Expert opinion

Lilian Edwards

Professor of Law, Innovation and Society,
Newcastle University

Regulating AI in Europe: four problems and four solutions

Ada Lovelace Institute

March 2022

Introduction

The subject of the paper is the European Commission proposal for the Artificial Intelligence Act ('the AI Act'), published on the 21 April 2021 and the draft Council position also since published. [1]

The aim of this paper, is to help create 'trustworthy AI', which balances proportionately the social interest in innovation and better delivery of public services from AI, with adverse impacts on fundamental rights and societal values. This aim aligns with that of the proposed AI Act, which we welcome in principle as the first comprehensive attempt in the world to regulate AI from a fundamental rights perspective.

We have closely followed from their inception the European Union's plans to regulate AI. The Act sets out harmonised rules for the development, placing on the market, and use of 'AI systems' in the European Union (EU). However its area of impact is wider than the EU, since it governs any provider who places AI systems on the market or into service in the EU. The AI Act, like the GDPR before it, is explicitly positioned to become a global model and given first-mover advantage, it is quite likely this will become the case.

The Ada Lovelace Institute's particular interest, as a UK-based independent research institute working in data, policy and regulation, is to consider how to build on, develop – or perhaps reject – this model, before it becomes entrenched.

Major structural change is politically unlikely within the AI Act legislative process at this stage. A great deal of effort has already been sunk into it by the Commission and Council, and shortly the Parliament, which will make fundamental changes in structure or goal implausible. While fundamental changes to the AI Act – such as the addition of a true *ex ante* fundamental rights impact assessment, discussed in detail below – may be regarded at this point as unrealistic, this is only the start of

1 Council of the European Union (2021). *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts – Progress report*. Available at: https://data.consilium.europa.eu/doc/document/ST-13802-2021-REV-1/en/pdf

regulating AI, both in the EU and globally. We feel it is important to have an eye to the horizon, as well as the ground.

Another limitation to acknowledge is that the Act is trammelled by the requirements of EU constitutional and internal market law, and understandably tends to make use of already existing infrastructure paradigms, especially those provided by the New Legislative Framework (NLF), such as Market Surveillance Authorities (MSAs).

In short, the AI Act is itself an excellent starting point for a holistic approach to AI regulation. However, there is no reason why the rest of the globe should unquestioningly follow an ambitious, yet flawed, regime held in place by the twin constraints of the New Legislative Framework (see below) and the legislative basis of EU law. Therefore, in this paper, it seems important to flag the debates both EU policymakers and the world beyond the EU should be having at this crucial regulatory turning point.

This paper is – therefore – primarily a critique of the existing AI Act, which we hope is of relevance in the EU legislative process, informed especially by Ada's practical and theoretical track record in impact assessments, data stewardship, public participatory methods and regulatory policy. It is also a survey of its flaws as a potential global model for 'getting AI right'.

There are several things this paper does *not* do. First, there is a serious debate to be had about whether a holistic instrument for regulating AI systems as opposed to governance of sectors such as labour, health or military applications (which might then also include codes of conduct, ethical principles and technical standards, as well as command-and-control law) is the right way to go, and if so, what its full scope should be. Second, there are similar and more fundamental debates to be had about whether 'AI' actually exists or is merely a term for advanced software or data engineering; and if it does exist, again, what is its useful scope?

Thirdly, there are several jurisdictions and international organisations that are beginning also to regulate AI, and which contribute to the discourse on a global model for AI regulation. These – very non-exhaustively – include the Council of Europe's new proposals on AI and human rights, currently being finalised by the CAHAI and aiming to be

open for ratification in 2024;[2] the Canadian Directive on Automated Decision-Making, which imposes a requirement for a questionnaire-based algorithmic impact assessment in certain cases;[3] and the exciting new Chinese law, which regulates recommender algorithms and contains in many ways a distinctly non-Western and socialist perspective.[4] Furthermore, a very large number of non-state bodies have also developed governance tools for AI, including principles, codes of ethics and, notably, models for algorithmic impact assessment.[5, 6]

These issues and models will be addressed in the next phase of research, which convenes international experts to discuss:

1. Alternative global models for governance of AI
2. AI and labour
3. Emotion ID (biometrics) and general-purpose AI
4. Technical mandates and standards (in conjunction with the Alan Turing Institute)

In this paper however, we primarily seek to point out key flaws with the AI Act as a whole and where possible, suggest solutions drawn from the experience of an independent organisation with a mission to make data and AI work for people and society. As many reading this paper will already be familiar with the AI Act proposal, it has not been reprised in detail below. We will shortly be issuing a short summary/explainer of the AI Act proposal so far.

---

2 MAIA Grenoble Alpes. (2021). 'The Council of Europe's recommendation for a legal framework on AI'. *AI-Regulation.com*. Available at: https://ai-regulation.com/council-of-europe-cahai-ai-recommendation/

3 Government of Canada. (2020). *Algorithmic impact assessment tool*. Available at: https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

4 Creemers, R., Webster, G., and Toner, H. (2022). 'Translation: Internet Information Service Algorithmic Recommendation Management Provisions – Effective March 1, 2022'. *DigiChina*. Available at: https://digichina.stanford.edu/work/translation-internet-information-service-algorithmic-recommendation-management-provisions-effective-march-1-2022/

5 There is of course a further debate about whether AI needs regulation at all or should be entirely governed by the market. We discard this debate as having already concluded in favour of regulation of some kind.

6 A leading model here is that developed by Data and Society, a US NGO : https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/ . In the UK, as a sectoral example, the IFOW is working to promote a statutory algorithmic impact assessment for public sector automated decision-making. See: Hansard. (2021). Public Authority Algorithm Bill [HL]. [Hansard]. (Vol. 816). https://hansard.parliament.uk/lords/2021-11-29/debates/E07A5CBD-A767-4D35-9261-37B35BA086BB/PublicAuthorityAlgorithmBill(HL)

Critiquing the EU AI Act

Our starting point is that the Proposal takes an unhelpfully oversimplified view of the way in which AI systems function. This problem derives from the origins of the Act's structure in the New Legislative Framework (NLF) adopted in 2008, which sought to ensure the safety of consumer-facing products entering and circulating in the internal market. While this scheme has worked relatively well for tangible products, the division of duties seems much more questionable in a world of (a) AI as a service which learns and changes, (c) 'AI as a service' or 'upstream' AI services,[7] (c) general purpose AI and (d) AI as part of the services of a platform (the 'AI lifecycle').

