Comments from the Governance in AI Research Group (GAIRG) on the proposed EU AI Regulation

The Governance in AI Research Group (GAIRG) consists of:

- James Thornton LLB PhD, Lecturer in Law, Nottingham Law School, Nottingham Trent University, England, UK.
- Caroline Jones LLB PhD, Associate Professor in Law, Hillary Rodham Clinton School of Law, Swansea University, Wales, UK.
- Prof Jeremy Wyatt DM FRCP, Fellow of ACMI, IAHSI & FCI; Emeritus Professor of Digital Healthcare, University of Southampton, England, UK; Chair, AI Special Interest Group, Faculty of Clinical Informatics, London; Adviser, NHSX AI Lab; Past President, European Society for AI in Medicine.

The opinions expressed below are those of the individuals named above, not of our institutions.

General approach

We welcome the proportionate risk-based approach and attempts to support innovation while protecting safety. We also welcome the important observation that the legislation shall apply to users of AI systems originating from third countries outside the EU (para 10 p 20).

Scope and definition of AI:

While we welcome the proposed wide-ranging definition of AI in paragraph 6, sections 5.2.1, 53, this fails to capture an important subcategory of AI, the more traditional **knowledge-based systems** (KBS) based on expert knowledge or practice guidelines. These KBS are already widespread in medicine (especially in their simplified form as rule-based alerts and reminders) but can pose the same risks to health & safety as data-derived AI. Since knowledge-based systems can generate much more persuasive explanations than data-derived systems (eg. those based on deep learning) they may pose more risks, as users may assume that the system is correct because its explanation is plausible, and then follow incorrect advice - an example of automation bias. So, we would argue that knowledge-based systems should be included within the scope of the proposed legislation. This would be consistent with the requirements for a definition set out in para 6, page 18.

Including knowledge-based systems in the definition of AI has implications for other parts of the legislation. For example, paragraph 43 (page 29) states that: "Requirements should apply to high-risk AI systems as regards the quality of data sets used, …". This should be reworded to: "Requirements should apply to high-risk AI systems as regards the quality of data sets and knowledge base used, …"

Definition of high-risk AI

Paragraphs 33-37 pages 26 & 27 list a large number of AI systems considered to be high risk, but only one use case relevant to healthcare: "*Finally, AI systems used to dispatch or establish priority in the* *dispatching of emergency first response services should also be classified as high-risk since they make decisions in very critical situations for the life and health of persons and their property*."

We would argue that most (if not all) applications of AI in healthcare are inherently high risk. For example, see the table on the following page listing health and safety risks associated with different AI use cases in healthcare.

Application of AI in healthcare	Risk to health or safety of patients or populations
AI supporting the remote capture or processing of patient data embedded in devices or via online forms in apps / telemedicine / telehealth	a) Delays in or failure to capture or communicate important data or events to health professionals that would otherwise lead to effective action b) False reliance by health professionals on data from remote monitoring tools
Online or app-based symptom checkers used by the public or health professionals	False reassurance due to the failure to recognise a dangerous but treatable condition
Diagnostic decision support systems	False reassurance due to the failure to recognise a dangerous, treatable condition
AI supporting professionals with the choice or interpretation of lab or imaging investigations	a) Incorrectly ordering a risky / invasive test (eg. lumbar puncture) when not needed b) Failure to carry out a diagnostic test that is needed c) Incorrect interpretation of test / imaging results, leading to a missed therapy opportunity
AI supporting patients or professionals with the choice, dosing or discontinuation of therapy	a) Choosing an ineffective or toxic therapy b) Choosing the wrong dose of a therapy with a narrow range between benefits and side effects (eg. insulin dosage adjustment) c) Premature discontinuation of a drug risking relapse or antibiotic resistance.
AI giving an outcome risk estimate to patients or health professionals	a) Giving a risk that is too high: may result in invasive tests or toxic treatment that are not needed; or in agreement to discontinue all therapy and end life when treatment appears futile, but is not. b) Giving a risk that is too low: may result in failure to change health-related risk behaviour, give necessary treatment; failure to counsel family about genetic risks, etc.

Proposed conformity assessment procedures related to high risk AI

a) If the AI supplied is part of a device, eg. a medical device:

- We welcome the proposal for high risk AI systems related to products already covered by existing Union harmonisation legislation that the assessment of compliance with this Regulation will be addressed under the existing conformity procedures provided for by that legislation; and agree that this pragmatic approach should minimise burdens on providers and the potential for duplication.

