National Engineering Policy Centre

Royal Academy of Engineering

Regulation of artificial intelligence (AI)

Response to the Department for Science and Technology's consultation

Engineering for a sustainable society and inclusive economy

Introduction

Since 2019, the National Engineering Policy Centre (NEPC) has been exploring the safety and ethics of autonomous systems to understand the risks and benefits associated with this technology across different sectors. The project seeks to understand how autonomous systems can be ethically designed, developed and deployed to ensure benefits are widely distributed and no one is disadvantaged. We have undertaken a series of sector specific deep dives to understand the opportunities and challenges within different sectors such as transport and healthcare. The below comments are drawn from this work and the viewpoints of the expert working group that steers it. It is important to point out that this work on autonomous systems is distinct from the explicit subject of artificial intelligence (AI), but they are closely related. This work therefore has important and insightful implications for the pro-innovation framework to regulation that is being proposed in the government's AI white paper since the majority of autonomous systems make use of AI, and this perspective is being addressed in the work.

This response also references an NEPC roundtable led by the BCS (British Computer Society) on June 5th entitled 'NEPC round table with the Office for AI: AI Safety and Risk'. The roundtable was arranged to feed into this consultation and was attended by a number of experts with deep knowledge of a variety of different sectors who shared views on the AI white paper, and regulatory and non-regulatory mechanisms which could help to ensure safety while supporting innovation in the development and deployment of AI.

Lastly, it should be noted that this response is based on the following distinctions between automation, autonomy, AI and machine learning (ML):

Automation, autonomy and AI are closely linked concepts that relate to a range of different technologies. In sectors such as transport, the word autonomy has been associated with the embodied technological systems, such as autonomous vehicles which navigate or take action independently. However, these terms also relate to technologies that can underpin either fully autonomous decision-making, or automated systems that provide advice to human experts.

AI is the broader set of technologies that, to some degree, mimic human intelligence or decision-making, with ML referring to models and systems that learn independently using datasets, and that modify their operation on the basis of such learning. Not all AI-enabled and ML systems are autonomous (i.e., functioning independently of human operators). However most autonomous systems rely on these techniques to some extent. Some systems continuously learn from new data; others are locked, working on the basis of fixed algorithms derived from programming or learning. This is significant in terms of whether systems are deterministic – meaning that the same input will always yield the same output, or whether they may behave in unexpected ways. Non-deterministic AI presents significant challenges for validation and verification.

About the National Engineering Policy Centre

We are a unified voice for 43 professional engineering organisations, representing 450,000 engineers, a partnership led by the Royal Academy of Engineering. We give policymakers a single route to advice from across the engineering profession. We inform and respond to policy issues of national importance, for the benefit of society.

About the Royal Academy of Engineering

The Royal Academy of Engineering is harnessing the power of engineering to build a sustainable society and an inclusive economy that works for everyone. In collaboration with our Fellows and partners, we're growing talent and developing skills for the future, driving innovation and building global partnerships, and influencing policy and engaging the public. Together we're working to tackle the greatest challenges of our age.

1. Do you agree that requiring organisations to make it clear when they are using AI would improve transparency?

There would be some merit in doing this, e.g. if the output from a large language model were suitably and reliably "watermarked" then individuals, e.g. clinicians, or organisations, e.g. regulators, could exercise an appropriate degree of caution. It is also important to flag when an algorithm is used to make a decision, rather than it being AI per se. Sometimes the main issue is the fact that decision making is computer-driven rather than AI-driven, and so restricting this requirement to AI would be unhelpful.

However, if, for example, AI is used in perception systems in a self-driving vehicle then what value would this be? What difference would it make to the user of the vehicle (the regulator will already have taken that into account in assessing the vehicle)? Further, the "labelling" would only make clear the presence of AI, not how it might fail (what harms it might contribute to). Thus, this is desirable, especially in cases where outcomes of decision making could be challenged or re-assessed, but far from sufficient to ensure trustworthiness of AI, especially in life-critical applications.