These issues are unpacked further below as a series of issues:

1. **AI is not a product** nor a 'one-off' service, but a system delivered dynamically through multiple hands ('the AI lifecycle') in different contexts with different impacts on various individuals and groups. This derives from various features of the current AI market discussed in detail below.

2. **Those impacted by AI systems** – sometimes thought of as end-users, data subjects or consumers – **have no rights, and almost no role in the AI Act**. This is incompatible with an instrument whose function is to safeguard fundamental rights.

3. The alleged **'risk-based'** nature of the Act is illusory and arbitrary. A genuine assessment of risk based on reviewable criteria is necessary.

4. The Act **lacks a general fundamental rights risk assessment**, for *all* AI systems in scope of the Act, not just 'high-risk' AI.

7 We discuss this further below. Broadly 'upstream' AI services involve 'Artificial Intelligence as a Service', which in accordance with dominant industry use of the term, refers to pre-trained models provided to customers on a commercial basis (see Cobbe, J. and Singh, J., 2021. Artificial Intelligence as a Service: Legal Responsibilities, Liabilities, and Policy Challenges. *Computer Law and Security Review*, v. 42).

1. AI is not a product nor a 'one-off' service, but a system delivered dynamically through multiple hands ('the AI lifecycle') in different contexts with different impacts on various individuals and groups.

The Act draws its inspiration from existing product safety legislation, and largely conceives of AI 'providers' as the equivalent of the manufacturers of real-world products like dishwashers or toys. For these kinds of products, it is indubitably the initial manufacturer who is the person who knows best how to make the product safe. Thus, most duties are placed on these 'manufacturers', at the very beginning of the AI lifecycle.

But AI is not a dishwasher and the way downstream deployers use it and adapt it, may be as significant as how it is originally built. The AI Act takes some notice of this but not nearly enough, and therefore fails to appropriately regulate the many actors who get involved in various ways 'downstream' in the AI supply chain. This manifests in a number of different ways:

- Many AI products are **dynamic, not static** products – their behaviour (and successful implementation) will change with new data, new uses, and new integrations, which in turn changes their risk profiles and requires continuous evaluation.

- Many AI products are **not produced by a single organisation**, but involve a complex web of procurement, outsourcing, re-use of data from a variety of sources, etc. This changes the question of who is in scope, and who should be accountable, for different parts of the AI lifecycle. Notably, smaller 'downstream' providers are likely to save time, resources and maintenance obligations by relying heavily on AI services delivered by the large tech firms such as Google, Microsoft and Amazon. This follows the same path as cloud computing has already trodden, but – as we shall see – creates substantial issues for regulation of AI through its lifecycle.

- AI systems can be **general purpose**, meaning the same system can be applied to different contexts and raise different impacts for different individuals and groups. For example, a developer of a facial recognition system could sell their product to authenticate entry to prisons or to surveil customers for targeted advertising. Holistically evaluating the risk of such a system in the abstract is an impossibility. Again, general-purpose AI is predominantly delivered by the tech giants with dominant market share.

- An AI system is not necessarily a creature in their own right, but can be a function of a **larger system or platform**. Facebook is a social media platform, but it runs a wide variety of AI systems on its services at any given time, including content moderation algorithms, recommendation engines, advertising algorithms, search functions and others, some of which may belong to Facebook, and some of which may be affiliated to and run by a third party.

Translating this complex web of actors, data, models and services into a legal regime that places duties and rights on certain identifiable actors is extremely hard. In the AI Act, primary responsibility is, by analogy to the manufacturers of physical goods, placed on an initial **'provider'**. Those who place AI systems provided by others into operation are confusingly termed **'users'** (we suggest, alongside others, renaming as 'deployers') and have a highly limited, regulated role in the AI Act, which comes into play principally when a 'substantial modification' is made to the upstream system.[8]

Yet many obligations in the AI Act scheme, such as ensuring that 'human oversight' is correctly implemented (in high-risk systems) can only effectively be put in place by users (deployers) who, often, will buy a system off the shelf and will not regard themselves as making the 'substantial modification' necessary to become regarded legally as providers (Article 3(23)). The Act fails to take on the work, which is admittedly difficult, of determining what the distribution of sole and joint responsibility should be contextually throughout the AI lifecycle, to protect the fundamental rights of end users most practically and completely. It can be compared unfavourably to recent developments in GDPR case law,[9] where courts are attempting to *distribute* responsibility for data protection among various controllers at the most relevant times.

8 Art 3(23) of the AI Act : 'substantial modification' means a change to the AI system following its placing on the market or putting into service which affects the compliance of the AI system with the requirements set out in Title III, Chapter 2 (essential requirements for high-risk AI) or results in a modification to the intended purpose for which the AI system has been assessed.

9 See C-210/16 *Wirtschaftsakademie Schleswig-Holstein* Judgment of the Court (Grand Chamber) of 5 June 2018 (Available at: https://curia.europa.eu/juris/liste.jsf?num=C-210%20/16) and subsequent CJEU case law.

Example 1: chains of providers

To give a detailed example, an algorithm that enables the training of a model in a novel and efficient way might be made freely available online by an academic researcher. It might then be adopted by a start-up delivering 'machine learning-as-a-service' for free. This trained model might then be incorporated, for a fee, by a commercial cloud provider offering software-as-a-service (SaaS). This SaaS might then be purchased by a government department to deliver a public-facing service. Datasets for both training and testing at various stages might be retrieved from various global providers, with varying degrees of access to how those datasets were constructed. Such a system might be characterised as 'high risk' (Annex III, paragraph 5) *only* at the point where it is put into service by the public body. Yet the system would be the product of many hands, not all in ongoing contractual relationships, and it would not be clear which had to, or should have had to, fulfil duties to certify the system as compatible with 'essential requirements' (Chapter II).

Furthermore, the characterisation as 'high risk' in our example might only cut in at the last step, yet that user/deployer might well not have access to either model or training data, or have technical resources, or legal rights under license, to assess and alter the system. Meanwhile, the initial provider could currently claim that at the time of provision there was no intended 'high-risk' use. A successful regime to regulate AI increasingly made from components supplied through chains of providers must grapple with this problem.

General purpose AI

The Act, in its failure to appropriately regulate the many actors who get involved in various ways in the AI lifecycle, will particularly struggle to regulate **general-purpose AI systems** appropriately. 'General purpose' means very loosely that the same system can be applied to different contexts and raise different impacts for different individuals and groups. A clearer definition is very much needed however.

