The Bar Council

Law Reform Essay Competition Winner 2023

‘Not OK Computer: A Proposed AI Transparency Framework for the UK’ by Louis Dejeu-Castang

1. Introduction

Louis Brandeis wrote that “*sunlight is said to be the best of disinfectants; electric light the most efficient policeman*”.[1] The following paper will argue for a transparency law to act as a harsh spotlight pointed directly at artificial intelligence (AI),[2] a technology that is both underregulated and shrouded by its technical complexity. Four basic ideas are at the centre of this proposal: (1) *higher risk AI systems require higher levels of transparency*,[3] (2) *the unforeseen consequences of AI requires a nuanced approach to assessing risk*, (3) *the public should be*

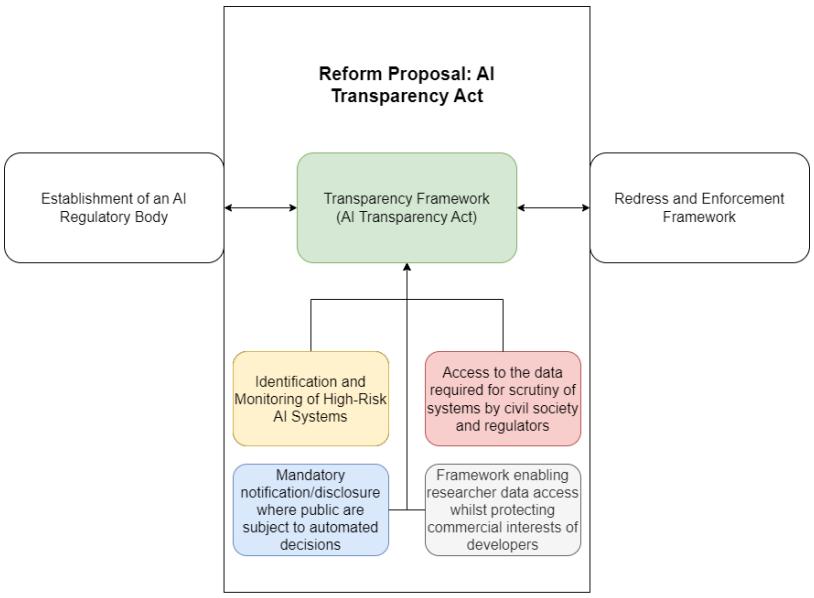

Figure 1: Scope of the Proposal

[1] Brandeis, L.D., 1914. Other People’s Money: And how the Bankers Use it, HeinOnline Legal classics library. Stokes.

[2] AI is an umbrella term encompassing machines that make decisions in an ‘intelligent’ way, often akin to human thought. See Manning, C., 2020. Artificial Intelligence Definitions. Stanford University Human-Centered Artificial Intelligence. Available at: https://hai.stanford.edu/sites/default/files/2020-09/AI-Definitions-HAI.pdf (accessed 10.10.23).

[3] AI system definition (OECD): “*machine-based system capable of influencing the environment by producing an output (predictions, recommendations or decisions) for a given set of objectives*”. AI-Principles Overview, OECD.AI. Available at: https://oecd.ai/en/principles (accessed 10.10.23).

aware of how automated decisions are made about important aspects of their life, and (4) *whilst researcher data access is important to scrutinise AI systems, this must be accompanied by protections for the commercial interests of AI developers*. This paper first addresses transparency conceptually, before examining gaps in existing law. Finally, it sets out a framework for allocating transparency obligations based on risk, whilst protecting commercially-sensitive information.

2. Transparency – the cornerstone of AI Regulation

Transparency generally revolves around increasing the visibility of a system/entity's workings, either by mandating the publication of comprehensible information for a lay audience, or by supplying experts with the information required for effective scrutiny. This section will examine how transparency ought to be conceptualised faced with the challenges of AI.