2. Are there other measures we could require of organisations to improve AI transparency?

Our work on autonomous systems argues that transparency is critical when it comes to building confidence and trust in AI and autonomous systems with users and the wider public (noting again that AI and autonomous systems are not one and the same).[1] Our approach to transparency assumes that the basis of a particular autonomous decision or action should always be discoverable. Therefore, understanding what a system is doing, why decisions are made, and what went wrong once autonomous systems fail are all absolutely crucial when it comes to increasing transparency and building trust. Requiring organisations to disclose use of AI goes some way in improving transparency as a principle by helping to establish a culture and practice of openness that would allow for public scrutiny and the prevention of misuse. However, in order for that disclosure to have any meaningful value, organisations would need to provide evidence of a system's reliability through verifying that a whole system meets its design specification, meaning that the basis for a system's action is discoverable.

There is, however, the challenge of assuring stakeholders of transparency when transparency often means something different to different stakeholders, who will require different information dependent on their level of knowledge. Therefore, having a standard definition that covers the breadth of transparency for expert stakeholders (safety certification engineers, accident investigators, lawyers or expert witnesses) as well as non-expert stakeholders (users, wider society) would be essential in enabling such standards to have a substantive impact. The National Engineering Policy Centre (NEPC) previously identified the Institute of Electrical and Electronics Engineers standard (IEEE P7001-2021 *transparency of autonomous systems*[2]) as an example of this,[3] and we are pleased to see that the white paper has incorporated this into its cross-sector principles.

This said, it is important to consider that AI systems can be designed to be non-deterministic. As a result, evidence cannot always be provided for how the system will function in every possible outcome. Related to this is the difficulty of assigning moral responsibility when harm occurs within an AI-enabled system where there are complex overlaps between human and machine. The various social, behavioural, cultural, and organisational issues that can come into play in such cases requires an interdisciplinary approach. To help address this, the UKRI-funded project *Assuring Responsibility for Trusted Autonomous Systems (AR-TAS)* aims to develop an interdisciplinary methodology to trace and allocate responsibility of the decisions

[1] https://kpmg.com/au/en/home/insights/2021/03/artificial-intelligence-five-country-study.html
[2] https://standards.ieee.org/ieee/7001/6929/
[3] https://nepc.raeng.org.uk/media/2hsh552k/autonomous-systems-workshop-report.pdf

and outcomes of autonomous systems.[4] These findings would be important to consider when outlining standards of transparency, as part of considering which standards should apply to computer systems more broadly and those which should apply to AI-enabled systems more specifically.

Safety Cases

At an NEPC roundtable on safety considerations for AI, led by the BCS (British Computer Society) and attended by the Office for AI, contributors noted that the publication of 'safety cases', and making these cases publicly available, could deliver improved transparency to the benefit of the public, developers and users alike.[5]

'Safety cases', in this context, refer to written demonstrations of the hazards presented by the application of AI in a specific context, as well as the level of risk associated with those hazards and how those risks are being mitigated. These 'safety cases' could serve as an exemplar to developers and users – and would also provide a level of transparency that could serve to engender greater public trust.

'Safety cases', as such, would not only be a means to improve the transparency of AI-enabled systems. They would also serve to compliment the effective application of the cross-sectoral principles outlined in the white paper. Whilst work on safety cases for AI is still an evolving subject, work at the Assuring Autonomy International Programme (AAIP) funded by the Lloyd's Register Foundation has published guidance on safety cases for machine learning (the typical form of AI used in critical applications) which is already influencing standards and regulation.[6]

Making safety cases publicly available, however, may conflict with the interest of private companies developing AI models and AI-enabled systems. Given that there is a high level of competition between companies who are developing and commercialising AI technologies, there is a strong incentive for these companies to protect their intellectual property (IP), which may limit the detail they would feel comfortable providing in publicly published 'safety cases'. It was therefore noted by one contributor to the NEPC roundtable that public bodies should provide leadership on this issue.[7]

5. Do you agree that, when implemented effectively, the revised cross-sectoral principles will cover the risks posed by AI technologies?

Whilst we find the revised cross-sectoral principles to be very helpful in covering the risks posed by AI technologies, there are a number ways in which some of them could be expanded. A recent NEPC roundtable discussion, led by the BCS (British Computer Society), entitled *How can AI Principles deliver for different communities and groups?* Explored ways in which AI-enabled systems could be improved to mitigate harm or prevent further marginalisation against disadvantaged and vulnerable people such as minority groups, those with disabilities and young people. Relevant findings are outlined below.