Example 2: 'general purpose' AI

For example, a developer of a facial recognition system could sell their product to authenticate entry to prisons or to surveil customers for targeted advertising. Holistically evaluating the risk of a system in the abstract is an impossibility. We have seen in recent Uber disputes in the UK,[10] how facial recognition systems used to verify identity can, in context of deployment, discriminate against workers of colour who make-up the majority of the Uber workforce. If we want to make these systems operate fairly and in a non-discriminatory fashion, we must be as careful about how they are deployed and embedded in existing processes downstream, as how they are built upstream.

Example 3: large language models

One key case study for general purpose AI is large language models such as Open AI's GPT3. These systems or services are often incorporated downstream into multiple AI systems for multiple purposes, in contexts not supervised or imagined by the upstream providers. These include mundane 'plug-ins' such as for analytics and language translation as well as services such as large language models, which may allow the automated generation of text, translation of speech to text, automated bot assistants etc. Large language models, while extremely useful for, among other things, speech synthesis, generation and translation, are known to be dangerous sources of errors, discrimination, and other adverse effects.[11] However, such systems may largely fall out of the controls of the AI Act because their uses, and thus their impacts, are determined not by the initial provider but by downstream deployers.

We are pleased to note in the draft Council position (Article 52a) that it is clarified that any person (a deployer, in effect) who 'puts into service or uses' a general-purpose AI system for an intended high-risk purpose comes under duties to certify confirmation with the essential requirements of Chapter III and does not seem to need a 'substantial

10 See Butler, S. (2021). 'Uber facing new UK driver claims of racial discrimination'. *The Guardian*. Available at: https://www.theguardian.com/technology/2021/oct/06/uber-facing-new-uk-driver-claims-of-racial-discrimination

11 See Bender et al 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?'. *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, March 2021*, pp. 610–623. Available at: https://doi.org/10.1145/3442188.3445922

modification'. This is clearly aimed at catching the downstream adapter or deployer.

But in so doing, the text seems to have done nothing to meet the problem that the user/deployer almost certainly lacks mandatory access to training-set or testing data, or ability to compel changes in the upstream service (unless these are built in as rights into a contract which is highly unlikely, especially when there are chains of providers as in example 1). At the same time, the Council proposal removes liability for the upstream provider of the general-purpose AI (Article 52a (1)). This exculpates the large tech suppliers like Amazon, Google and Microsoft, whose involvement in certification of AI as safe is, as discussed above, vital, since they have effective control over the technical infrastructure, training data and models, as well as the resources and power to modify and test them.

2. Users in the conventional sense of 'end-users', have no rights and almost no role in the AI Act scheme at all.

The contrast here to the European data protection regime, also intended to protect fundamental rights, is palpable. By deriving the design of the AI Act primarily from product safety and not from other instruments, the role of end users of AI systems as subjects of rights, not just as objects impacted, has been obscured and their human dignity neglected. This is incompatible with an instrument whose function is ostensibly to safeguard fundamental rights. The current proposal fails at many key points of the regulation and enforcement cycle. It does not consult users at the very start when providers of 'high risk' AI have to certify that they meet various fundamental rights requirements, even though the users will suffer potential impacts; does not give users a chance to make points when unelected industry-dominated technical bodies turn democratically made rules into the standards that actually tell companies making AI how to build it; and, most importantly, does not allow users to challenge or complain about AI systems down the line when they do go wrong and infringe their rights. The GDPR has already shown, in areas like targeted advertising and lack of protection of data transfers out of the EU, that users as activists and complainants are as crucial to post-launch enforcement as regulators (and the AI Act has far weaker enforcement structures than the GDPR). Why has their voice been cut out of the AI Act?

Users must be given a chance to have their views considered both *before* the product is certified as valid to enter the market, as well as rights to challenge the legality of a system *after* it is placed on the market. Civil society as the representatives of users, must be empowered and resourced to enter the standard-setting process on their behalf.

We want to start a debate about whether regulators like existing Data Protection Authorities, already struggling to cope with policing the GDPR, can really manage also to represent the voice of the AI user, or whether we should look to building in extra capacity for a central European body to become a champion for users and a central source of expertise about what users throughout the EU really need in order to trust AI.

We would draw here on the experience of consumer law and the use in many countries of an ombudsman-like figure, who could not only receive and further user complaints but, on an EU-wide basis, group them, spot patterns of complaint, and possibly instruct or aid regulators or civil society in taking representative actions (which are also currently not part of the framework of the Act).[12] This could also assist in reducing the impact of a state Market Surveillance Authority (MSA), which acts a single point of permission to circulate AI systems in the EU single market, but for whatever reason was unable to fulfil its regulatory role properly. We also discuss below at point 5 how users could become directly involved in the initial impact assessment of an AI system.

3. The alleged 'risk-based' nature of the Act is illusory and arbitrary. Impacts on groups and on society as a whole need to be considered, as well as risks to individuals and their rights, and risks should be considered throughout the AI lifecycle not just at market entry.

The AI Act does not lay down criteria for when AI poses unacceptable risks to society and individuals. It merely designates set lists of what categories of AI systems are deemed 'unacceptable risk' and thus banned from the EU (a small number of systems, notably including some public-space, real-time, law-enforcement biometric systems); and

12 The Ombudsman model was developed for users of public sector services to make complaints but has spread successfully to the private sector in consumer law, especially in digital spheres: see chapter 4 in Hertogh, M. and Kirkham, R. (eds.), (2018). *Research Handbook on the Ombudsman (Research Handbooks in Law and Politics)*. Edward Elgar.

which should be allowed on to market only if certain safeguards are put in place ('essential requirements'), known as 'high-risk' AI. A further few systems are even more arbitrarily designated as limited risk, although the obligations associated with this (basically, transparency in labelling as machine-made) are minimal and to some extent duplicate existing requirements in the GDPR).

These lists are not justified by externally reviewable criteria, and thus can only be regarded as political compromises at one point in time – leaving it difficult-to-impossible to challenge the legal validity of AI systems in principle rather on point of detail. The draft text added by the Council as of 30 November 2021 illustrates this point well: insurance systems and digital critical infrastructure have been added, but with little precision as to what the latter means, and little or no justification as to why these were selected, and not, say, 'emotion ID recognition' systems which many regard as pernicious and unscientific.[13] In practical terms, if it is uncertain why certain systems are on the red or 'high-risk' lists *now*, it will be difficult-to-impossible to argue that new systems should be added *in future* according to the criteria in Article 7. We regard this as unacceptably arbitrary, denying justiciability and lacking futureproofing.

