on fundamental rights, discrimination, and explainability. These concerns cannot be adequately addressed by current legislation, so we need new rules: mandatory requirements and conformity assessments by independent testing centres. The introduction of these new rules should begin with AI systems used in sectors and high-risk application areas where the general concerns are particularly acute. It is a sensible approach to determine high-risk application areas with the two cumulative criteria, while providing for exceptional instances that should be considered high-risk as such. It is reasonable to begin with high-risk sectors including healthcare, transport, energy, and parts of the public sector, while retaining the flexibility to review and amend this list. The proposed mandatory requirements are proportionate, important, and practical on a technical and operational level. Self-assessment cannot ensure that AI is trustworthy. Compliance should mainly be assessed ex-ante by means of an external conformity assessment procedure by an independent testing centre. There should also be provision for additional ex-post market surveillance and enforcement, for example for software updates and systems that keep learning during operation.

With this framework, the EU will demonstrate global leadership and profoundly shape global standards. It will help dissolve the presumed tension between respecting citizen's rights and company competitiveness, creating a more level playing field for EU companies. This framework strikes the right balance between overly burdensome regulation and allowing citizens to be placed in harm's way.

Our recommendations are summarised here:

- Concerns about AI
    - Consider including security explicitly in the list of concerns.
- High-risk application areas
    - Keep "significant immaterial damage".
    - Keep "periodic reviews and amendments".
    - Keep the "exceptional instances" clause.
    - Consider other sectors, such as political advertising on social networks and airborne swarms.
    - Consider adding 'replacing professionals with specific legal responsibilities' as an 'exceptional instance'.
    - Consider extending to a multi-tier risk determination.
    - Consider societal risks and how they scale.
- Mandatory requirements
    - Data quality
        - Keep the requirements around broad data sets - this can and should be done in ways that respect GDPR and copyright
        - Further specify and provide clear benchmarks for "all relevant scenarios", "dangerous situations" and "sufficiently representative".
    - Record-keeping
        - Keep the requirements around data set record-keeping and sharing - this can and should be done in ways that respect GDPR and copyright.
        - Consider adding record-keeping of the computational resources (or 'compute') used to build, test and validate the AI systems.
        - Consider adding record-keeping of AI incidents.
        - Consider options for audit trail requirements and documentation of AI systems.
    - Information
        - Keep the requirement for citizens to be *proactively* informed *whenever* they are interacting with an AI system.
        - Consider adding that if a citizen is interacting with a human, but an AI system is playing a substantive role in decision-making, then the citizen should be informed.
    - Robustness and accuracy
        - Robustness should be interpreted as whether it *in fact* adequately deals with errors (safety) and attacks (security)
        - Consider requirements that prohibit unexpected system behaviours, including preventing the system from operating, if inputs or outputs fall outside a predefined "safe" range.
        - Consider also paying attention to 'systemic' safety risks.
        - Further specify and provide clear benchmarks for "reproducible outcomes".

        - Further specify and provide clear benchmarks for “mitigating measures”, such as red-teaming, bias or safety ‘bug bounties’, and hardware security.
    - Human oversight
        - Consider requirements to ensure the efficacy of human oversight.
        - Further specify in which areas particular forms of human oversight will be required.
    - General
        - ‘Fauxtomation’ should not allow companies to dodge the requirements.
        - These requirements should also apply to suppliers of multi-purpose AI components.
- Conformity assessments
    - Keep the commitment to ex-ante, external conformity assessments by independent testing centre(s), supplemented with additional ex-post market surveillance and enforcement.
    - Do not consider a ‘grandfather clause’ for AI systems deployed before the regulation comes into place.
    - Consider only allowing limited confidential testing and piloting in coordination with the independent testing centres.

More detail on all of these points can be found below, and we would be delighted to answer any further questions the Commission might have.

An ecosystem of trust

Concerns about AI

In your opinion, how important are the following concerns about AI?
All very important: Safety, fundamental rights, discrimination, explainability, redress (compensation) and accuracy.
Do you think that the concerns expressed above can be addressed by applicable EU legislation? If not, do you think that there should be specific new rules for AI systems?
There is a need for a new legislation

There are several very important concerns about current AI systems. To a limited extent, these concerns can be addressed through applying and adapting existing EU legislation. However, these concerns cannot be adequately addressed by current legislation, so we need specific new rules.