Safety, Security and Robustness

Participants argued that there is a pressing need to safeguard young people as malicious actors could explore AI technologies to inflict harm. Due regard would therefore need to be in place to protect vulnerable users in particular, and the proper processes and skills in place to enable developers and regulators identify and differentiate between real and fake threats.

---

[4] https://www.cs.york.ac.uk/research/trusted-autonomous-systems/
[5] "Office for AI: AI Safety and Risk" (NEPC Roundtable, Online, June 5, 2023).
[6] https://www.york.ac.uk/assuring-autonomy/ and https://www.assuringautonomy.com
[7] Ibid.

Empowering stakeholders such as children and organisations like Citizens Advice Bureau (CAB) can play a significant role in preventing harm.

We note, however, that the term 'AI safety' should not be used just to address on-line harms and should include the potential for physical harm and fatalities. As noted above, this is where safety cases have a role to play in demonstrating that the risks to life from AI-based systems are adequately controlled.

Fairness

The workshop participants argued that AI-enabled systems need to act as a counter-system against marginalisation of disadvantaged and vulnerable groups, driving progress and actively curating fairness. A "good work charter" can serve as a framework for defining what "good" looks like in the context of work, but AI principles must go beyond their current scope to explicitly outline how we can achieve positive outcomes. Achieving this requires a collaborative effort through co-creation, where disadvantaged individuals are actively involved in shaping the technology that will help make their lives easier. Empowering such communities to co-create, where possible[8], giving them the agency to challenge and push back against negative impacts, and establishing a feedback loop to ensure their perspectives are considered, is therefore essential.

Discussion at the workshop explored the idea that erosion of trust in Government measures was wider within underrepresented communities, which dropped significantly during the COVID-19 pandemic.[9] To rebuild trust, regulatory measures should include a sufficient representation of individuals from diverse backgrounds, including minority and marginalised communities. Where this has not happened – such as in clinical trials that often exclude these communities – the result is the production of technologies that do not adequately serve their needs. Therefore, setting out guidelines to ensure inclusive data collection is necessary.

Another barrier to AI empowerment is with literacy and understanding. Enabling people to make informed decisions regarding their engagement with AI technology is essential. However, literacy poses a challenge since the labelled data used to develop AI models and the parameters of AI-enabled systems often marginalises certain groups. It is therefore imperative to make the language more accessible and inclusive for all communities, ensuring a baseline level of adequate understanding.

It is also important to consider that the scaling up the application of a single machine learning model will inevitably reduce its effectiveness for individuals. Vital to mitigating this is careful monitoring and documentation of algorithmic unfairness and exploring legal and statistical measures that can help alleviate disparities through AI.

6. What, if anything, is missing from the revised principles?

In addition to expanding the revised principles, there are several other principles that the NEPC has previously identified that would be important to consider for inclusion.[10] Several key standards for understanding and regulating autonomous systems, such as management systems, failsafe design in baseline AI-enabled system development, the verifiability of autonomous systems, and risk management, could helpfully sit beneath the high-level principles.

[8] This is unlikely to be possible in all cases, e.g. in developing self-driving (autonomous) vehicles.
[9] https://www.lshtm.ac.uk/newsevents/news/2021/high-level-public-support-strict-covid-control-measures-much-lower-level-trust
[10] https://nepc.raeng.org.uk/media/2hsh552k/autonomous-systems-workshop-report.pdf

Management System

The International Organisation for Standardisation and International Electrotechnical Commission (ISO/IEC) 42001 standard on AI Management System (due to be available in late 2023) specifies the requirements and provides guidance for establishing, implementing, maintaining and continually improving an AI management system within the context of an organisation. ISO/IEC 42001 will help the organisation develop or use AI responsibly in pursuing its objectives, and to meet applicable regulatory requirements, obligations related to interested parties and expectations from them. ISO/IEC 42001 will provide a step-change in how organisations approach AI and what they expect from customers/supplier/partners.