2.1. What should Transparency mean in an AI context?

2.1.1. Safety by Design

A significant obstacle to regulating technology is its rapid rate of change. Parliament faces an uphill struggle to create legislation with both the breadth to allow flexibility, and precision to address specific harms and nuances. Consequently, dialogue around regulating AI takes on a technological-determinist tone by referring to law "*keeping pace*" with developments.[4] Transparency provides an alternative. Regulating how AI systems are trained, designed, and deployed can help regulators drive AI development around pro-social norms rather than attempt to impose standards on the technology *ex post-facto*. For example, Aizenberg frames stakeholder debate about a technology's harms and "*investigations*" into technological

[4] Donelan, M., 2023. UK unveils world leading approach to innovation in first artificial intelligence white paper to turbocharge growth [WWW Document]. GOV.UK. Available at: https://www.gov.uk/government/news/uk-unveils-world-leading-approach-to-innovation-in-first-artificial-intelligence-white-paper-to-turbocharge-growth (accessed 10.9.23).

solutions as crucial to the translation of values into design principles.[5] Transparency can therefore be thought of as crucial to stakeholders feeding into AI development and legislation.

2.1.2. **Measurement of Performance/Accountability**

AI is overwhelmingly developed in the private sector by small teams of engineers.[6] Anticipating economic and social impacts of the model rests with a small group of individuals who are unlikely to have a background in these fields. [7][8] Lack of diversity among technologists hinders recognition of biases at a design stage and increases the likelihood that models incorporate the worldview of their predominantly white male developers.[9] Transparency enables oversight from cross-sector stakeholders capable of recognising the impact of models on society, including communities most likely to be harmed by bias, in the absence of commercial motivations. Transparency is also critical to assessing AI performance; a general-purpose model might have unforeseen consequences in specific fields. For example, an algorithm intended to reduce the number of medical scans to manually examine saw radiologists spend "*more time analysing 40 AI-flagged x-rays than they did 100 non-flagged x-rays*"[10] to determine what the algorithm might have spotted. Accordingly, companies and the public sector ought to have some means of independently assessing the risks of different models.

2.1.3. **Transparency and agency**

Transparency about automated decisions enhances freedom in two senses: firstly, it facilitates attempts to seek recourse for potential biases by appeal or raising the issue with civil society, and secondly, understanding the algorithm's parameters helps individuals to make informed

---

[5] Aizenberg, E., Hoven, J. van den, 2020. Designing for Human Rights in AI. Big Data & Society 7, 205395172094956. https://doi.org/10.1177/2053951720949566.
[6] OpenAI has approximately 375 employees as of 2023.
[7] Much AI terminology is disputed – for simplicity I use 'AI system' and 'model' interchangeably.
[8] Blackman, R., Ammanath, B., 2022. Building Transparency into AI Projects. Harvard Business Review.
[9] Only 20% of AI and data professionals are women. Women in Data Science and AI [WWW Document], The Alan Turing Institute. Available at: https://www.turing.ac.uk/research/research-projects/women-data-science-and-ai-new (accessed 8.9.23).
[10] Magnus, D., 2019. The Ethics of Artificial Intelligence in Medicine. https://www.youtube.com/watch?v=6NkFPJfSyRM; Blackman, R., Ammanath, B., 2022. Building Transparency into AI Projects. Harvard Business Review.

decisions about participating in otherwise lawful activities that could adversely affect automated decisions about them. Broadly, transparency enables individuals to better understand the consequences of their actions, rather than be the passive subject of a faceless algorithm. As Berlin writes, central to freedom is the desire to be "*a doer--deciding, not being decided for, self-directed and not acted upon by external nature"*.[11]

2.2. The Black Box: Obstacles to Transparency

Technical complexity in AI models is an obstacle to lawmakers assessing their impact. Deep learning algorithms consist of layers of interconnected neurons resembling the structure of the human brain.[12] Per Mahapatra, even when possible to determine which nodes are activated "*we don't* [necessarily] *know what their neurons were supposed to model and what these layers of neurons were doing collectively*".[13] De Bruijn, Warnier and Janssen further question the feasibility of abstracting and simplifying AI decisions. They argue that models rarely resemble static decision-making processes; instead, more dynamic algorithms will constantly alter their parameters to learn from past decisions and new data.[14] This uncertain causal relationship between inputs and outputs in a constantly changing system presents a significant challenge to experts attempting to explain AI decisions.