There are serious concerns about AI, as shown by a series of scandals over the last few years[1]. AI systems have been criticised for contributing to the addiction of consumers[2] and

[1] Whittaker, Meredith, et al. *AI Now Report 2018*. Dec. 2018, https://ainowinstitute.org/AI_Now_2018_Report.pdf. https://ainowinstitute.org/AI_Now_2019_Report.pdf

[2] Alter, Adam. *Irresistible - The Rise of Addictive Technology and the Business of Keeping Us Hooked*. Penguin Random House Usa Ex, 2017.

the hijacking of their attention[3]. AI business models have been criticised for relying on work that is exploitative, insecure, low-paid, and degrading[4], as well as engaging in extractive "surveillance capitalism."[5] AI systems have been criticised for fueling polarisation and political violence, upsetting elections and referendums[6]. Other concerns include the amplification of biases in datasets,[7] contributing to discriminatory hiring[8], loss of privacy,[9] emergent manipulative behavior,[10] at-scale generation of disinformation,[11] and social harms associated with facial recognition,[12] predictive policing[13], migration management[14] and criminal risk assessment.[15] Other risks could involve powerful systems adopted for malicious use,[16] safety failures in critical infrastructure (such as major failures in traffic safety or energy grids),[17] the development of advanced systems that are difficult to understand and control,[18] and the disruption of economic and social stability,[19] including through large-scale work displacement and changes in the nature of work itself[2021]. A recent survey of businesses suggests that, despite pushes for responsibility, the majority of organisations using AI are

---

[3] Williams, J. 2018. *Stand Out Of Our Light: Freedom and Resistance in the Attention Economy*. Cambridge Univeristy Press.
[4] Khan, L. M. 2016. Amazon's antitrust paradox. *Yale Law Journal*, *126*, 710.
[5] Zuboff, S. 2019. The age of surveillance capitalism. Profile Books.
[6] Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives*, *31*(2), 211-36.
Fink, C. 2018. Dangerous speech, anti-muslim violence, and facebook in myanmar. *Journal of International Affairs*, *71*(1.5), 43-52.
Klein, E. 2020. Why We're Polarised. Simon & Schuster.
Cadwalladr, C. (2019). Facebook's role in Brexit — and the threat to democracy. *TED*.
[7]Barocas, Solon and Selbst, Andrew D., Big Data's Disparate Impact (2016). 104 California Law Review 671 (2016).
Hao, Karen. (2019). This Is How AI Bias Really Happens-and Why It's so Hard to Fix. *MIT Technology Review*.
[8] Ajunwa, Ifeoma. (2019). Automated Employment Discrimination.
[9] ScienceDaily. (2019). Artificial Intelligence Advances Threaten Privacy of Health Data. *ScienceDaily*.
[10] Russell, Stuart. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control.* Viking.
[11] Solaiman, Irene et al., 2019. Release Strategies and the Social Impacts of Language Models. arxiv:1908.09203.
[12] Stanley, Jay. (2019). The Dawn of Robot Surveillance: AI, Video Analytics, and Privacy. ACLU.
Stark, Luke. 2019. Facial recognition is the plutonium of AI. XRDS 25, 3 (Spring 2019), 50–55.
Joy Buolamwini, Joy & Gebru, Timnit. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of Machine Learning Research 81:1–15, Conference on Fairness, Accountability, and Transparency.
[13] Broadhurst, Roderic, Maxim, Donald, Brown, Paige & Trivedi, Harshit. (2019). Artificial Intelligence and Crime: A Report for the Korean Institute of Criminology.
UNICRI & INTERPOL. (2019). Artificial Intelligence and Robotics for Law Enforcement.
Vestby, Annette & Vestby, Jonas. (2019). Machine Learning and the Police: Asking the Right Questions. Policing: A Journal of Policy and Practice.
[14] Molnar, Petra. (2019). Technology on the margins: AI and global migration management from a human rights perspective. Cambridge International Law Journal.
Beduschi, Ana. (2020). International migration management in the age of artificial intelligence. Migration Studies.
[15] Angwin, Julia, et al. (2019). Machine Bias. *ProPublica*; (2018).
The Use of Pretrial 'Risk Assessment' Instruments: A Shared Statement of Civil Rights Concerns.
Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System. *The Partnership on AI*, 2 Aug. 2019, .
[16] Brundage, Miles & Avin, Shahar et al. (2018) The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. arXiv:1802.07228.
[17] Osoba, Osonde & William Welser. The Risks of AI to Security and the Future of Work. *RAND Corporation*.
[18] Castelvecchi, Davide. (2016). Can We Open the Black Box of AI? *Nature*, 538:7623:20–23.
[19] Bostrom, Nick, et al. (2018). *Public Policy and Superintelligent AI: A Vector Field Approach*.
[20] Gray, Mary and Suri, Siddharth. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass.*
[21] Frey, Carl Benedikt & Osborne, Michael. (2018). The Future of Employment. Oxford Martin Programme on Technology and Employment.

not taking efforts to mitigate the full range of risks that they may cause[22]. Studies and polls indicate a lack of trust in AI development[23][24][25].

It is our assessment that these concerns are not being, and cannot be, adequately addressed through existing EU legislation. As such, we need new legislation: mandatory conformity assessments for high-risk AI applications carried out by independent testing centres.

