ONLINE SAFETY ACT
NETWORK

RESPONSE TO UNESCO CONSULTATION ON EMERGING APPROACHES TO AI REGULATION

1. We welcome the opportunity to contribute to UNESCO's work and commend the clarity of the global overview set out in the Emerging Approaches to AI Regulation document.

2. Our interest in this agenda is founded in our work in recent years to influence the development of online safety legislation in the UK, resulting in our recommendation for a "duty of care" approach to harm minimisation in relation to social media underpinning the foundations of the Online Safety Act 2023. Our work was carried out as part of a project for Carnegie UK from 2017-2023, and is continued by Prof Lorna Woods OBE (Essex University) and Maeve Walsh at the Online Safety Act Network. Some of the largest AI systems are arguably those used in social media and search for curation and moderation of content; concerns about synthetic media and their impact on individuals and society are also relevant here.

3. In recent years, we have contributed to the UK's debate on the best approach to AI regulation and our response draws on previous submissions to Parliamentary and Government consultations. The material here sits best within consideration of the third question posed in the consultation: **"Do the nine AI regulation approaches capture all ways in which AI is regulated? Which ones may be missing?"** Our suggestion fits closest with approach 7) – a risk-based approach – and is therefore applicable, we believe, to many different jurisdictions and sectors.

4. The UK has not introduced any AI-specific regulation, though existing laws apply where relevant; see, for example, the written Parliamentary answer from the Government in 2018 confirming that the Health and Safety at Work Act 1974 applies to the deployment of AI in the workplace. The previous Conservative Government proposed a principles-based approach that prioritised innovation, with some small-scale regulatory interventions proposed with regard to Large Language Models (LLMs) – so a mix of approaches 1) and 2) in the UNESCO taxonomy – with some elements of approach 5) (adapting existing laws) in that existing sectoral regulators were to be tasked with setting out approaches to regulate AI within their existing frameworks.

5. Following the election of the Labour government in July 2024, a debate on the best way to regulate AI is back on the agenda in the UK. An AI Bill was promised in the Labour manifesto, along with a ban on the creation of sexually explicit deepfakes, but there was no commitment to bring forward a Bill in the first term within the recent Kings Speech. It referred, instead, to plans to "seek to establish the most appropriate legislation to place requirements on those working to develop the most powerful AI models", though detail as to what might be envisaged is absent.

6. Our recommendations below remain relevant to UK thinking as much as they are – we hope – useful to UNESCO's consideration at a global level.

Risk-based approaches to AI regulation

7. We provide some thoughts below on two risk-based approaches:
   i. Product safety and a duty of care
   ii. The precautionary principle.

Product safety and a duty of care

8. Discussions about AI regulation imply a presumption that AI might require regulation. In our wide-ranging work on regulation of software in social media we have always gone back to fundamentals. A need for regulation implies that there are costs arising from the use of AI in a production decision which do not fall on the company but on wider society – such as workers, customers and third parties. This causes allocative inefficiency or individual or social harm. There are many regulatory mechanisms for regulation of external costs – these should be explored before reaching for new uncertain models.

9. A common and highly successful approach to returning external costs to the production decision is risk-based, proportionate regulation or self-regulation focused on the outcomes of a company process – and could apply irrespective of the type of AI used. In this model the obligation is placed on the developer of the product to understand the product, how it could be used and the risks arising therefrom – and to take steps to mitigate those risks. This basic model has been successfully deployed across a range of contexts, including in relation to quite general forums of risk.

10. A strong and effective example of this model is the UK's Health and Safety at Work Act 1974. A statutory duty of care enforced by a regulator has proven effective and future-proofed. The Act firmly applies to AI in the workplace as we describe below. This approach is flexible and future-proofed, focusing as it does on the outcomes that arise from service design - the company systems and processes, rather than the specifics of the technology that underpins them. One advantage with this approach is that it could

apply as a base-level principle across the use of AI generally but could also be deployed in sector-specific contexts, so as to allow the particular characteristics and risks of those contexts to be taken into account. Indeed, insofar as the UK's Online Safety Act 2023 looks at the operation of filters, algorithmic recommendation tools and automated content moderation, it could be said to be a sector-specific example of AI regulation. Having a common approach potentially allows the interconnection between sector specific rules and general AI rules to occur, and potentially between those developing AI and those deploying them.