Further, there are risks to sharing information about a model with the wrong groups. Caplan et al. highlight the "gaming" of Twitter's trending algorithm to promote political messages as

---

[11] Berlin, I., 2002. Two Concepts of Liberty, in: Berlin, I., Hardy, E. by H. (Eds.), Liberty. Oxford University Press. https://doi.org/10.1093/019924989X.003.0004.

[12] Deep learning: The use of multiple-layered neural networks to enable machines to "learn" from large datasets that currently underpins most sophisticated AI systems. What is Deep Learning?, IBM. Available at: https://www.ibm.com/topics/deep-learning (accessed 10.9.23).

[13] Mahapatra, S., 2019. Why Deep Learning over Traditional Machine Learning? [WWW Document]. Medium. Available at: https://towardsdatascience.com/why-deep-learning-is-needed-over-traditional-machine-learning-1b6a99177063 (accessed 10.1.23).

[14] De Bruijn, H., Warnier, M., Janssen, M., 2022. The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. Government Information Quarterly 39, 101666. https://doi.org/10.1016/j.giq.2021.101666, 4.

an example of bad actors exploiting information about an algorithm to their own ends.[15][16] Additionally, developers investing in new models have a legitimate expectation of ownership over their code and trade secrets. Transparency laws should therefore respect the commercial interests of developers, lest transparency come at the expense of innovation.

Finally, effective regulation of AI is predicated on regulators and lawmakers having access to the expertise and information required to understand how models work. Currently developers hold more information about how their models work than anyone tasked with holding them accountable.[17] Businesses and lawmakers lack the technical awareness to evaluate claims made by developers about the effectiveness of their systems.

3. Limitations in the current regulatory framework

Per the government AI whitepaper, there is unlikely to be any upcoming AI-specific legislation.[18] Accordingly, the following largely ineffective UK GDPR/DPDI transparency provisions are central to the current framework:[19]

- **Articles 22A-C UK GDPR/DPDI** set out restrictions and safeguards for automated-decision making, stipulating that where a "*significant decision is taken by or on behalf of a controller […] based solely on automated processing*" the controller must protect the data subject's rights through certain safeguards.[20] Veale and Edwards argue that limiting safeguards under Article 22 to decisions based *solely* on automated processing

[15] Caplan, R., Donovan, J., Hanson, L., Matthews, J., 2018. Algorithmic Accountability: A Primer. Data & Society.
[16] Macaulay, T., 2021. How hackers have manipulated Twitter's trending algorithm for years [WWW Document]. TNW | Deep-Tech. Available at: https://thenextweb.com/news/twitter-trending-topics-algorithm-has-vulnerability-hackers-using-ephemeral-astroturfing-attacks (accessed 8.11.23).
[17] Rong, R., 2023. Information Asymmetry in Artificial Intelligence: Consumers Bear the Brunt When Good AI is Driven…. Medium. Available at: https://medium.com/@rebecca.r.rong/information-asymmetry-in-artificial-intelligence-consumers-bear-the-brunt-when-good-ai-is-driven-399578e00ba9 (accessed 10.16.23).
[18] "Initially, we do not intend to introduce new legislation." A pro-innovation approach to AI regulation, 2023. Department for Science, Innovation & Technology. Available at: https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper.
[19] The Data Protection and Digital Information (No. 2) Bill, currently at the report stage in Parliament, is expected to amend the UK GDPR.
[20] Articles 22A-C, Data Protection and Digital Information (No. 2) Bill 2022-23. Available at: https://publications.parliament.uk/pa/bills/cbill/58-03/0314/220314.pdf.