As AI is increasingly used in real-world contexts, processes using AI should be subject to similar safety and impact expectations as other engineering processes. In other engineering disciplines, it is reasonable and expected to fulfill stringent requirements when operating in a high-risk application area. One has to fulfill these requirements before building a bridge or a power plant. It does not make sense for the self-driving car driving over the bridge, or the AI system operating the plant, to have to fulfill less stringent - or even no - requirements.

Recommendation: Consider including security explicitly in the list of concerns.

The concerns raised in the White Paper - safety, fundamental rights, discrimination, explainability, redress (compensation) and accuracy - are all very important. It would be appropriate to also include security as one of the concerns. “Safety” generally refers to proper internal functioning of an AI system and the avoidance of unintended harms, while “security” addresses external threats to an AI system and the malicious use of AI as a tool for attacks[26]. Security challenges are a major concern, and the problem surface of security threats (the actors and kind of solutions needed) is arguably quite distinct from the problem surface of safety or ethics challenges. The importance of security to attacks and manipulation attempts is mentioned several times throughout the White Paper, especially in the mandatory requirements around robustness (5.D.d).

High risk applications of AI

If you think that new rules are necessary for AI system, do you agree that the introduction of new compulsory requirements should be limited to high-risk applications (where the possible harm caused by the AI system is particularly high)?
Yes.
Do you agree with the approach to determine “high-risk” applications of AI?
Yes.

[22] Cam, Arif, Chui, Michael, & Hall, Bryce. (2019). Global AI Survey: AI proves its worth, but few scale impact. McKinsey.
[23] Edelman. (2019). Edelman Trust Barometer.
[24] Zhang, Baobao & Dafoe, Allan. (2019). Artificial Intelligence: American Attitudes and Trends. Future of Humanity Institute, University of Oxford.
[25] Belfield, Haydn. (2020). Activism by the AI Community: Analysing Recent Achievements and Future Prospects. Proceedings of the AAAI/ACM Annual Conference on AI, Ethics, and Society.
[26] Fjeld, Jessica, Nele Achten, Hannah Hilligoss, Adam Nagy & Madhulika Srikumar. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI. Berkman Klein Center for Internet & Society.
Hayward, Keith J., & Maas, Matthijs M.. (2020). Artificial Intelligence and Crime: A Primer for Criminologists. Crime, Media, Culture.

Concerns about AI are particularly acute in high-risk application areas, so new compulsory requirements should be introduced in these high-risk areas. It is a sensible approach to determine high-risk application areas with the two cumulative criteria, while making provision for exceptional instances. It is reasonable to begin with healthcare, transport, energy, and parts of the public sector, while keeping the flexibility to review and amend this list.

It is reasonable that the *introduction* of new requirements be limited to high-risk application areas. The high-risk areas are those where it is particularly urgent and pressing that we act now to prevent serious harm to EU citizens. Furthermore, the concerns about applying AI in a range of areas are often particularly concerning in high-risk areas. We should be concerned by AI-enabled discrimination, or any safety failure, but they are particularly concerning in a situation where people's liberty or lives are at stake. On a practical level, we can make sure the system of requirements and conformity assessments is working well, then consider to what extent to extend.

Determining high-risk application areas using the two cumulative criteria and "exceptional instances" clause is a flexible, sensible, and proportionate approach. Identifying application areas within sectors is sensible, as it incorporates an assessment of severity and likelihood within itself - the application areas are those in which severe harm is so probable that conformity assessments are the only way to protect EU citizens[27]. It is good that the White Paper specifies how it will determine 'high-risk'. It is not enough for case-by-case assessments of whether a product or service is high risk to be made by the developers and deployers themselves, as this creates legal, regulatory and operational uncertainty.

The *determination* of high-risk sectors and application areas should not be affected by operational factors, such as the degree of human control and internal governance. These operational factors form part of the conformity assessment as to whether a particular AI system should be allowed in the EU, but they should not contribute to the determination of whether a particular application area is high-risk. The application area *as such* is so high-risk that only systems that meet the mandatory requirements are acceptable - fulfilling these requirements doesn't make the application area itself less high-risk.

Recommendation: Keep "significant immaterial damage"

This is needed to cover significant effects for the rights of an individual or company, damages in the form of significant economic loss or reputational effects, and damages relating to data protection and privacy, non-discrimination, defamation and freedom of expression.

Recommendation: Keep "periodic reviews and amendments"

We need a flexible approach to determining risk for this fast-changing and rapidly-improving technology. It is especially welcome that the White Paper foresees that the framework will be "periodically reviewed and amended where necessary in function of relevant developments

[27] Schuett, Jonas. (2019). A Legal Definition of AI. arXiv:1909.01095.

in practice". Frequent and timely evaluations - indeed ongoing evaluation[28] - of the regulatory framework are appropriate for this rapidly-developing issue-area.