Failsafe Design

The Institute of Electrical and Electronics Engineers' (IEEE) P7009 standard for *fail-safe design for autonomous and semi-autonomous systems* establishes a baseline for the development, implementation and use of fail-safe mechanisms in these complex systems. It describes some of the key requirements and properties of these systems and provides tools to implement fail-safe mechanisms and methods to measure and certify the ability to fail safely. The standard informs the design, testing and analysis of the failsafe mechanisms and the organisational safety processes, should a system fail. These mechanisms are essential as autonomous systems can fail, often without a human being on hand to recover, and there is a need to help mitigate risk of harm to people, society or the environment. The intent is that this standard is adapted for different sectors so they can define what is "safe enough" in each specific context. For example, the safety requirements for a self-driving car on a public road may be different to an autonomous robot in a nuclear facility.[11]

Verifiability

The development of IEEE's P2817 *guide for the verification of autonomous systems* will enable users to define an appropriate multistep verification process for autonomous systems, based on the available tools, levels of transparency, and good practice. The guide provides resources on: formal methods to provide strong evidence (mathematical proof) for the systems; simulation to understand the behaviours in specific scenarios; stochastic methods for probabilistic estimates of system behaviour; real world testing for higher risk scenarios; and runtime verification to ensures the system remains within predicted boundaries. The guide helps developers avoid common pitfalls in the collection, analysis and inbuilt assumptions underlying the evidence that the integrated system meets the design specification. It focuses on the functionality of, and decision-making processes within, an autonomous system, and not the outcome.[12]

Risk Management

The ISO/IEC 23894 AI standard on risk management, due to be published shortly, will provide guidelines on how organisations that develop, produce, deploy and use AI products, systems and services can manage risk specifically related to AI.

There are also other principles, such as design practice, operational contexts, and human interaction (outside of human factors or machine learning explainability), that would also be worth considering for inclusion.

Clarity on What Constitutes 'Safety'

Contributors to the NEPC roundtable noted that while the safety of an AI-enabled system is contingent on a number of factors specific to the application of AI technologies, greater clarity on the Government's definition of what constitutes 'safety' may support the development of

[11] https://ieeexplore.ieee.org/document/9700402
[12] https://standards.ieee.org/ieee/2817/7644/

effective regulation from individual regulators. See also the note above on different interpretations of the term 'AI safety'.

A more robust description of 'safety' could provide information of cross-cutting hazards that should be given special consideration, or the level to which regulators should seek to mitigate certain risks. Contributors noted, however, that such a description would need to clarify whether it was referring to the safety of AI technologies themselves or AI-enabled systems.

Human Control

Whilst not addressed at the NEPC roundtable, there is a strong argument that there should be meaningful human control related to the implementation of AI. This might simply be the ability to refuse the use of an AI-based decision-aid in a particular circumstance, or the ability to receive an explanation before agreeing on the implementation of a recommendation from such a decision-aid. This was certainly highlighted as a need in terms of the use of AI-enabled systems in healthcare.

7. Do you agree that introducing a statutory duty on regulators to have due regard to the principles would clarify and strengthen regulators' mandates to implement our principles while retaining a flexible approach to implementation?

One of our Fellows noted that based on their discussions with a wide range of regulators, especially the HSE, there is a clear need for clarity of guidance on how to deal with AI and AI-enabled systems. It is likely, therefore, that the regulators will adopt or be influenced by these principles anyway. The introduction of a statutory duty therefore may not make much difference in practice. We also note that duties around principles will be hard to realise without associated guidance, so there would be a need for regulators to produce domain-specific guidance to implement these principles.

8. Is there an alternative statutory intervention that would be more effective?

The potential introduction of a statutory duty with a flexible approach to implementation could help strengthen the implementation of the principles outlined in the AI white paper. However, there remain several other barriers beyond statutory requirements that would require addressing if the implementation of the principles by regulators is to be effective. Regulators currently face several challenges that could impede the enforcement of the white paper's revised principles.