removes much of its force, noting that the majority of *"ML systems that affect people's lives significantly are usually not fully automated*", instead usually acting as "*decision support*" in a system involving some human input.[21] Furthermore, human input should not be the deciding factor in setting transparency requirements; many AI harms have been recorded in partially automated systems.[22]

- **Articles 5, 13, and 14 UK GDPR:**[23] Subsection 2 of Articles 13 and 14 requires that controllers provide data subjects with "*meaningful information*"[24] to the "*extent necessary*" for fair and transparent processing. Limitations of these provisions include the vague scope of "*meaningful information*" and "*to the extent necessary*". Lawrence-Archer and Naik note that in practice these requirements are usually discharged via a "*high-level privacy notice that discusses in general terms (albeit at great length) the kinds of processing that a controller engages in*."[25] Much like the often ignored "terms and conditions" attached to online contracts, documents of this nature require patience and legal knowledge to overcome the many pages of dense jargon. Moreover, companies will simply cite impracticality to avoid releasing information about specific decisions, despite specificity being critical to individuals understanding how decisions have affected them. Likewise, Article 5 is typically satisfied by lengthy non-specific privacy notices.[26]

---

[21] Edwards, L., Veale, M., 2017. Slave to the Algorithm? Why a Right to Explanation is Probably Not the Remedy You are Looking for. Duke Law & Technology Review 16, p.45. Available at: https://doi.org/10.2139/ssrn.2972855.
[22] Veale and Edwards provide examples at p.46-48, Ibid. Additionally, "automation bias" sees humans over-rely on decisions made by automated systems.
[23] Articles 5, 13, and 14, Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation). Available at: https://www.legislation.gov.uk/eur/2016/679/contents#.
[24] Article 13(2) and 14(2) UK GDPR.
[25] Lawrence-Archer, A., Naik, R., 2023. Effective protection against AI harms. AWO, p. 9-11. Available at: https://www.awo.agency/files/AWO%20Analysis%20-%20Effective%20Protection%20against%20AI%20Harms.pdf.
[26] Requirement that data be processed "*in a transparent manner in relation to the data subject*". Article 5, UK GDPR. Ibid.

Similarly, the government AI white paper sets out guiding principles for pre-existing regulators.[27] Though including useful guidance, including on transparency and explainability, without new powers or incentives to encourage industry compliance the paper lacks the force to underpin a robust transparency regime.

The current framework is therefore flawed. Firstly, it **fails to provide data subjects with the transparency necessary to understand how they are affected by automated decisions**. Secondly, there is an **absence of law aimed at AI developers** despite them being best placed to anticipate and mitigate risk at a pre-market stage.[28] Additionally, **researchers lack tools to scrutinise and evaluate the impact of AI**. Finally, **existing regulators have neither the resources nor expertise to adequately deal with the issues that AI presents**. Research commissioned by the Ada Lovelace Institute found the ICO and EHRC "*do not have sufficient powers, resources, or sources of information*" to enforce UK law "*with a completeness that will reliably protect against AI harms*".[29]

4. Transparency in practice: a reform proposal

[27] A pro-innovation approach to AI regulation, 2023. Department for Science, Innovation & Technology. Available at: https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper.
[28] Regulating AI in the UK, 2023. Ada Lovelace Institute, p.49. Available at: https://www.adalovelaceinstitute.org/wp-content/uploads/2023/09/ALI_Regulating-AI-in-the-UK_2023.pdf.
[29] Lawrence-Archer, A., Naik, R., 2023. Effective protection against AI harms. AWO. Available at: https://www.awo.agency/files/AWO%20Analysis%20-%20Effective%20Protection%20against%20AI%20Harms.pdf.

I propose a tiered approach to transparency, consisting of a hierarchy of risk and a hierarchy of access, to be incorporated into an AI Transparency Act.[30] The former resembles the framework of risk categorisation in the AI Act,[31] but subdivided into risk by *use*, *scale*, and *proximity*. The latter outlines the level of access to AI technology afforded to different groups, aiming to mitigate the risk of revealing sensitive information.