On the specific proposed sectors: it is sensible for the initial sector list to be healthcare, transport, energy, and parts of the public sector (asylum, migration, border controls, judiciary, social security, employment services). However, "parts of the public sector" should include policing.

Sensibly, the White Paper does not exhaustively list all the high-risk application area within these sectors. However as indicative examples, some of the high-risk application areas within these sectors include:

- Healthcare:
    - Application of 'black-box medicine'
    - Public health (e.g. pandemic) modelling
- Transport:
    - Self-driving cars
    - Autonomous shipping
    - 'Last-mile' delivery robots in the public space
- Energy:
    - Grid management
    - Nuclear power plant operations
- Public sector
    - Policing
        - Predictive policing
        - Bomb disposal robots

Recommendation: Keep the "exceptional instances" clause

It is also sensible to have an 'exceptional instances' clause, determining that the use of AI applications for certain purposes should be considered as high-risk as such, irrespective of the sector concerned. It covers cross-cutting exceptional instances, while providing sufficient clarity for companies to have confidence about whether a specific application is in or out of scope. It is also a flexible structure that could be added to with future 'exceptional instances' as and when these emerge. In particular, due to their impact on fundamental rights it is sensible to always consider as "high-risk" the use of AI applications for recruitment processes as well as in situations impacting workers' rights, and for the purposes of remote biometric identification and other intrusive surveillance technologies. This latter exceptional instance clearly addresses concerns about some forms of persistent facial recognition in public spaces. But it is also usefully 'technology-neutral', as it prima facie applies to other concerning forms of remote, continuous AI surveillance, including gait analysis[29] and laser-measured heartbeat identification[30].

---

[28] Including anticipation, foresight and futureproofing.
Laurie, Graeme, Harmon, Shawn HE & Arzuaga, Fabiana. (2012). Foresighting Futures: Law, New Technologies, and the Challenges of Regulating for Uncertainty, Law, Innovation and Technology, 4:1, 1-33.
Ranchordas, Sofia & van 't Schip, Mattis. (2020). Future-Proofing Legislation for the Digital Age, in S. Ranchordas & Y. Roznai (Eds), *Time, Law, and Change*.

[29] Horst, F., Lapuschkin, S., Samek, W. et al. (2019). Explaining the unique nature of individual gait patterns with deep learning. Sci Rep 9, 2391.

[30] Hambling, David. (2019). The Pentagon has a laser that can identify people from a distance. MIT Technology Review.

We are strongly supportive of the introduction of this framework. As time goes on and the framework proves its usefulness, the Commission should expand the framework. We make the following recommendations for areas the Commission should consider in the near future.

Recommendation: Consider other sectors.
Other sectors and application areas worth future consideration as either 'high-risk application areas', or as 'exceptional instances' include: automated trading,[31] political advertising on social networks, cybersecurity and airborne swarms[32].

Recommendation: Consider adding 'replacing professionals with specific legal responsibilities' as an 'exceptional instance'.

The use of AI applications to replace professionals with specific responsibilities could always be considered as "high-risk", under 'exceptional instances'. If a system performs a portion of, or replaces entirely, a function of a human actor that has certain legal responsibilities[33] to another individual, those legal responsibilities should be reciprocated by the responsible agent making use of the AI system. To be specific, if a system is providing financial, legal or medical information, advice or recommendations (such as through personal assistants or other 'AI extenders'[34]) then a human responsible for that system should have the same fiduciary obligations, client obligations or patient obligations as a human financial advisor, lawyer or doctor would. The Commission should specify how those obligations are distributed amongst the developer and operator of the AI system.

Recommendation: Consider extending to a multi-tier risk determination.

The Commission should consider extending its framework to include different tiers of risk, as proposed by the German Data Ethics Commission[35]. Moving beyond a binary 'high-risk or not' determination would allow more nuance and distinctions, and enable the application of more precise and targeted regulatory interventions. It would also be consistent with the EU's approach to a multi-layered risk assessment framework for medical devices in the Medical Device Directive (MDD) and Medical Device Regulations (MDR).

Recommendation: consider societal risks and how they scale.