The Speed of Technological Innovation

The speed of technological progress makes it difficult to ensure that regulators have enough understanding to be able to competently assess compliance. As with previous technology evolutions, competing with industry for a limited pool of skilled professionals will affect regulatory capacity.[13] The increasing convergence of technological innovations across different sectors are blurring the lines between regulatory systems and between sectors. The result is a mismatch between innovations and our regulatory systems, which can slow down new products being brought to market, compromise safety when they get there, or introduce inconsistencies that weaken the clarity and strength of the implementation of the principles.[14]

Capacity Limitations

Many regulators are now trying to take an enabling approach, with early engagement to try and understand the issues, agree on solutions with the industry and enable joint research. However, regulators and developers still lack the resources and time to learn about standards, which means they do not necessarily know what standards already exist or how best to apply

[13] https://nepc.raeng.org.uk/media/kzujkic2/nepc-the-journey-to-an-autonomous-transport-system.pdf
[14] Ibid.

them. With that lack of understanding, it was felt that both confidence and trust to implement and rely on a collection of technical standards was also missing. Additionally, regulators lack understanding of AI, Machine Learning (ML) and autonomous systems and are unable to keep up with technological developments. Therefore, there is a need for CPD courses that help regulators to better understand AI, ML and autonomous systems and existing and emerging standards and how to adopt them. There is potential for a body such as the Institute for Regulators to collaborate with members of the National Engineering Policy Centre on designing and developing CPD courses that help regulators to better understand AI, ML and autonomous systems and existing and emerging standards, and how to adopt them.[15] [16]

Consistency Across Sectors

It will also be important to consider how language and approaches across sectors differ. Language used in standards poses a challenge, as it is either inconsistent or too complex, resulting in difficulty of, or varying, interpretation of standards. Ideally, language across standards should be made consistent to make it easier for users to effectively understand and interpret between standards produced by different bodies, this may require standardised terminology and collaboration to build unified understanding. There is a potential role here for the emerging Institute for Regulators.[17]

9. Do you agree that the functions outlined in Box 3.1 would benefit our AI regulation framework if delivered centrally?

There is merit in establishing frameworks and principles centrally but given the breadth of applications of AI and AI-enabled systems there will still be a need for individual regulators to undertake substantial amounts of work. This is likely to have to include guidance on risk acceptance and the forms of evidence (including safety cases) need to support decisions on whether or not to approve systems for use. Thus, centralised functions are likely to be helpful but they will not be sufficient.

10. What, if anything, is missing from the central functions?

Accreditation

Government may wish to engage with academic institutions, accrediting institutions and the private sector to explore opportunities to develop accreditations for both the development and use of AI models and AI-enabled systems. Contributors to the NEPC roundtable noted that this would help to ensure competency in both development and deployment.

It was noted that the increasing availability of tools to support development was currently serving to reduce the 'barriers to entry' for developers– and such a trend could impact the quality of models and the systems in which they are applied. Given the cross-sectoral applications of these systems, it would be important for these accreditations to also assess and reflect the life cycle of models and systems.

11. Do you know of any existing organisations who should deliver one or more of our proposed central functions?

Similar to the AI white paper, we have also observed the demand by developers for a central guide to help navigate UK regulatory requirements. The functions outlined in Box 3.1 do provide us with confidence that these would benefit AI regulation. Additionally, having something like the AI Multi-Agency Advisory Service (AI MAAS) which is being established

---

[15] https://nepc.raeng.org.uk/media/nqnhktgq/nepc-safety-and-ethics-of-autonomous-systems.pdf
[16] We understand that the AAIP (https://www.york.ac.uk/assuring-autonomy/) has already run SPD course for regulators including the MCA, MHRA and VCA.
[17] https://nepc.raeng.org.uk/media/2hsh552k/autonomous-systems-workshop-report.pdf

through the NHS Transformation Directorate and AI Lab regulatory programme would be welcome in executing these duties.

Government may also wish to engage local authorities in these discussions, as end uses may be impacted by local policies. For example, Integrated Care Systems (ICSs) may wish to use AI-enabled systems to support the delivery of their specific health equity goals – and as such they would likely be engaged with the delivery of the functions described in Box 3.1. – particularly 'horizon scanning'.

12. Are there additional activities that would help businesses confidently innovate and use AI technologies?

Yes, see below.