- The proposal is that certification will rely on a third party ex ante conformity assessment – ie. predicted conformity rather than actual ("ex post") conformity. However, given that this AI will be used in a high-risk scenario with the potential for causing harm to people or organisations, is ex ante conformity assessment enough ? We would argue that it is impossible to predict the actual performance of an AI algorithm from the quality of the data used to train the algorithm, which is why external validation using a new, unseen dataset is regarded by the NHSX AI Lab as a core requirement for algorithms to be adopted by the NHS. There are many examples of AI algorithms that performed extremely well in house but failed to deliver on their potential when used elsewhere. There are even examples in which the algorithm was validated on datasets which were later shown to include information about the reference standard (such as images that included the ruler used by dermatologists to measure the size of a melanoma). So, to protect public safety for high risk algorithms that pass the conformity assessment process, an independent assessment of **actual performance** is needed, using a new external validation dataset. We do not feel that the requirements described in paragraph 46 page 30 ("Such information should include the general characteristics, capabilities and limitations of the system, algorithms, data, training, testing and validation processes used as well as documentation on the relevant risk management system ") gives enough detail on methods to promote supplier adherence to good practice. Good practice methods for use by a trustworthy AI provider are described in the EQUATOR group's TRIPOD reporting guideline and supporting materials https://www.equator-network.org/reporting-guidelines/tripod-statement/
- It is reassuring that this check will be carried out by a third party, ie. a Notified Body. However, are there enough sufficiently skilled people working in European Notified Bodies to carry out this assessment ? How will the EU ensure that enough Europeans are trained in future to carry out this assessment task ?

b) For standalone high-risk AI:

- We are pleased to note that conformity assessment procedures in Articles 19, 43, 48 and 49 required for high-risk AI systems prior to being put into service or placed on the market.
- However, we have concerns regarding the potential for some high-risk AI systems (ie. those listed in points 2-8 of Annex 3) to be assessed by the provider for compliance with the requirements in Chapter 2 of the Title via an internal control procedure (outlined in Annex VI). A systematic review of empirical research on computerised clinical decision support systems has shown that when developers themselves carry out studies on the performance of their products that they are three times as likely to indicate positive results than when an independent assessor did so (Garg AX, Adhikari NKJ, McDonald H, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA 2005;293:1223–38). This systematic review provides strong evidence that third party assessment of the quality, performance or impact of high-risk AI will be more rigorous than first party assessment. We believe there is distinct risk that similar findings, with regard to compliance with the Regulation, could eventuate if the compliance procedure is left solely to internal control for these particular high-risk AI systems.
- There is a risk of providers believing that they were impartial when checking AI quality but may not be. A (non AI) example of the naivety of relying on self-certification is how VW managers persuaded themselves that it was acceptable to develop software to defeat diesel vehicle

emissions checks to meet an internal corporate goal, when it led to massive fines for VW and hundreds of thousands of Europeans being affected by respiratory conditions related to illegal nitrous oxide emissions.

For further discussion on trustworthiness and developers, see our discussion in: Jones C, et al. BMJ Health Care Inform 2021;28:e100247. doi:10.1136/bmjhci-2020-100247 https://informatics.bmj.com/content/bmjhci/28/1/e100247.full.pdf

Cost implications of the legislation

We are surprised at the low estimated costs of verification for the providers of high-risk AI systems (3.3: "Verification costs could amount to another EUR € 3000 to EUR € 7500 for suppliers of high-risk AI."). Even a simple study to check the accuracy of output / advice requires a test dataset of hundreds to thousands of cases similar to those in which the AI will be used, each of which requires a robust reference or gold standard. Assembling the gold standard for each of these test cases can take an hour of work, so even at EUR € 25 per hour, a sizable test set could cost 5000 X 25 = EUR € 125000. To ensure rigorous evaluation, this test set can only be used once to validate a given AI.

Transparency & labelling

We strongly agree with Articles 13 & 52, as expanded on in Section 5.2.4, that "When persons interact with an AI system or their emotions or characteristics are recognised through automated means, people must be informed of that circumstance." and believe that this will help promote the uptake of AI, and also its quality.