It will be important to consider application areas that may not pose immediate "high-risk" to an individual, but do pose substantial societal risks because of their wider effect. Beyond accidents and misuse, AI poses societal (or 'structural'[36]) risks that change the environment and incentives people face. Widespread disinformation and manipulation, amongst other

---

[31] Given algorithmic flash crashes, and the propensity of reinforcement learning trading algorithms to (albeit in simulation) self-learn fraudulent or illegal trading strategies.
[32] Brundage, Miles & Avin, Shahar et al. (2018) The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. arXiv:1802.07228.
[33] Such as decisions affecting fundamental rights, and more broadly where 'administrative discretion' is involved.
[34] Hernández-Orallo, José & Vold, Karina. (2019). AI Extenders: The Ethical and Societal Implications of Humans Cognitively Extended by AI. Proceedings of the AAAI/ACM Annual Conference on AI, Ethics, and Society.
[35] Data Ethics Commission of the Federal Government. (2019). Opinion of the Data Ethics Commission.
[36] Remco Zwetsloot, Allan Dafoe (2019). Thinking About Risks From AI: Accidents, Misuse and Structure. Lawfare.
Schim van der Loeff, Agnes, Bassi, Iggy, Kapila, Sachin & Gamper, Jevgenij. (2019). AI Ethics for Systemic Issues: A Structural Approach. arXiv:1911.03216

psychological or cultural effects, should also be considered high-risk. For example, synthetic media such as deepfakes and AI-generated text can undermine our "political security"[37].

This consideration should take into account how scale affects risk. It seems reasonable that systems that affect tens of millions of EU citizens should receive more scrutiny than those that affect thousands. Some systems (e.g. recommender systems used in social media) may not be high-risk when they affect thousands, but pose high societal risks when they affect tens of millions. Acknowledging scale could also help SMEs.

Feedback on mandatory requirements

In your opinion, how important are the following mandatory requirements of a possible future regulatory framework for AI?
All very important: Data quality, record-keeping, information, robustness and accuracy, human oversight, clear liability and safety.

The proposed mandatory requirements are proportionate, important, and practical on a technical and operational level. In the below we offer further commentary on the specific mandatory requirements.

Data quality

> *"reasonable assurances that the subsequent use of the products or services that the AI system enables is safe. AI systems should be trained on data sets that are sufficiently broad and cover all relevant scenarios needed to avoid dangerous situations."*
> *"obligations to use data sets that are sufficiently representative, especially to ensure that all relevant dimensions of gender, ethnicity and other possible grounds of prohibited discrimination are appropriately reflected in those data sets"*

Recommendation: Further specify and provide clear benchmarks for "all relevant scenarios", "dangerous situations" and "sufficiently representative".

As the overall framework is put in place, it will be helpful, as the White Paper suggests, to provide further specification of these terms. It would also be useful to specify the treatment of data sets where synthetic data is used to complement training data.

Recommendation: Keep the requirements around broad data sets - this can and should be done in ways that respect GDPR and copyright

Requiring the use of broad data sets, that promote safety and reduce discrimination and bias, can and should be done in ways that respect GDPR and copyright. As systems are often trained on several datasets, the Commission could clarify that all relevant dimensions should be reflected *across* those data sets.

Record-keeping

> *prescribe that the following should be kept:*
> *"accurate records regarding the data set used to train and test the AI systems, including a description of the main characteristics and how the data set was selected"*

[37] Brundage, Miles & Avin, Shahar et al. (2018) The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. arXiv:1802.07228.

"in certain justified cases, the data sets themselves"
"documentation on the programming [objective function, weights] and training methodologies, processes and techniques used to build, test and validate the AI systems"

Record-keeping will be essential for fulfilling the other requirements, providing information to the independent testing centres on, for example, the extent to (and methods by) which the AI systems were tested for safety, security or ethical concerns, and the sources of data, labour, and other resources used.

Recommendation: Keep the requirements around data set record-keeping and sharing - this can and should be done in ways that respect GDPR and copyright

This may not be used in all cases, but in some cases the full details about AI models, the underlying code or data sets themselves will need to be shared for testing and inspection by competent authorities. In many cases, testing the system (by an independent testing centre) against clear benchmarks, performance standards and confidence levels may be sufficient. But there will be some cases in which the training data itself will have to be assessed. As the White Paper notes, sensible safeguards should be put in place so that this can happen while protecting business confidentiality, not enabling adversarial gaming, and not infringing GDPR or copyright.

Recommendation: Consider adding record-keeping of the computational resources (or 'compute') used to build, test and validate the AI systems.

We strongly suggest that measures of the computational resources (or 'compute') used to build, test and validate the AI systems also be kept. Compute is a key driver of AI progress and performance, of similar importance to data and better algorithms. Better compute tracking and reporting was one of the key recommendations of our Toward Trustworthy AI: Mechanisms for Supporting Verifiable Claim report[38]. The report notes that compute reporting is a "building block of credible third party oversight of AI projects: an auditor might note, for example, that an organization has more computing power available to it than was reportedly used on an audited project, and thereby surface unreported activities relevant to that project."

Recommendation: Consider adding record-keeping of AI incidents.