12.1. If so, should these activities be delivered by government, regulators or a different organisation?

The NEPC's workshop report on regulation and cross-cutting standards for autonomous systems within AI discusses several key standards for regulating and understanding autonomous systems – transparency, management system, failsafe design, verifiability, and risk management (see responses to questions 1 and 6) – all of which would be important to consider here. The workshop also identified a set of recommendations to enable the development of regulation through the adoption of standards which could therefore help encourage innovation.[18] These are outlined below.

Cross-Sector Community Collaboration

The Better Regulation Executive should work with The UK Regulator Network to encourage greater cross-sector collaboration on AI, ML, and autonomous systems, to build a community to understand and tackle common challenges and opportunities.

Regulator Upskilling

There is a need for continuous professional development (CPD) courses that help regulators to better understand AI, ML and autonomous systems, and to be aware of existing and emerging standards and how to adopt them. There is potential for the Institute for Regulators to collaborate with members of the National Engineering Policy Centre on designing and developing such courses. Language across standards should also be made consistent to make it easier for users to effectively understand and interpret between standards produced by different bodies.

Principles and New Standards

Standards bodies and regulators, alongside the AI Standards Hub, should work together to identify and develop usable standards beyond transparency, verification and failsafe design. This might include principles such as: design practice, principles of operational context, human interaction and security.

Industry Uptake

Regulators, Professional Engineering Institutions, Catapults and public procurement bodies should promote the adoption of standards that encourage safe and ethical development of autonomous systems.

[18] https://nepc.raeng.org.uk/media/2hsh552k/autonomous-systems-workshop-report.pdf

13. Are there additional activities that would help individuals and consumers confidently use AI technologies?

The NEPC's programme of work on safety and ethics of autonomous systems has found that past examples of technological change suggest that public acceptance of controversial technologies depends on complex sociotechnical factors including the technological expectations, societal and cultural structures, and the role of related industries that may be disrupted.[19] Lessons should be learned from the success and failure of other technologies. Scale demonstrations with a user-centric approach, such as living labs, may provide a way to introduce autonomous systems to those who will live and work alongside them. This would allow bounded experimentation with the technology and gauge the public acceptability and delivered benefits.

Digital decision-making systems that rely on personal data will be deployed to make decisions in high-impact areas such as welfare. To develop systems that are considered trustworthy in these contexts, data security will need to be rethought to extend beyond protection from harm and towards delivering wider benefits. Achieving benefit requires high-quality information to be collected. When relying on population data, it is important to work with communities to understand their responses to autonomous decision-making and enable forms of cooperation and trust to be built up between individuals and the service provider. Collaboration allows the perspectives of both groups to be included when considering who, or what, is being secured in this environment. That means considering whether these systems offer communities security, and with that the ability to live free from fear, so that they can engage and positively contribute to realise the benefits from the system. This also influences resilience, adding social and economic factors to the technical resilience, to allow services to cope with problems within that system.

The NEPC has also undertaken sector specific deep dives to understand the opportunities and challenges within different sectors such as transport and healthcare. Some lessons could be drawn from these to help increase user confidence in AI technologies. With specific regard to the healthcare sector, the NEPC's work has found that different types of users would have different perceptions and therefore require different assurances.[20]

Autonomous Systems in Healthcare

Patients

Patients will be exposed to the largest risk and the greatest benefit from the deployment of autonomous systems in healthcare. They look to clinicians to deliver treatment and care safely, and so clinicians take on and accept a level of moral responsibility. This becomes more complex when automated systems support the clinician in their decision making or performs some of the tasks, which may erode trust between clinician and patient. Patients may also require a level of explainability to feel comfortable and safe with the use of automated systems along their care pathway.[21] There is also a need to better frame automated systems in this sector, particularly with regard to apps and personal services as an opportunity for patients to have more agency in their own health and the decision-making.

Despite these systems having the potential to increase efficiency, patients may also be concerned that there may be fewer clinicians available with an increased use of automated systems in hospitals.[22] However, such systems may address a severe skills shortage and may help to address waiting times.