We suggest therefore that *initial* criteria for assessing what is high risk be developed, possibly mirroring the criteria for adding *new* systems to Annex III under existing categories in Article 7. Examining if systems meet or escape the scope of these criteria should be an essential precursor to, or part of, the impact assessment process discussed below.

Without legitimacy as to why certain systems are or are not on the red list, both public trust and the rule of law are inherently compromised.

13 See Heaven, D. (2020). 'Why faces don't always tell the truth about feelings'. *Nature.com*. Available at: https://www.nature.com/articles/d41586-020-00507-5. These systems typically take images of faces or data drawn from other bodily functions, e.g. sweat or body temperature, and interpret them using algorithmic models as indicating certain behavioural states, such as attentiveness or truth telling, or character features, such as reckless, or identity, such as gay.

Clearly, problems may arise from this suggestion. Given the internal market basis for the AI Act, it seems possible to argue that the 'red' list and 'high-risk' AI lists may have to be defined *ex ante* to fulfil the ostensible purpose of the Act to harmonise the placing of AI systems on to the single market. The Commission have furthermore argued it is a strength of the proposal, not a weakness, that providers are *not* asked to self-assess if their system is 'high risk' but only to check if they fall into one of the categories on the list. This they argue, creates certainty. Obviously, a criteria-based risk assessment process would run a risk of being gamed if based on self-assessment, or more charitably, of providers being over-optimistic or under-critical. Adding third-party audit of some kind to the risk-assessment process would help exclude this risk but would add cost and take up time. We address and to some extent repel this in point 4 below.

4. The Act also lacks a comprehensive process for rights-based assessment of the impacts and risks of an AI system

The AI Act is not ambitious enough at assessing and seeing off the risks caused by AI. The Act speaks continually of risks to fundamental rights as its prime reason for being (80 mentions in the initial proposal) yet contains no comprehensive *ex ante* fundamental rights-based impact assessment for all AI systems. Unpacking this statement, we pose two basic questions:

1. What criteria should we use to certify the safety of AI systems in society? Is certifying conformity with fundamental rights of the type protected by the EU Charter and the European Convention on Human Rights sufficient?

2. If we can agree on these criteria, should they be certified before the system is released on to the market, or into society ('*ex ante*' assessment) or after they have been put out and had impact ('*post-factum*' assessment or audit); or some combination of both?

Certifying the safety of AI systems: fundamental rights and beyond

The nearest the Act has to an *ex ante* assessment of compliance with fundamental rights is the need for certification for 'essential requirements' in Chapter III. However, this only applies to 'high-risk' AI (see above), which at present does not include many or most of the AI

systems consumers encounter on a daily basis, such as search engines, content moderation and profiling for targeted interventions. In this sense the AI Act is in fact a step backwards from the GDPR, where *all* machine-learning systems processing personal data are already required to carry out a data protection impact assessment (DPIA).[14]

Even where a system is subject to Chapter III requirements, they do not constitute a true fundamental Human Rights Impact Assessment (HRIA). Only three of the Chapter III articles refer to fundamental rights and, in most cases, only briefly. This is far from the kind of HRIA that is already required or recommended for specific classes of private and/or public sector activities in a number of states (e.g. Denmark, Scotland) or is being developed specifically for AI (algorithms) by the Council of Europe, who sponsor the European Convention on Human Rights (ECHR).[15] Many key interests which may be impacted by AI such as freedom of thought and conscience,[16] and due process[17] are not included at all.

Another problem with Chapter III is the lack of systematic concern for impacts on groups, particularly algorithmically constituted groups. Chapter III occasionally refers to risks to groups (e.g. Article 10(3)), but on the whole, its concentration is on individuals not society. Much scholarship in the human rights domain has argued that concentrating only on individual rights – as in the conventional ECHR human rights structure – leaves crucial gaps in relation to common and minority interests, and allows structural discrimination to persist and grow. Individual rights tend to empower those who are already most empowered to exercise their rights and fail to support marginalised and socio-economically impacted communities. The instrument that is most often cited to give rights to individuals in the AI society, at least in the EU, is data protection law, and a critique has built up which points out that it has a gaping gap around rights for groups and society as a

14 It may have been presumed that all high-risk AI systems will also be required to undergo a DPIA – but this is by no means certain, especially given likely assertions (probably wrong, but proof may be tiresome) that systems process only anonymised data, or possibly controversial 'synthetic' data.

15 See a useful summary in IFOW. (2021). *Policy Briefing – Building a systematic framework of accountability for algorithmic decision making.* Available at https://ifow.webflow.io/publications/policy-briefing-building-a-systematic-framework-of-accountability-for-algorithmic-decision-making .

16 For freedom of thought, conscience and religion, see art 9 of the European Convention on Human Rights. Available at: https://www.echr.coe.int/Documents/Guide_Art_9_ENG.pdf . It could be argued that the prohibition on AI systems in art 5 which subliminally manipulate users is based on freedom of thought; however this is not explicit and is not extended as a general principle throughout the Act, nor even throughout the Ch III essential criteria.

17 For due process and the rule of law concepts, see art 6 of the European Convention on Human Rights. Available at: https://www.echr.coe.int/documents/guide_art_6_criminal_eng.pdf

whole. Even more importantly, while class actions may help groups to get remedies based on individual rights, algorithmic systems construct new groups whose commonalities are not easily fitted into existing rules for discrimination and protected characteristics.[18] It would thus be unfortunate to see AI regulation mostly proceed down the same traditionalist individualised path.

Arguably an *ex ante* impact assessment and/or *post-factum* audit should, not only, more comprehensively take account of fundamental rights, but also move beyond fundamental rights, to scrutinise other important risks and impacts.[19] Ethical impact assessment work has already extensively explored these possibilities but not generally in the context of legal mandates. These include risks to groups and communities; individual and structural discrimination caused by contexts of deployment; environmental impacts; effectiveness; transparency; contestability; and the views and wishes of end users and affected communities. Most or many of these issues are not contained in the Chapter III conformity exercise. Some of these issues are already, or should be, raised as part of a DPIA, but as noted above, not all AI systems may require a DPIA; participation by affected users is desirable but not mandated; publication is not required; subsequent re-examination after a certain period is not clearly mandated.[20] We do not feel a DPIA is an adequate justification for the gaps in the 'essential requirements' scheme of the AI Act.