Records and data of AI 'incidents' (such as any notable bugs, unexpected behaviours, 'edge' cases', overfitting, or reward hacking) should also be kept. These incidents could be encountered during the development process or in the deployed final product. These could contribute to an AI incident-sharing database, in which incidents are shared in an anonymous (and even aggregated) way, to create common knowledge of risks, vulnerabilities and failure modes of AI systems[39]. Such information-sharing could improve system design and the conformity assessments. This could build on the pilot database started by the Partnership on AI, and several public compilations[40].

---

[38] Brundage, Miles, Avin, Shahar, Wang, Jasmine, Belfield, Haydn & Krueger, Gretchen et al. (2020). Toward Trustworthy AI: Mechanisms for Supporting Verifiable Claims. arXiv:2004.07213.
[39] Brundage, Miles, Avin, Shahar, Wang, Jasmine, Belfield, Haydn & Krueger, Gretchen et al. (2020). Toward Trustworthy AI: Mechanisms for Supporting Verifiable Claims. arXiv:2004.07213.
[40] Lehman, Joel et al. (2019). The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities. arXiv:1803.03453.

Recommendation: Consider options for audit trail requirements and documentation of AI systems.

Another recommendation of the report[41] was to adopt audit trail requirements. An audit trail is a traceable logs of steps taken in problem-definition, design, training and development, testing and system operation. Audit trails are used in several high-risk application areas, such as flight data recorders on commercial aircraft. IEC 61508 is a basic functional safety standard, used in many application areas, that specifies a required audit trail. Several tools can be useful in managing recordkeeping and compliance claims, such as the Assurance and Safety Case Environment (ACSE), or version control tools such as GitHub or GitLab, or proposed tools such as the Verifiable Data Audit. The Commission should allow some flexibility in the records or evidence that the independent testing centres can accept to satisfy traceability requirements (such as test logs or verification and validation activities).

The Partnership on AI has a useful workstream on Annotation and Benchmarking on Understanding and Transparency of Machine learning Lifecycles (AboutML). It is collating best-practice on record-keeping and documentation. Other interesting frameworks on which to build include outcomes- or claim-based "assurance frameworks" such as the Claims-Arguments-Evidence framework (CAE) and Goal Structuring Notation (GSN), which are already in wide use in safety-critical auditing contexts.

Information

> *"Ensuring clear information to be provided as to the AI system's capabilities and limitations, in particular the purpose for which the systems are intended, the conditions under which they can be expected to function as intended and the expected level of accuracy in achieving the specified purpose. This information is important especially for deployers of the systems, but it may also be relevant to competent authorities and affected parties."*
>
> *"Citizens should be clearly informed when they are interacting with an AI system and not a human being. Whilst EU data protection legislation already contain certain rules of this kind, additional requirements may be called for to achieve the above mentioned objectives."*

Recommendation: Keep the requirement for citizens to be *proactively* informed *whenever* they are interacting with an AI system.

Citizens should be clearly and proactively informed whenever they are interacting with an AI system. This should not be limited to that fact being merely discoverable (for example, buried in the terms and conditions), or to a situation in which an AI system is playing a substantive role in decision-making. Due to rapid improvements in video, voice and text generation, it is becoming less and less "immediately obvious" to citizens that they are interacting with an AI system. As a clear example, whenever a citizen is interacting with a chatbot (video, voice or text) they should be proactively informed of that fact.

Shankar Siva Kumar, Ram et al. (2019). Failure Modes in Machine Learning Systems. arXiv:1911.11034.
Krakovna, Victoria et al. (2020). Specification gaming: the flip side of AI ingenuity. DeepMind.

[41] Brundage, Miles, Avin, Shahar, Wang, Jasmine, Belfield, Haydn & Krueger, Gretchen et al. (2020). Toward Trustworthy AI: Mechanisms for Supporting Verifiable Claims. arXiv:2004.07213.

Recommendation: Consider adding that if a citizen is interacting with a human, but an AI system is playing a substantive role in decision-making, then the citizen should be informed.

If a citizen is interacting with a human, but an AI system is playing a substantive role in decision-making, then the citizen should also be clearly informed of that fact. A clear example would be when a citizen is informed of social security or employment services decisions by a human, but an AI system played a substantive role in those decisions.

Robustness and accuracy

> *"ensuring that the AI systems are robust and accurate, or at least correctly reflect their level of accuracy, during all life cycle phases"*
> *"Requirements ensuring that outcomes are reproducible;*
> *"Requirements ensuring that AI systems can adequately deal with errors or inconsistencies during all life cycle phases.*
> *"Requirements ensuring that AI systems are resilient against both overt attacks and more subtle attempts to manipulate data or algorithms themselves, and that mitigating measures are taken in such cases."*

Recommendation: Robustness should be interpreted as whether it *in fact* adequately deals with errors (safety) and attacks (security)

Robustness should not be interpreted as intentionally *engineering or designing* the AI system to cope with failure and adapt to new situations, but rather whether it *in fact* adequately deals with errors (safety) and attacks (security). There should be extensive processes for testing, evaluation, verification and validation (TEVV) processes for both safety and security risks. These should happen pre-launch, and at appropriate intervals post-launch.