---

[19] https://nepc.raeng.org.uk/media/nqnhktgq/nepc-safety-and-ethics-of-autonomous-systems.pdf
[20] https://nepc.raeng.org.uk/media/mmfbmnp0/towards_autonomous_systems_healthcare_report.pdf
[21] Although there is some evidence that patients will continue to rely on the clinicians.
[22] *Human factors challenges for the safe use of artificial intelligence in patient care*, Sujan et al. (2019).

General Public

The public tend to need a substantial guarantee beyond doubt, for example, clinical trials in large numbers. They will want to know that automated systems are trustworthy, safe, and ethical. There can also be a baseline belief that machine accuracy and reliability is lower than a human's or a low trustworthiness of AI advice.[23] It is therefore important to understand and address misconceptions, manage expectations, ensure the automated systems that are developed are trustworthy and well regulated.

Healthcare Staff

Some clinicians are hesitant to adopt automated systems because of the fear of being replaced.[24] There should be engagement from both users and non-users of the technology to build knowledge and better perception of the benefits. Clinicians also require a level of explainability to understand and have confidence in a system's decision making and to accept a degree of moral responsibility; there is emerging work in this area[25], but it remains a topic of active research. Clinicians may also want to a guarantee that the technology is well integrated into the hospital care pathways to ensure that the administration of care is uninterrupted, and that it still allows for an appropriate level of human – patient interaction.[26]

Administrators

Hospital managers and administrative staff may want to know that an automated system fits well within the clinical workflow. They will want to ensure that the care system remains uninterrupted with a guarantee that adding the system results in, for instance, shorter hospital stays and lowers costs.

Developers

Developers need to be clear that there is a medical need for the products that they are working on, so they are likely to be adopted and produce positive outcomes for the patients. Developers may benefit from a deeper understanding of the expectations and perceptions of all stakeholders in the healthcare system.

Considerations relating to autonomous systems in transport can be found in our report on *The Journey to an Autonomous Transport System*.[27]

L1. What challenges might arise when regulators apply the principles across different AI applications and systems? How could we address these challenges through our proposed AI regulatory framework?

In the NEPC's work on cross-cutting governance of autonomous systems, there was agreement that high level principles are needed as they capture the key issues arising from autonomous systems. Standards are helpful as they provide practical ways to assess the system and promote consistency. They are also not prescriptive, which allows for the consideration of specific context and encourages conversations about what is safe enough. Autonomous systems create similar ethical challenges (avoidance of harm, fairness, transparency) across sectors. However, different sectors do have different needs depending on the context of how autonomous systems will be developed and deployed. Some cross-cutting principles such as failsafes will be more familiar in safety critical domains like space or nuclear,

---

[23] *Exploring stakeholder attitudes towards AI in clinical practice,* Scott IA et al., BMJ Health & Care Informatics (2021), https://informatics.bmj.com/content/bmjhci/28/1/e100450.full.pdf.
[24] Ibid.
[25] For example, *The role of explainability in assuring safety of machine learning in healthcare*, Jia et al (2022), https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9769937.
[26] *Human factors challenges for the safe use of artificial intelligence in patient care*, Sujan et al. (2019).
[27] https://raeng.org.uk/media/kzujkic2/nepc-the-journey-to-an-autonomous-transport-system.pdf

but not all sectors will have this same starting point. So, while high-level principles are of value, it was agreed that sector-specific standards would be needed in addition.

There is also some potential to develop standards that work across different contexts. For example, as set out in an earlier answer, transparency often means something different to different stakeholders who will require different information, relayed in plain language. The P7001 standard covers transparency for expert stakeholders (safety certification engineers, accident investigators, lawyers or expert witnesses) as well as non-expert stakeholders (users, wider society). Also, the BSI is leading work on standards relating to the use of AI in a number of domains, including healthcare and self-driving vehicles, which can help the UK establish leadership in some key areas.

There is therefore a need to balance cross-sector standards and governance with the specific areas of application that will fall under the remit of different regulators.

20. Do you agree that a pooled team of AI experts would be the most effective way to address capability gaps and help regulators apply the principles?