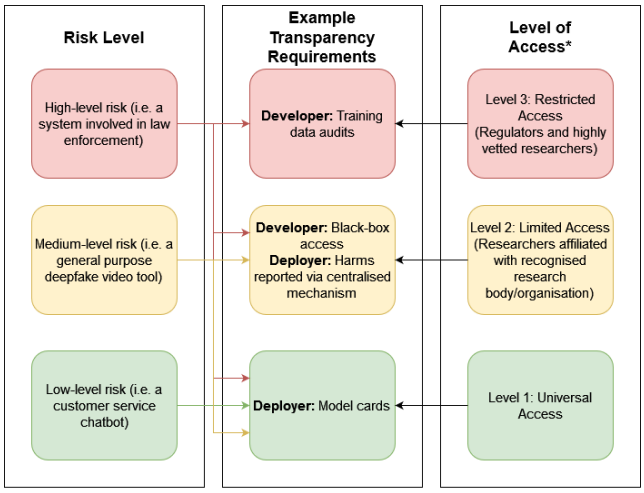

Figure 2: Relationship between hierarchies and transparency requirements

* Note that the access and risk hierarchies will not always align (i.e. some disclosures relating to high-risk systems should be universally accessible - such as the database of high-risk systems in Annex 1)

4.1. Risk categorisation

The opacity and complexity in how AI systems are developed and deployed creates the need for more flexibility in assessing a system's risk than under the EU framework. Rather than using a pre-determined list, updated via the adoption of delegated acts by the Commission,[32] to classify whether an AI system is high-risk, this proposal evaluates the risk-level of systems holistically.

[30] The scope of this proposal is confined to AI transparency; a robust regime would need other legislation (i.e. to establish redress mechanism for challenging decisions).
[31] The AI Act sorts systems into risk categories but lists certain types of system as innately high-risk. Annex III, 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 COM/2021/206 final, European Commission. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.
[32] Articles 7 and 73, Artificial Intelligence Act.

This section suggests three metrics anticipating a system's potential to harm. Taken together these factors will enable an AI regulator to determine a system's level of legal obligation under the framework.[33] Logically, heightened transparency should follow heightened risk.

4.1.1. **Risk by Use**

The intended use of an AI system is the logical starting point for assessing its potential to cause harm. An important factor in determining risk of "*harm to health and safety or risk of adverse impact on fundamental rights*"[34] in the AI Act is "*the intended purpose of the AI system*".[35] There are certain uses of AI that from the onset can be identified as having a high-risk to fundamental rights. One formulation of this test could be **a system whose use has a high risk of undermining the fundamental rights of citizens or causing injustice, but not so high as to be unacceptable in a free and fair society**.[36]

Drawing on the AI Act and the work of academics, examples of systems that might be considered high-risk by nature include:

- systems involved in "border monitoring and surveillance";[37]
- systems "used to evaluate the credit score or creditworthiness of natural persons";[38]
- "predictive analytic systems used in migration, asylum and border control";[39]
- systems used by law enforcement; and
- systems used to determine access to basic services, such as healthcare.[40]

4.1.2. **Risk by Scale**

---

[33] Though integral to an effective regulatory regime, outlining the role and powers of an AI-specific regulator is beyond the scope of this proposal.
[34] Recital 28, Artificial Intelligence Act.
[35] Article 7(2)(a), Artificial Intelligence Act.
[36] Systems with an unacceptable risk ought to be prohibited in separate legislation.
[37] Civil Society Joint Statement on the AI Act. Amnesty International. Available at: https://www.amnesty.eu/wp-content/uploads/2022/12/Open-letter_EU-AI-Act_migration_December-2022.pdf.
[38] Recital 37, Artificial Intelligence Act.
[39] Civil Society Joint Statement on the AI Act. Amnesty International. Available at: https://www.amnesty.eu/wp-content/uploads/2022/12/Open-letter_EU-AI-Act_migration_December-2022.pdf.
[40] Recital 37, Artificial Intelligence Act.