Recommendation: Consider requirements that prohibit some unexpected system behaviours, including preventing the system from operating, if inputs or outputs fall outside a predefined "safe" range.

On adequately dealing with errors or inconsistencies: the high-risk application areas to which these requirements apply are definitionally those in which mistakes are hard to reverse and have extreme consequences. In recognition of that fact, there need to be measures in place that prohibit some classes of unexpected system behaviours, including preventing the system from operating, if inputs or outputs fall outside a predefined "safe" range. It is not adequate to just check for anomalies and errors and have remediation processes.

Recommendation: Consider also paying attention to 'systemic' safety risks.
Focus should also be paid to 'systemic' safety risks from interaction *amongst* AI systems. For example, an algorithmic trading AI system on its own may be considered 'safe', but interactions between many such systems can lead to emergent behaviour such as 'flash crashes'. AI services and products that are tightly coupled with other AI systems should not be assessed or tested in isolation.

Recommendation: Further specify and provide clear benchmarks for "reproducible outcomes".

As the overall framework is put in place, it will be helpful, as the White Paper suggests, to further specify and provide clear benchmarks for "reproducible outcomes". Most AI systems are not deterministic, so outcomes may not be exactly reproducible. Operationalisation could address which outcomes should be reproducible (e.g. the underlying techniques, or the individual results or patterns of a behaviour of a specific system), and what standards of reproducibility (as there are different statistical measures).

Recommendation: Further specify and provide clear benchmarks for "mitigating measures", such as red-teaming, bias or safety 'bug bounties', and hardware security.

We suggest several "mitigating measures" to attacks and manipulation attempts in Toward Trustworthy AI: Mechanisms for Supporting Verifiable Claims[42], such as red-teaming and bias or safety 'bug bounties'. We also made suggestions at the hardware level, such as new hardware security features for specialised AI chips, or the use of secure general computing hardware, including secure enclaves (also known as Trusted Execution Environments). Also useful are threat modelling, and testing in an adversarial setting.

Human oversight

> *Human oversight could have the following, non-exhaustive, manifestations:*
> *"the output of the AI system does not become effective unless it has been previously reviewed and validated by a human (e.g. the rejection of an application for social security benefits may be taken by a human only)"*
> *"the output of the AI system becomes immediately effective, but human intervention is ensured afterwards (e.g. the rejection of an application for a credit card may be processed by an AI system, but human review must be possible afterwards)"*
> *"monitoring of the AI system while in operation and the ability to intervene in real time and deactivate (e.g. a stop button or procedure is available in a driverless car when a human determines that car operation is not safe)"*
> *"in the design phase, by imposing operational constraints on the AI system (e.g. a driverless car shall stop operating in certain conditions of low visibility when sensors may become less reliable or shall maintain a certain distance in any given condition from the preceding vehicle)"*

Recommendation: Consider requirements to ensure the efficacy of human oversight.

AI systems interact with humans and organizations in complicated ways that can lead to new risks[43]. This problem has been tackled by the wide and vibrant fields of, amongst others, human-computer interaction (HCI), user experience (UX) and user interface (UI) design - which we hope the Commission is drawing from. One particular sub-problem that we wish to highlight is the tendency of AI systems to induce 'automation bias': errors made when decision makers rely on AI cues and systems instead of vigilantly seeking and processing information[44]. These issues may become more acute as the average performance of AI

---

[42] Brundage, Miles, Avin, Shahar, Wang, Jasmine, Belfield, Haydn & Krueger, Gretchen et al. (2020). Toward Trustworthy AI: Mechanisms for Supporting Verifiable Claims. arXiv:2004.07213.
[43] Rahwan, I., Cebrian, M., Obradovich, N. et al. (2019). Machine behaviour. Nature 568, 477–486.
[44] Mosier, K. L., Skitka, L. J., Burdick, M. D., & Heers, S. T. (1996). Automation Bias, Accountability, and Verification Behaviors. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 40(4), 204–208.
Parasuraman, R., & Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. Human Factors, 52(3), 381–410.

systems increase, leaving operators unprepared for unlikely, unexpected errors. Absent broader measures, human oversight can simply set up an operator to take the fall - to serve as 'moral crumple zone'[45] - once accidents happen.

The Commission should consider requirements to counteract automation bias. This could include prompting for human oversight on a random subset of outputs, requirements around the system's opacity, or limits on the system's autonomy (or speed) when deployed outside of intended environments.[46]

Recommendation: Further specify in which areas particular forms of human oversight will be required.