Participation

A key point where the Ada Lovelace Institute has already conducted considerable prior research concerns user participation in impact assessment, the lack of which is already highlighted above as a fatal weakness of the Act at point 2. Research shows that AI development teams tend to be non-diverse and particularly rarely include representation from the marginalised groups most impacted by biased or unfair systems. Giving access to individuals affected by AI systems

18 See Mantelero A. (2016), 'Personal data for decisional purposes in the age of analytics: From an individual to a collective dimension of data protection' *Computer Law & Security Review*, Volume 32, Issue 2. pp. 238-255. Available at: https://doi.org/10.1016/j.clsr.2016.01.014 Even outside the data protection sphere, US scholars have recognised that a severe weakness of privacy law is its failure to support privacy as a social good as well as an individual remedy: see Regan, P. (1995), *Legislating Privacy*. UNC Press.

19 Ada Lovelace Institute. (2021). *Technical methods for regulatory inspection of algorithmic systems.* Available at: https://www.adalovelaceinstitute.org/report/technical-methods-regulatory-inspection/

20 See GDPR art 35 (9)(11).

to point out flaws at the design, pre-market stage of AI development is crucial. Conventionally this has been done by civil society intervention. But perhaps in a time of AI regulation we should consider adding more direct ways for users to intervene, in addition to properly financing and giving access to civil society.

Public scrutiny would be assisted by mandatory publication of completed impact assessments. Publication would also assist in representative group challenges and be available as a resource to other actors down the AI supply chain, to minimise the burden of subsequent impact assessments. Social media also offers a potential route for users to get directly involved as opposed to via the mediation of civil society, who are crippled by resource constraints. Such approaches could of course be vulnerable to trolling, abuse and co-option and would have to be very carefully piloted and policed. Representation would be haphazard as we see on social media where some voices, often coming from privilege, drown out others. These approaches might expand some types of participation; but could only ever be additional to more traditional routes as some communities will not be well represented on various social media platforms for a number of obvious socio-economic reasons.

Alternately then, perhaps the EU AI Board should maintain a standing panel of representative users – a type of 'citizens jury' – who could be mandated to comment at impact assessment stage. Ada Lovelace has already set out a framework for participatory modes of data stewardship.[21] These methods could be applied to the governance and design of AI algorithms, in addition to the data that underpins them. Ada has developed a participatory process for the use of AIAs in a particular context, and the EU process, or the European Commission, could explore how a similar process could be adapted for their needs here. We call for imaginative thinking here as to how to get those most affected by the deployment of AI systems, traditionally regarded as inert consumers, to become truly involved.

21 Ada Lovelace Institute. (2021). *Participatory Data Stewardship*. Available at: https://www.adalovelaceinstitute.org/report/participatory-data-stewardship/

Self-certification vs third party certification; ex ante scrutiny vs p ost-market scrutiny

The Chapter III approach has a big problem around enforcement. As has been well ventilated by civil society, the requirements of Chapter III can, for most 'high-risk' AI systems, be met by self-certification; there is real risk of this being exploited adversely to avoid true scrutiny or reflection, and a good argument can be made that – given the history of inadequate self-regulation online relating to privacy,[22] *ex ante* certification by an external third party should be required. On the other hand, this may be regarded as disproportionately costly and restrictive of innovation and might encourage regulatory arbitrage to less stringent jurisdictions.[23]

The question of how much burden of prior certification to impose on the production and distribution of various high-risk products and services such as medicines, vaccines, environmentally toxic products such as chemicals, cars and, now, AI is globally conflicted, with multiple different models in operation, from the extensive prior vetting of drugs by the US Food and Drugs Agency (FDA) to the largely self-certificatory and private Data Protection Impact Assessment (DPIA) set out in the GDPR for certain types of high-risk personal data processing. A number of jurisdictions are experimenting with algorithmic impact assessments of various types, such as, notably, Canada.[24] Sectoral assessments are also being proposed.[25] In the private sector many companies including tech giants are implementing or trialling various types of differently scoped AI impact assessment tests.[26] In the Ada Lovelace Institute's current work within the UK for the NHS AI Lab,[27] they have explored as a practical use case whether risk and impact assessment can only best be delivered by *ex ante* external certification, or if internal *ex ante* self-certification backed by external *post-factum* audit,

22 See most notably, the failure of safe harbor as a largely self-certificated scheme for guaranteeing the privacy safeguards for EU data sent to US companies: see CJEU C-362/14 *Schrems v DPC*.

23 The EU AI Act avoids this threat by demanding certification as a condition of entry to the EU internal market thus effectively applying its rules extraterritorially; this might however lead to it being abandoned as a market by some providers (or by seeking entry to the Single Market via 'passporting' by less demanding member states).

24 See *footnote* 3 above. The US has also explored the concept both at state and federal level; The Algorithmic Accountability Act, proposed in the US Congress in 2019, would have require companies with large userbases to conduct impact assessments of their automated systems that affect certain sensitive domains of people's lives

25 Most notably, in respect of digital labour and performance management – see *footnote* 13.

26 See Watkins, E. A., Moss, E., Metcalf, J., Singh, R. and Elish, M. C. (2021).'Governing Algorithmic Systems with Impact Assessments: Six Observations'. *AIES'21: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp1010-1022. Available at: https://doi.org/10.1145/3461702.3462580

27 Ada Lovelace Institute. (2022). *Algorithmic impact assessment: a case study in healthcare.* Available at: https://www.adalovelaceinstitute.org/report/algorithmic-impact-assessment-case-study-healthcare/

might do, with audit being on a regular schedule or possibly triggered by an external risk-displaying event. These types of post-audit or assurance are also being extensively investigated in the UK by *inter alia* the Information Commissioner's Office and the Centre for Data Ethics and Innovation.[28]

The debate about *ex ante* assessment of AI systems versus post-audit is a difficult and complex one. *Ex ante* algorithmic impact assessment for every AI system, especially if defined as widely as in the AI Act will be a source of uncertainty for investors, developers and customers, and will be costly, especially for deployers rather than providers, for whom it might be regarded as a reasonable cost of business (see Figure 1 below). Third-party certification, rather than internal or self-certification, will further add costs. It seems an impact assessment of adding an impact assessment might need to be done!

This proposal may already be seen as particularly pre-market assessment-heavy, given that we have already suggested that a prior assessment by providers, possibly with external certification, must be made of whether an AI system falls into the high-risk category which triggers the application of Chapter III (see point 3 above). We nonetheless argue (as indeed do the drafters of the AI Act) that economics and promotion of AI need to be balanced with the social value of public trust and, hence, uptake as well as protection from AI-inflicted harms.

Grafting a true *ex ante* impact assessment and/or *post-factum* audit on to the AI Act structure may simply not be politically feasible, which would explain the partial human rights scrutiny compromise of Chapter III. If that is true, an entirely new structure needs to be considered for EU AI regulation.