The scale at which an AI model is deployed is also indicative of its potential to cause harm, and consequently should be considered when determining its regulatory obligations. As Edwards notes, AI is not a "one-off service" but a "system delivered dynamically through different hands".[41] Alongside specific products aimed at serving particular industries are the "foundation models", such as GPT-4, that are capable of general tasks and are therefore often integrated into other systems.[42][43] It is therefore significantly harder to anticipate the risk of a powerful general purpose model which might be the foundation for a multitude of unforeseen applications.

Similarly, even if a model is narrow in purpose, its use in a way that impacts a large number of people can be indicative of its capacity to harm. Accordingly risk by scale should also include an assessment of whether a model is used in a way that impacts a significant number of people.

4.1.3. **Risk by Proximity**

Proximity, here, refers to the degree of involvement that an AI system has in making decisions that are likely to have a substantial impact on access to an essential service by individuals from marginalised or vulnerable groups. For instance, an algorithm being used by a local authority in social housing decision-making is making determinations about individuals with a low-level of income who may be at risk of homelessness should their access to a basic service (housing) be denied. The reasoning behind the proximity metric is to mandate transparency in systems whose intended purpose may seem innocuous but have the potential to cause significant harm if an underlying flaw or bias in its design is left uncorrected.

4.2. Access Hierarchy

---

[41] Edwards, L., 2022. Regulating AI in Europe: four problems and four solutions. Ada Lovelace Institute. Available at: https://www.adalovelaceinstitute.org/report/regulatingai-in-europe/.
[42] Example: GPT-3 serves as the basis for a multitude of products ranging from Dall-E (image generation) to YouChat (chatbot).
[43] Jones, E., 2023. Explainer: What is a foundation model? [WWW Document]. Ada Lovelace Institute. Available at: https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/ (accessed 9.18.23).

Per 2.2, granting universal access to all information about every AI model is clearly undesirable. Accordingly, the UK could draw from the researcher access framework outlined in Article 40 DSA, under which researchers apply to access data from platforms.[44] In the application they must demonstrate affiliation to a research organisation (as defined in an EU Directive on copyright),[45] independence from commercial interests, funding sources, the ability to fulfil confidentiality and data security requirements, that access to the data is necessary and proportionate to their research purposes, and that their research is "*for the sole purpose of conducting research that contributes to the detection, identification and understanding of systemic risks*" and the assessment of risk mitigation measures taken by platforms.[46]

I propose a similar system, with vetting requirements becoming more stringent based on the level of access requested. The more commercially sensitive data is, the higher the security risk of its misuse, and the more sensitive the nature of any personal data, the greater the rigour of the vetting process.

4.2.1. **Level 1: Universal Access**

On a basic level, all individuals who are subjected to automated decision-making where the outcome will involve a significant determination about their health, education, finances, or impact their fundamental rights, should be informed that the decision was either wholly or partially determined by a machine. High-risk decisions should have higher disclosure requirements, even to the public; public sector bodies for instance should disclose where automation is used, the risk assessment carrier out prior to deployment of the system, and what safeguards exist to prevent unfair outcomes. Per section 3, for this information to be useful, it will need to be accompanied by a requirement that it be relayed comprehensibly.

---

[44] Specifically, Very Large Online Platforms (VLOPs) – see Article 40, Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) PE/30/2022/REV/1. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32022R2065.

[45] Article 2(1), Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC PE/51/2019/REV/1. Available at: https://eur-lex.europa.eu/eli/dir/2019/790/oj.

[46] Article 40(4) and (8)(a)-(g), Digital Services Act.

Additionally, the final reports of studies by vetted researchers should be released to the public by default with sensitive data redacted.[47]

Universal access is most relevant to the *agency* and *accountability* aspects of transparency discussed previously. It allows the subjects of automated decisions to identify wrongful decisions and seek legal advice on the matter.[48] A widely accessible repertoire of information about AI models, and absolute transparency as to where such models are used, may also unlock the ability of the public to actively engage in political debate on AI ethics.