Human oversight will be necessary for trustworthy AI in many areas. However, the clarification that human oversight might take very different manifestations is very positive, to avoid a simplistic human-in-the-loop approach, which in some high-risk areas might even be dangerous (e.g., transportation, energy, etc.)

For the second manifestation listed, it will be important to specify whether the human intervention is required afterwards in all cases, or should merely be possible afterwards. There should also be a requirement for this intervention to occur within a reasonable time period. For example if a job applicant is appealing a decision taken by an AI system, the applicant could suffer if it takes too long for the human to complete the review of the decision.

General

Recommendation: 'Fauxtomation' should not allow companies to dodge the requirements.

'Fauxtomation', also known as 'clickwork' or 'ghost work', is when products or services that present themselves (or are perceived) as AI systems actually rely heavily on human labour, normally poorly paid and insecure[47]. A fauxtomation system should also meet some of the mandatory requirements. The requirements in terms of the information they must provide, their robustness and oversight by peers should not vanish merely because of a homunculus inside the system. Otherwise, this raises what we would call the 'puppet problem' - that companies may use human 'puppets' to circumvent some AI regulations. These considerations will become more and more important as hybrid and collective intelligences abound, such as through the use of human computation (e.g., Amazon mechanical turks)[48].

Recommendation: These requirements should also apply to suppliers of multi-purpose AI components.

A complication is raised by the (increasingly common) use by an AI application provider of multi-purpose AI components from a third-party. One could propose that the provider is solely responsible, and the only requirement for suppliers of multi-purpose AI components

[45] Elish, Madeleine Clare. (2019). Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. Engaging Science, Technology, and Society.
[46] Maas, Matthijs M. (2018). Regulating for 'Normal AI Accidents': Operational Lessons for the Responsible Governance of Artificial Intelligence Deployment. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society Pages 223–228.
[47] Taylor, Astra. (2018). The Automation Charade. *Logic Magazine*.
[48] Crawford, Kate & Joler, Vladan. (2018). Anatomy of an AI System. AI Now Institute.

should be to ensure that the terms and conditions of sale do not prevent meeting such obligations. However this is clearly insufficient, as it may be precisely these multi-purpose component that require testing and inspection. The direction of progress in the field is for AI systems and components to become more general and to become integrated and combined into larger systems. So suppliers should also be required to follow these requirements.

Conformity assessments

What is the best way to ensure that AI is trustworthy, secure and in respect of European values and rules?
Compliance of high-risk applications should be assessed ex-ante by means of an external conformity assessment procedure
Ex-post market surveillance after the AI-enabled high-risk product or service has been put on the market and, where needed, enforcement by relevant competent authorities
A combination of ex-ante compliance and ex-post enforcement mechanisms

To ensure that AI is trustworthy, compliance should mainly be assessed ex-ante by means of an external conformity assessment procedure by an independent testing centre. There should also be provision for additional ex-post market surveillance and enforcement, for example for software updates and systems that keep learning during operation.

We also strongly support the other proposals around governance in section H. The wider governance structure should exchange information and best practice, identify emerging trends, advise on standardisation activity, and issue guidance, opinions and expertise on implementation. This should include the meaningful participation of affected communities.

Recommendation: Keep the commitment to ex-ante, external conformity assessments by independent testing centre(s), supplemented with additional ex-post market surveillance and enforcement.

Self-assessment cannot ensure trustworthy AI. Firstly, AI development needs to be a level playing field - SMEs should not be disadvantaged by having to build a self-assessment system. Secondly, developers are often not experts in the risks their systems could cause. It is better to build up assessment expertise in dedicated, independent centres. Thirdly and unfortunately, there are many examples of a failure of self-assessment, including Volkswagen, Boeing, Google and Amazon. These companies shouldn't be "marking their own homework" for AI systems used in high-risk application areas.

The testing centres should focus on certification, but they should do this in a collaborative way, especially with SMEs. Before, during and after the conformity assessment they should be able to provide advice and suggestions. The assessment should not just be a simple pass/fail but a way to support companies, especially SMEs, towards developing trustworthy AI systems.

It is important that the testing centres have stringent cybersecurity and confidentiality standards, to allay any concerns about trade secrets or industrial espionage.

Recommendation: Do not consider a 'grandfather clause' for AI systems deployed before the regulation comes into place.

The conformity assessments will need to be retroactive. The Commission should not include a grandfather clause. If an AI system is being used in a high-risk application area, it needs to pass the requirements. It is irrelevant whether it was deployed before or after the regulation came into force. This is also needed to ensure technological neutrality.

Recommendation: Consider only allowing limited confidential testing and piloting in coordination with the independent testing centres.

While there should be some allowance for confidential testing and piloting of an AI application, this should be done in coordination with the independent testing centre. When used in a high-risk application area, an AI system could still cause unacceptable harm at a testing and piloting stage.