11. Our proposal for a statutory duty of care for online harm reduction which has been adopted, in part, by the UK Government in the Online Safety Act, drew on the approach that, for nearly 50 years, has underpinned Health and Safety legislation in the UK. We are of the view that this approach is applicable to AI and its application in many different industrial sectors, albeit that some of those sectors may require additional considerations or refinements to be added to the regulatory framework to take account of their specific risks and potential harms. We also believe that it has international application, being adaptable to many different jurisdictions and legal frameworks. While the EU's AI Act does not fully embrace a risk assessment and mitigation approach (instead pre-determining the risk level of certain categories of AI and AI use), it still requires certain due diligence obligations around risk and testing, especially as regards data governance and certain aspects of safety by design (obligations regarding accuracy, robustness and cybersecurity) as well as a human rights impact assessment.

12. While the EU's AI Act envisages that the deployer of high-risk AI systems must follow instructions and ensure the AI systems have human oversight, it could be possible also to place more general due diligence obligations on the deployer. Indeed, such steps might be envisaged by pre-existing sectoral regulation. For example, the UK's Health and Safety at Work Act 1974 places duties on any person who designs, manufacturers, imports or supplies any article for use at work to ensure that it will be safe and without risks to health. The UK Government has confirmed in a written Parliamentary question that this applies to artificial intelligence and machine learning software. Section 6(1)(b) requires such testing and examination as may be necessary to ensure that any article for use at work is safe and without risks but does not specify specific testing regimes. It is for the designer, manufacturer, importer or supplier to develop tests that are sufficient to demonstrate that their product is safe.

13. It is understandable that policymakers and Parliamentarians, in a rush to provide responses to new or innovative technologies, often overlook existing systems that can be used to make them work safely and well. Extending speculation about new laws rather than complying with existing ones also suits many of the biggest technology companies as it delays and diverts scrutiny of the development of their products in real

time – putting off the day that they are accountable to regulators for the safety of their services and products and, as we have seen with social media, allowing significant harm to be caused to individuals and society in the meantime.

14. It is important to reiterate that AI regimes, especially when in the form of guidance, aspirational standards or best practice, do not displace or downgrade existing legal regimes which apply generally, whether this be in terms of how the models and tools are developed, or deployed. So, in the UK and likely elsewhere, existing data protection regimes apply to AI. (The same is true of data protection as it is for user safety; the UK Information Commissioner has set out clearly how “the underlying data protection questions for even the most complex AI project are much the same as with any new project. Is data being used fairly, lawfully and transparently? Do people understand how their data is being used? How is data being kept secure?”).

15. Copyright has become another fundamental issue, as have questions surrounding bias and the impact of bad data. Beyond this we can see the potential for synthetic media outputs in particular to raise questions for numerous laws, including private law concerns (eg defamation; confidentiality), administrative law (eg misleading advertising) as well as criminal (eg deepfake porn and sextortion; fraud). AI decisions may challenge underlying constitutional principles – notably fair decision-making (and decisions that can be challenged)[1]; these concerns lead on to more fundamental questions about societal values. Moreover, both developers and deployers should be aware of these risks when thinking about the need for safeguards.

The precautionary principle

16. There is a second aspect in a long-established, risk-based approach that has relevance to debates about AI regulation: the precautionary principle, which has its roots in environmental protection but has relevance to scientific developments more generally. It provides a mechanism for dealing with situations where risk of harm is evident but the precise causality is not known, yet waiting for evidence can result in more work both in terms of correcting damage caused in the interim as well as setting the course for the future. The precautionary principle, though much adopted, is somewhat uncertain in its ambit. In some views, the precautionary principle operates to stop development, but it does not require the banning of products. It can rather be used to provide a frame for development, and indeed is closely linked to risk governance approaches.

17. The lack of scientific certainty (or at least consensus) should not be deemed to be a barrier to a decision to take action to prevent the damage. Work undertaken within the UK Government in the 1990s in response to a series of public safety and public health

scares arising from new scientific advancements looked at potential responses from regulators and innovators in such scenarios. It sets out two conditions:

i. that there is good reason to believe that harmful effects may occur to fundamental interests (eg environment, health); and
ii. that the consequences or likelihood of the risk cannot be assessed with sufficient confidence to inform decision making.[2]

18. Note the precautionary principle here does not require action but justifies the policy choice if made, a decision made on balancing the risks and harms in issue – a similar approach to that in the UNESCO World Commission on the Ethics of Scientific Knowledge and Technology. Although prohibiting products or reversing the burden of proof where the precautionary principle applies is not required, these steps both remain as possible policy interventions, especially where there are significant hazards identified. Otherwise risk-based regulation is a central mechanism.