However, there is a risk of proliferation in the methods used by suppliers to inform the user that AI is present. We would therefore value an EU common, normative standard on how the presence of AI in a complex app or other software – or a medical device – is communicated to the user.

The EU has previously used its welcome power to regulate on the use of standard trust marks such as the CE mark, and standard labels such as that used on tyres showing the purchaser the impact of a tyre on fuel consumption, noise and braking effectiveness. We believe that this same approach could be used to label AI (and we have recently recommend the tyre label approach for AI in the healthcare context - Jones C, et al. BMJ Health Care Inform 2021;28:e100247. doi:10.1136/bmjhci-2020-100247). There is an opportunity to both develop a standard mark to denote that there is "AI Inside", and also a summary label informing the purchaser / user about key aspects of an AI. Such aspects could include:

1. The source or nationality of the datasets or knowledgebase used to develop the AI
2. The date of the last revision to the AI algorithm (possibly including a version number and a "Do not use after" date)
3. The intended use of the algorithm / AI

4. The intended user of the algorithm / AI (eg. member of the public, trained person or professional user)
5. The original publisher or developer and their contact details for enquiries or to whom comments or complaints should be addressed
6. Depending on context, other key details such as the actual performance of the AI on an independent test set, in the case of a diagnostic decision support system in medicine (eg. to support the interpretation of mammographic images)

We welcome the statements in paragraph 47 page 22: "Users should be able to interpret the system output and use it appropriately. High-risk AI systems should therefore be accompanied by relevant documentation and instructions of use and include concise and clear information, including in relation to possible risks to fundamental rights and discrimination, where appropriate". However, we again believe that users making high risk decisions such as in healthcare need **immediate and permanent access** to the 6 items of information listed above on which to base their decision to trust the AI output or not. We propose that this core information should be included in a standard label attached to each high risk AI system to ensure that users can immediately access the necessary data they require in standard format, no matter where in the EU they are working.

We agree that "The level of accuracy and accuracy metrics should be communicated to the users." (articles 13 & 15 & para 49, page 30), but think that this information should be available **in standard format** (like the excellent EU tyre label) at the point of use, not buried on the supplier website nor in instructional materials.

We welcome the robust transparency requirements in article 13 and (for human oversight) article 14, point 4. However, we are concerned that the issue of some AI systems being a so-called "black box" (such as those using deep learning) is not adequately addressed by the Regulation. By "black box", we mean that the reason or mechanism for the AI system coming to a particular conclusion/output is unclear, even if the final conclusion/recommendation itself *is* clear. This has implications for user trust, particularly in healthcare. Eg. a survey of senior physicians ranked "black box" features as their 2nd greatest concern about professional practice, ethics and liability in using clinical decision support systems (Petkus H, et al. *Clinical Medicine* 2020;20:324-8). The requirements in articles 13 and 14 currently focus on transparency of *output* (eg. article 13, point 1), but this may not go far enough. We would encourage consideration and provisions **directly addressing** transparency requirements as to the underlying mechanism(s) by which the AI system comes to conclusions, not just the conclusions themselves.

Self-learning systems and substantial modification:

We are concerned about Article 3 – point 23 read in conjunction with the statement in para 66, page 33: "*…as regards AI systems which continue to 'learn' after being placed on the market or* *put into service (i.e. they automatically adapt how functions are carried out), it is necessary to provide rules establishing that changes to the algorithm and its performance that have been pre-determined by the provider and assessed at the moment of the conformity assessment should not constitute a substantial modification*."

Our concern originates in the fact that the performance of learning systems after release on the market is inherently unknown, so clearly needs to be checked at intervals with the interval between the checks reducing as the consequence(s) of an erroneous output increases. There will also be some high-risk use cases (eg. closed loop control systems in nuclear power stations or a closed loop insulin infusion pump) in which systems that learn (ie. change after their release on the market) should be prohibited.

Language used

We are concerned about the extensive use in this proposed legislation of the outmoded Latin phrases "ex ante" and "ex post". These are confusing at least, and could be misleading to some stakeholders. The phrases derive from economic history a century ago which is not related to AI, and were actually rejected by world leading economists such as John Maynard Keynes. A clearer wording would be to substitute "predicted" for ex ante and "actual" for ex post (eg. "predicted conformity" and "actual performance"). Use of modern language would make the EU's proposed approach to high-risk AI much clearer and more accessible to non-economist stakeholders.