28 See Ahamat, G., Chang, M. and Thomas, C. (2021). 'The need for effective AI assurance'. *Centre for Data Ethics and Innovation*. Available at: https://cdei.blog.gov.uk/2021/04/15/the-need-for-effective-ai-assurance/

One way to reduce the burdens of *ex ante*, and especially external, certification would be to reduce the overall scope of the AI Act, which is currently very wide, and embraces traditional software systems based on rules or logic, as well as the machine-learning systems that we mostly now think about when we say 'AI'.[29]

The aim of the Act, we suggest, should not be to regulate every piece of software in a digital world; this would be better approached by proportionate sectoral legislation. Scope reduction is included to a small extent in the draft Council position. Some organisations such as EDRi and Access Now[30] have already argued that an arbitrary reduction of the scope of the Act to machine learning only might be highly damaging for fundamental rights protection.[31] On the other hand, as Veale and Borgesius[32] have highlighted, the current very wide scope of the AI Act may pose a threat to more effective regulation of sectors of AI largely untouched by any mandatory safeguards within the Act scheme, given the potential pre-emption effect of a maximum harmonisation measure.

Finally, we suggest that real consideration needs to be given to how best to incentivise full engagement with an algorithmic impact assessment (AIA). A market-based solution would be to tie any eventual liability under reformed product liability law[33] (or, indeed, national tort laws) to whether the obligations under the AI Act, including to carry out the AIA, were judged as fully and with due diligence met. Another would be to insist on third-party audit and publication of the AIA. These different levers for compliance will be investigated in further work from Ada.

29 AI Act, art 2.

30 EDRi. (2022). 'Open Letter – Civil society calls for AI red lines in the European Union's Artificial Intelligence proposal'. Available at: https://edri.org/our-work/civil-society-call-for-ai-red-lines-in-the-european-unions-artificial-intelligence-proposal/

31 See for example the Dutch SYRi benefits system scandal, which did not involve an advanced ML system: Vervloesem, K. (2020). 'How Dutch activists got an invasive fraud detection system banned'. *AlgorithmWatch*. Available at:https://algorithmwatch.org/en/syri-netherlands-algorithm/

32 Veale, M., and Borgesius, F. Z., (2021) 'Demystifying the Draft EU Artificial Intelligence Act'. *Computer Law Review International*, 22(4), pp. 97-112.

33 See *Evaluation of Council Directive 85/374/EEC of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products*, (Available at: https://ec.europa.eu/transparency/documents-register/detail?ref=SWD(2018)157&lang=en) and *Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics* (Available at: https://ec.europa.eu/info/publications/commission-report-safety-and-liability-implications-ai-internet-things-and-robotics-0_en). Following a consultation questionnaire launched in 2021 which closed in January 2022, a full proposal for how product liability should be adapted for AI is now awaited.

Some proposed solutions

1. The AI Act should be restructured to provide adequate oversight of general-purpose AI systems by providers and deployers

The current term in the AI Act, 'user' should be renamed *deployer*. Providers and deployers of general-purpose AI should share responsibility for assessing its conformity with fundamental rights and with the safety standards of the Chapter III essential requirements where applicable, and without need to prove the deployer has made a 'substantial modification'. In the draft Council position (inserted Article 52a), this is partly adopted: it is clarified that any person (a deployer, in effect) who 'puts into service or uses' a general-purpose AI system for an intended high-risk purpose comes under duties to certify confirmation with the essential requirements of Chapter III without need to prove substantial modification. At the same time, however, the Council proposal removes liability for the upstream provider of the general purpose AI (Article 52a(1)).

We propose that responsibility cannot and should not be allocated to the deployer alone, since the power to control and modify such infrastructure, alongside technical resources, largely lies with the upstream provider. And that responsibility should be joint with the provider. As with the recent GDPR jurisprudence on joint data controllers, a much more nuanced appraisal must be made of what duties should lie where at what point in time, and who is empowered either legally or by practical control, power or access to data and models, to make changes. As a start, mandatory access to training-set or testing data must be provided where a downstream deployer takes a general-purpose AI system into a high-risk category (see Example 1 above at p. 8).

2. Participation of those impacted by AI systems in their design and safeguarding must be enabled, and due regard given as appropriate, to their views, at all stages of oversight of AI systems.

We propose that:

- Those most affected by the impacts of high-risk AI systems – both individuals and groups – must have input at the time when those systems are certified as compliant with the requirements of Chapter III. If an *ex ante* impact assessment is additionally introduced as canvassed below, particularly where it refers to fundamental rights, those affected must be consulted during that process and their views given due regard. We have suggested that innovative methods such as standing representative panels or 'citizens juries' might be explored, as well as conventional representation through civil society.

- Similar rights to participate must be made available in the standard-setting activity envisaged in Chapter 5, either directly or via civil society representatives. Again, public deliberation mechanisms such as standing panels or citizens juries could be used to provide input efficiently and democratically. We call nonethless for better resourcing and technical resources for civil society, so they can properly fulfil their advocacy role. We note the suggestion that a central technical task force should be established to assist state Market Surveillance Authorities (MSAs)[34] but also argue such a task force would be even better established to aid civil society, where resources and technical expertise are thinly spread.

- Those impacted should have rights to make complaints about all AI systems when they have been put into operation or on the market, to a national regulator and/or a central EU AI Ombudsman (see below). This right should not just apply to 'high-risk' systems. Individual redress should always be available for algorithmic harms, whether through national law, product liability rules or otherwise. Representational actions akin to those under the GDPR Article 80 should also be available. Currently the AI Act envisages only that reports on flaws of systems once on the market are fed back from system *deployers* ('users'): this is both insufficient and unenforceable, with few incentives

34 The European Consumer Organisation (BEUC). (2021). *Regulating AI to Protect the Consumer – Position Paper on the AI Act*. Available at: https://www.beuc.eu/publications/beuc-x-2021-088_regulating_ai_to_protect_the_consumer.pdf

for deployers to comply. Vitally, complaints from those impacted by systems, as well as their deployers, must also be fed back into the design of the system in question. This is particularly important for general-purpose AI systems where we have seen that bias and unfairness may be embedded 'upstream' before deployment in particular contexts and so such alerting is particularly crucial.