4.2.2. **Level 2: Limited Access**

The second access tier should be restricted to researchers affiliated with a recognised academic institution or a non-profit organisation,[49] with approval from a board of independent experts.[50] Until the establishment of an AI regulator, the board could be convened by academic institutions drawing from academics with relevant expertise and no conflicts of interest. Drawing on the DSA and EDMO Code of Conduct,[5152] the role of this panel will be to evaluate the proportionality and necessity of the data access sought, whether appropriate safeguards regarding confidentiality and data protection are in place,[53] and the independence of the project from commercial interests.[54] Simply being affiliated with a research body should not be enough to be granted access to all data below Level 3; researchers will still need to demonstrate that the specific data they seek is necessary to the project. The sensitivity of

[47] See Article 40(8)(g), Digital Services Act.
[48] It is noted that even with increased access to information, the technical complexity of AI will necessitate expert/legal advice to engage with questions of legality.
[49] This could include accredited universities and UKRI research institutions. See: Guidance UKRI list of approved research organisations, 2023. UK Immigration and Visas. Available at: https://www.gov.uk/government/publications/ukri-endorsement-employing-or-hosting-institutions-global-talent-visa/ukri-list-of-approved-research-organisations.
[50] Note that there ought to be exception to this process for regulators, who should have access to Level 2 data subject to internal policies.
[51] Article 40(8)(e), Digital Services Act.
[52] Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access. European Digital Media Observatory. Available at: https://edmo.eu/wp-content/uploads/2022/02/Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf (accessed 10.9.23).
[53] This proposal is intended to complement and not displace existing data protection requirements under UK law.
[54] Article 31, Digital Services Act.

information is contextual; therefore the proportionality test affords the flexibility to rapidly approve low-risk projects whilst applying greater rigour to projects requesting high-risk data.

4.2.3. **Level 3: Restricted Access**

Per section 2.2, some aspects of how AI models operate are highly sensitive. Access to information that poses the highest risk of interference should be confined to a relatively small group of regulators and academics. This could include access to a model's training data or even access to the model itself for sandbox testing.[55] As with Level 2, projects seeking access to data covered at this level would have to submit proposals to be evaluated by a panel convened by the AI regulator. Additional factors for consideration at this level could include the work's methodology, to consider whether they "*are broadly suited to achieve the research objectives*" and the "*researcher(s) involved are qualified to perform the research in question*".[56] On top of a more rigorous approval process, the panel should also have the power to impose conditions on data access to limit the risk of leaks. Drawing on the CMA's procedure for handling confidential information,[57] the panel could impose access conditions such as the use of confidentiality rings or data rooms. Given that a concern is the release of trade secrets, the panel could impose a prohibition on working for a competing company within 3 years of being granted access to highly sensitive data. Furthermore, under this regime researchers in breach of confidentiality rules could be prosecuted and face substantial fines. Finally, prior to publication any work involving Level 3 access should be reviewed to ensure it does not contain

---

[55] Cen, S.H., Fabrizio, C.L., Siderius, J., Madry, A., Minow, M., 2023. Auditing AI: How Much Access Is Needed to Audit an AI System? Thoughts on AI Policy. Available at: https://aipolicy.substack.com/p/ai-accountability-transparency-2?utm_medium=reader2 (accessed 10.9.23).

[56] Drawn from the methodological review at 7.2 of the EDMO Code of Conduct. See 2022. Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access. European Digital Media Observatory. Available at: https://edmo.eu/wp-content/uploads/2022/02/Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf (accessed 10.9.23).

[57] Guidance on the CMA's investigation procedures in Competition Act 1998 cases: CMA8, 2021. Available at: https://www.gov.uk/government/publications/guidance-on-the-cmas-investigation-procedures-in-competition-act-1998-cases/guidance-on-the-cmas-investigation-procedures-in-competition-act-1998-cases#limits-on-the-cmas-powers-of-investigation (accessed 10.9.23).

commercial information whose disclosure *"might significantly harm the legitimate business interests",*[58] or information that risks compromising a model's safe operation.