19. Significantly, however, as explained by ILGRA in the UK, the precautionary principle expects that the hazard creator should provide, as a minimum, the information needed for decision-making and that "[d]ecision-making should bring together all relevant social, political, economic, and ethical factors in selecting an appropriate risk management option". With regard to social media regulation, we advised that "companies should embrace the precautionary principle" because it prevented the need for banning particular types of content (especially those that did not trigger specific treatment under the law) and instead took a systemic approach to regulation, founded on risk assessment. The argument that perverse incentives impacting content creation (e.g. revenue sharing schemes forming the basis for clickbait farms; impact of likes and similar metrics on user behaviour) and distribution (e.g prioritising outrage and extreme content) should be removed the system, reducing the problem in the first place as well as ensuring adequate safeguards.

20. There is much to learn in these historic science/risk dilemmas. We were heartened to see that one of the leading UK-based AI thinkers and developers, Demis Hassabis of Deep Mind, had recently argued that "as with any transformative technology we should apply the precautionary principle, and build & deploy it with exceptional care". This is particularly the case given, as the IMF has noted, that "[t]he risk-reward profile of AI is asymmetric; although there are vast benefits to AI's potential, policymakers must guard against its potentially catastrophic downsides", especially given AI's ability to proliferate and spread.

21. The relevance of the precautionary principle to technology is set out in our 2019 exposition on this topic. We quote it in full here for ease of reference:

a. *One of the recurrent arguments put forward for not regulating social media and other online companies is that they are unique or special: a complex, fast-moving area where traditional regulatory approaches will be blunt instruments that stifle innovation and require platform operators to take on the role of police and/or censors. Another is that the technology is so new, sufficient evidence has not yet been gathered to provide a reliable foundation for legislation; where there is a body of evidence of harm, in most cases the best it can do is prove a correlation between social media use and the identified harm, but not causation. We believe that the traditional approach of not regulating innovative technologies needs to be balanced with acting where there is good evidence of harm.*

22. Finally, we include some material here from our UK work which may also be relevant to UNESCO's consideration of this topic as to which body or bodies should provide regulatory oversight for new regulation. In our view, regulatory bodies with oversight of individual industrial sectors should retain the lead in the oversight of how industries and companies within those sectors are using AI. This might involve, for example, scrutiny of the risk assessment and mitigation processes in place for the development and updating of new industrial techniques using AI software; risk assessments of (a) the potential for harm arising from the operation of AI, ML and other software that controls systems and processes and (b) monitoring and performance management of staff. We acknowledge that there may be cross-cutting issues across sectors and that a mechanism may need to be found to ensure coherence between them; it may also be that for the models themselves, and specifically general models, some baseline common principles (eg around bias; sourcing of data; sustainability etc) will be needed. Nonetheless, while each of these domains faces different challenges and pose a range of threats, the same risk-based approach can be taken; responses may however vary.

23. If the widespread deployment of AI in an industrial sector requires more resources and skills for the regulator then, on the "polluter pays" principle (that is, the idea that the person causing the problem should pay for rectification rather than those who suffer as a result), these should be raised from the regulated services or the government if regulation is direct funded. A mechanism for the regulators in different sectors to cooperate with each other, share insights and information and undertake horizon-scanning would also be advisable, as is the case in the digital field more generally.

24. Using HSAW74 as the baseline regulation simplifies the regulatory approach and prevents companies using arguments that AI-driven technology is somehow novel or special in order to avoid scrutiny or oversight. It avoids the need for AI-specific primary legislation and/or multiple sector-specific regulatory frameworks or different compliance requirements. Existing expert regulators in each sector will have the freedom as well as

the authority to ensure that their oversight of industries within their remit can keep up with the pace of technological development, with the onus being firmly on the regulated industries to design in AI risk-assessments and safety testing alongside their existing regulatory compliance duties rather than as an add-on or afterthought.

Conclusion

25. We hope that this response is helpful and look forward to seeing the next stage of UNESCO’s work on this topic. We would be happy to continue to contribute or to speak further to UNESCO officials, if helpful.

Online Safety Act Network

Contact: hello@onlinesafetyact.net

[1] Taeihagh A., Ramesh M. and Howlett M. “Assessing the regulatory challenges of emerging disruptive technologies”, (2021) 15 (4) Regulation and Governance,, 1009-1019. https://doi.org/10.1111/rego.12392

[2] The EU has set 3 principles, but covering similar ground: a) identification of potentially adverse effects; (b) evaluation of the scientific data available; (c) the extent of scientific uncertainty - European Commission (2000) Communication from the Commission on the precautionary principle, COM(2000) 1 final, 2.2.2000. https://op.europa.eu/en/publication-detail/-/publication/21676661-a79f-4153-b984-aeb28f07c80a/language-en