- Existing state regulators such as DPAs and MSAs under the AI Act may be overwhelmed, under-resourced and inaccessible to those who are impacted by AI systems. The GDPR experience has shown us that state level regulators may become bottlenecks to action, or subject to industry capture, which are some of the reasons that enforcement mechanisms are currently being reconsidered.[35] This leads us to consider models for complaint and redress from, in particular, administrative and consumer law in the form of an independent EU AI Ombudsman. Their role could include a cross-national element, e.g. collating complaints from national regulators, producing transparency reports, and assessing and grouping repeated complaints and passing them on as 'super-complaints'[36] for priority action to the relevant regulator or court, as well as assisting civil society in preparing representative actions.

3. Justifiable and reviewable criteria should be set for categorising AI systems as 'high risk' rather than an arbitrary list.

We suggest that the criteria for the Commission adding new AI systems under the existing 'high-risk' categories, laid out in Article 7, should be adopted with appropriate modifications to become the criteria for categorisation of systems as 'high risk'. Ideally this self-categorisation should be certified by a third party. These criteria will then be available to courts or regulators to assess certification when or if issues arise. We also think consideration should be given to applying a similar classification regime to prohibited-risk AI and limited-risk AI (where the systems included are particularly arbitrary – though our view is that that category as currently designed is of little value anyway in terms of fundamental rights protection).

35 See International Association of Privacy Professionals. (2021). 'EDPS discusses GDPR enforcement review proposal'. *iapp.org*. Available at: https://iapp.org/news/a/edps-discusses-gdpr-enforcement-review-proposal/ /

36 In the UK, see section 11(1) of the Enterprise Act 2002. Discussion at: Competition and Markets Authority. (2015). 'What are supercomplaints?'. Available at: https://www.gov.uk/government/publications/what-are-super-complaints/what-are-super-complaints

4. Providers and deployers of AI systems must participate in assessing the risks, impacts and potential harms of AI systems to both individuals and society, including impacts on fundamental/ human rights

- An emerging mechanism for this is an *ex ante* fundamental rights impact assessment. The value of this in the AI Act scheme, and the potential for overly burdensome and duplicative regulation, needs to be assessed in the context of the already existing requirement for high-risk AI to certify conformity with Chapter III, as well as the likely possibility of a DPIA being required. Duplication of existing obligations under Chapter III for high-risk AI, alongside new obligations to undertake a fundamental rights impact assessment, *and* to assess as high-risk, should be avoided (see Figure 1 below).

- We do not feel self-certification alone for essential requirements in Chapter III can produce the safe, high-quality AI systems that society needs; external scrutiny by accredited third parties is necessary **unless some other equally effective safeguarding mechanism is provided by addition post-market audit obligations**.

- If *ex ante* impact assessments are introduced in addition to Chapter III for high risk AI, these should:

  - consider group and societal values as well as fundamental rights and environmental impacts

  - demonstrate efficacy and be determined by whether an AI system is in fact needed at all and if it is consonant with human dignity

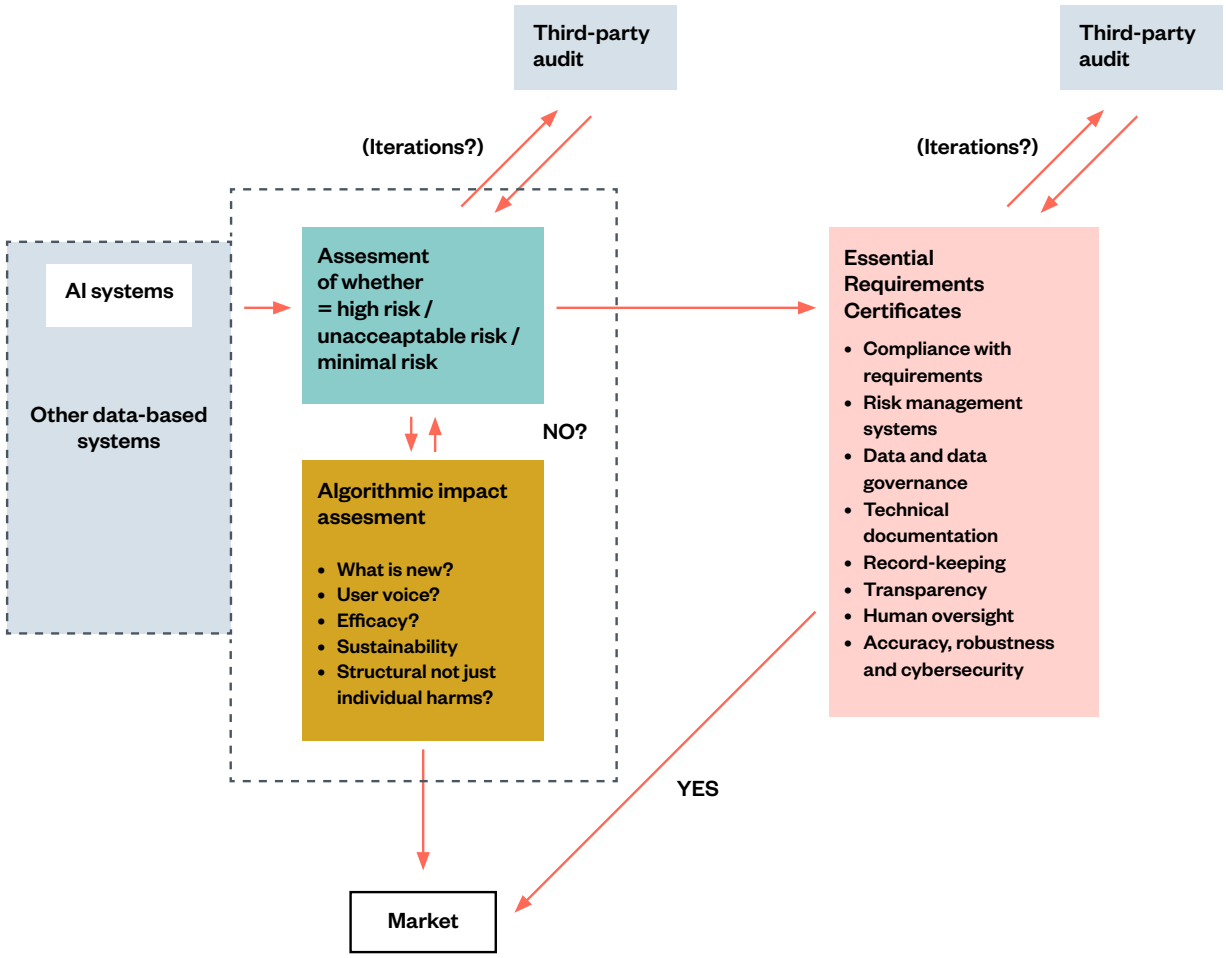
  - consider the views of individuals, groups and communities actually and potentially affected. In particular, designers should be required by participatory processes to examine how AI systems might be misused in deployment contexts to harmfully impact the vulnerable

  - be made public to encourage accuracy, enable scrutiny and provide templates for other providers, especially SMEs. Research should consider if external scrutiny can be supplied by regular audit after the system is deployed instead of, or to supplement, *ex ante* assessment.