5. Conclusion

In sum, AI presents specific challenges that require robust transparency laws. As the scale of AI deployment increases, the crucial machinery of scrutiny and accountability, in the form of regulatory bodies as well as the media, legislature and courts, must not be impeded by a veil of technical complexity and trade secrecy. Rather than wait for AI technological developments to set the tempo of regulation, this transparency framework opens the door for civil society and public engagement in discussions on AI, in turn helping lawmakers to shape AI around pro-social norms and values.

Annex 1: Suggested Transparency Obligations

The following examples help clarify how the framework could correspond with legal obligations for developers[59] and deployers[60] of AI systems.

Low-risk systems

Developers:

- Environmental impact assessment
- Model cards[61]

[58] Definition of "confidential information" at 4.14: 2014, Transparency and disclosure: Statement of the CMA's policy and approach. Competition and Markets Authority. Available at: https://assets.publishing.service.gov.uk/media/5a7cc94aed915d63cc65cd6e/CMA6_Transparency_Statement.pdf (accessed 10.9.23).
[59] Developers: entities/persons that develop an AI system either to place on the market or publish free of charge. See Article 3(2), Artificial Intelligence Act.
[60] Entities or persons using AI in a non-personal context. Amended form of the 'user' definition, based on recommendations in Edwards, L. (2022). Regulating AI in Europe: four problems and four solutions. Ada Lovelace Institute, p.20. Available at: https://www.adalovelaceinstitute.org/report/regulatingai-in-europe/.
[61] Document containing basic information about models, including intended uses, evaluation metric, ethical considerations, factors affecting model performance, and recommendations from developers. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D., Gebru, T., 2019. Model Cards for Model Reporting, in: Proceedings of the Conference on Fairness, Accountability, and Transparency. p. 220–229. https://arxiv.org/pdf/1810.03993.pdf.

Deployers:

- Bot-or-not bot disclosure[62]

Medium-risk systems

Developers:

- Provision of a safety document to deployers of the AI system[63]
- Datasheets for datasets [64]
- Black-box access[65]
- Explainability-by-design

Deployers:

- Declaration of usage[66]
- Post-market monitoring[67]
- Human rights impact assessment

High-risk systems

Developers:

- Training procedure auditing[68]

---

[62] See Article 52(1), Artificial Intelligence Act: *"designed and developed in such a way that natural persons are informed that they are interacting with an AI system*".
[63] See "instructions for use": Article 13(2)-(3), Artificial Intelligence Act.
[64] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé III, H., Crawford, K., 2021. Datasheets for Datasets. https://arxiv.org/pdf/1803.09010.pdf.
[65] Researchers and auditors with black-box access can test the outputs of a model by feeding it different inputs. Cen, S.H., Fabrizio, C.L., Siderius, J., Madry, A., Minow, M., 2023. Auditing AI: How Much Access Is Needed to Audit an AI System? Thoughts on AI Policy. Available at: https://aipolicy.substack.com/p/ai-accountability-transparency-2?utm_medium=reader2 (accessed 10.9.23).
[66] Any systems with a moderate or above level of risk should be declared to the regulator, along with the intended use of the system.
[67] Leufer, D., 2022. Why we need human rights impact assessments for AI. Access Now. Available at: https://www.accessnow.org/human-rights-impact-assessment-ai/ (accessed 10.6.23).
[68] Cen, S.H., Fabrizio, C.L., Siderius, J., Madry, A., Minow, M., 2023. Auditing AI: How Much Access Is Needed to Audit an AI System? Thoughts on AI Policy. Available at: https://aipolicy.substack.com/p/ai-accountability-transparency-2?utm_medium=reader2 (accessed 10.20.23).

- Training data auditing[69]
- Full-model sandbox access

Deployers:

- Inclusion on a publicly-accessible database[70]
- Complaints/reporting mechanism

[69] Training data auditing by an AI regulator could be triggered through the reporting mechanism. Ibid.
[70] With exceptions for systems essential to national security.