- In an aspirational model not constrained by the EU's New Legislative Framework (NLF), it is possible that all the following might be combined as a layered set of compliance processes:
  - a regime for categorisation as 'prohibited/high/limited/minimal risk' AI
  - an *ex ante* impact assessment process, and
  - 'essential requirements' certification.

Within the current EU AI Act structure, the process would then look like this:

Figure 1: Flow chart of potential new system incorporating risk-categorisation, and impact assessment modules



This is not ideal and may be seen as an over-burdened and potentially repetitive process. But commonalities and overlaps can be established across the risk-assessment, impact-assessment and 'essential requirements' processes, so that compliance becomes a systematic, layered process of classification, design, testing, monitoring and building in of safeguards, as opposed to a simple checklist for an arbitrary and relatively small list of systems. Such a layered process could save developers time and money. Early red flags from the risk-assessment module could be a signal to redevelop the system before putting it on to market, saving time and liability later.

It is probably too late in the AI Act process to embed such a model as it would mean tearing up Chapter III, but it is not too late for other states considering regulation.

On the other hand, *ex ante* algorithmic impact assessments (AIAs), however formulated, can never alone be a 'silver bullet'. We recommend them with some hesitancy, knowing that a number of key problems will need navigated. First, 'AIA' is a contested and emerging term that is poorly defined and, as discussed above, is not clearly pinned down in its interaction with *post-factum* audit. The objectives, requirements and limits of an impact assessment process need in-depth discussion. In the Ada Lovelace Institute's NHS AI Lab AIA, the objectives included encouraging more reflexivity with product teams, documenting key decisions and making them transparent, and enabling affected communities of these systems to have more of a say in the construction of impacts. It is possible that no over-arching aims of this kind can be conceived of for something as widely scoped as 'AI' overall, especially as defined in the AI Act. An answer may be the development of sectoral AIAs; work is already advancing on these in fields like labour.

Secondly, we must acknowledge the limitations of AIAs. These are not a form of accountability in themselves, but they can enable a more accountable relationship between regulators, members of the public and providers/deployers, by empowering the former two to ask questions,

pass judgement, and enforce sanctions on the latter. And they are not a crystal ball – they must be deployed as part of a wider set of accountability practices. In particular, the post-market enforcement processes of the AI Act as it currently stands are particularly weak. Market surveillance authorities (MSAs) are far less 'hands on' in enforcement than state Data Protection Authorities (DPAs) were under the GDPR. DPAs are themselves coming under a barrage of criticism, and as noted above, the role of user complaints in provoking enforcement, which has worked well in the GDPR, is entirely absent from the AI Act. There is considerable inequality of power between state bodies and tech giants, and a recognised danger of regulatory capture and possible capitulation in state MSAs, especially in smaller states. In the EU system, furthermore, the overarching body envisaged to bring harmonisation to enforcement (the EU AI Board) has so far, unlike its DP equivalent, no clear purpose or powers.

About the author

Lilian Edwards is a leading academic in the field of Internet law. She has taught information technology law, e-commerce law, privacy law and Internet law at undergraduate and postgraduate level since 1996 and been involved with law and artificial intelligence (AI) since 1985.

She worked at the University of Strathclyde from 1986–1988 and the University of Edinburgh from 1989 to 2006. She became Chair of Internet Law at the University of Southampton from 2006–2008, and then Professor of Internet Law at the University of Sheffield until late 2010, when she returned to Scotland to become Professor of E-Governance at the University of Strathclyde, while retaining close links with the renamed SCRIPT (AHRC Centre) at the University of Edinburgh. She resigned from that role in 2018 to take up a new Chair in Law, Innovation and Society at Newcastle University.

She is the editor and major author of *Law, Policy and the Internet*, one of the leading textbooks in the field of Internet law. She won the Future of Privacy Forum award in 2019 for best paper ('Slave to the Algorithm' with Michael Veale) and the award for best non-technical paper at FAccT in 2020, on automated hiring. In 2004 she won the Barbara Wellberry Memorial Prize in 2004 for work on online privacy where she invented the notion of data trusts, a concept which ten years later has been proposed in EU legislation. She is a partner in the Horizon Digital Economy Hub at Nottingham, the lead for the Alan Turing Institute on Law and AI, a Turing fellow, and a fellow of the Institute for the Future of Work. At Newcastle, she is the theme lead in the data NUCore for the Regulation of Data. Edwards has consulted for, inter alia, the EU Commission, the OECD, and WIPO. In 2021-22, she is part-seconded to the Ada Lovelace Institute to lead their work on the future of global AI regulation.

About the Ada Lovelace Institute

The Ada Lovelace Institute was established by the Nuffield Foundation in early 2018, in collaboration with the Alan Turing Institute, the Royal Society, the British Academy, the Royal Statistical Society, the Wellcome Trust, Luminate, techUK and the Nuffield Council on Bioethics.

The mission of the Ada Lovelace Institute is to ensure that data and AI work for people and society. We believe that a world where data and AI work for people and society is a world in which the opportunities, benefits and privileges generated by data and AI are justly and equitably distributed and experienced.

We recognise the power asymmetries that exist in ethical and legal debates around the development of data-driven technologies, and will represent people in those conversations. We focus not on the types of technologies we want to build, but on the types of societies we want to build.

Through research, policy and practice, we aim to ensure that the transformative power of data and AI is used and harnessed in ways that maximise social wellbeing and put technology at the service of humanity.

We are funded by the Nuffield Foundation, an independent charitable trust with a mission to advance social well-being. The Foundation funds research that informs social policy, primarily in education, welfare and justice. It also provides opportunities for young people to develop skills and confidence in STEM and research. In addition to the Ada Lovelace Institute, the Foundation is also the founder and co-funder of the Nuffield Council on Bioethics and the Nuffield Family Justice Observatory.

Find out more:

Website: Adalovelaceinstitute.org
Twitter: @AdaLovelaceInst
Email: hello@adalovelaceinstitute.org

Preferred citation: Edwards, L. (2022). *Regulating AI in Europe: four problems and four solutions.* Ada Lovelace Institute. Available at: https://www.adalovelaceinstitute.org/report/regulating-ai-in-europe/

ISBN: 978-1-7397950-0-9