There is value in cross-sector conversation, as well as collaboration between all parties (regulators; researchers and innovators; professional bodies, institutions and National Academies; standards developers; insurers; the legal profession), to understand and tackle challenges in ensuring the safety and effectiveness of autonomous systems. Often there are only a few individuals with expertise in autonomous systems in each organisation, and such connectivity would help to make best use of these scarce skills. Cross-sector collaboration may also help to navigate international regulatory differences by sharing an understanding of where the overarching principles remain the same which can help build confidence in safety processes and encourage transferable learning. This could be done through a stakeholder mapping and systems approach to help determine who would be the most effective groups to address capability gaps and apply principles. Additionally, end-user engagement in this process would be vital.

Currently there is no mandate for cross-sector collaboration between regulators and it would be useful to encourage this. The Better Regulation Executive should work with the UK Regulator Network to encourage greater cross-sector collaboration on artificial intelligence, machine learning and autonomous systems, to build a community to understand and tackle common challenges.

21. Which non-regulatory tools for trustworthy AI would most help organisations to embed the AI regulation principles into existing business processes?

There are several additional non-regulatory mechanisms that the NEPC has identified in its publication, *Safety and Ethics of Autonomous Systems*,[28] that would be relevant to highlight here in the effort to help embed the AI regulation principles into existing business processes. These include having in place ethical design standards and codes of conduct.

Ethical Design Standards

While adherence to standards is commonplace for safety compliance, autonomous systems create more than just technical problems. As such, the Institute of Electrical and Electronics Engineers (IEEE) has a global initiative to develop a cross-sector Ethically Aligned Design standard for intelligent and autonomous systems. This aims to embed an ethical approach into the way a product is designed. The components of the standard encourage consideration of the ethical risks throughout the design process and development of appropriate measures to ensure transparency and privacy and be aware of system biases.

[28] https://nepc.raeng.org.uk/media/nqnhktgq/nepc-safety-and-ethics-of-autonomous-systems.pdf

This raises a wider question about whether the technological system or the organisations that developed it are responsible for ethical governance. This decision requires either setting expectations for ethical products, or the requirement that companies who design autonomous systems have an ethical approach inbuilt to their wider company structure.

Codes of Conduct

Codes of Practice, or Conduct, are not legally binding in themselves but they can point to related legislation and provide guidance for using autonomous systems and encourage responsible behaviours. For both maritime and vehicles, Codes of Practice have been developed as a flexible way to ensure safety when trialling this technology while the future regulation and legislation develops. These codes can build trust and push a culture change within the profession.

Recognising the Limitations of Technical Standards

Whilst we recognise technical standards to be an important non-regulatory tool to support regulatory systems - encouraging the use of good practice and defining the conditions that systems must be tested under as outlined in the White Paper (part 4) – it is important to point out that these standards are known to have limitations. Firstly, they tend to be set by the incumbents because standards' committees are often populated by those who can afford to attend. Secondly, irrespective of the standard, there will always be those who try to manipulate it to their own advantage. It is important that standards bodies work together to develop global standards, as is already the case, and the UK should play an active role in setting the IEEE and other international standards.

Other Factors to Consider

There are a range of existing frameworks and tools to support responsible innovation. Alongside use of such frameworks and tools, product teams will need to move towards metrics that understand the change that technology is creating in the world so that informed goals can be set. There are limits to the use of these optional nonregulatory mechanisms. Too much reliance on industry generated codes of practice or standards may create risk for the system's users. We need to define the extent to which industry should be responsible for setting the bar for public trust.

Contributors to the NEPC roundtable on safety in AI also noted that the development of guidance for DevSecOps (i.e., the practice of integrating security testing at every stage of the software development process), or an equivalent specific to the development of AI models and systems, may improve the ability of organisations to embed the cross-sectoral principles identified in the AI white paper.[29]

22. Do you have any other thoughts on our overall approach? Please include any missed opportunities, flaws, and gaps in our framework.

Clarifying between AI Models and AI-Enabled Systems

There is a need for clarity on the distinction between what constitutes an AI model and an AI-enabled system to enable regulators, standards bodies and other stakeholders to effectively respond to the cross-sectoral principles. This includes setting out when and why AI requires different approaches to computer systems more broadly.

---

[29] "NEPC round table with the Office for AI: AI Safety and Risk" (NEPC Roundtable, Online, June 5, 2023), https://aws.amazon.com/what-is/devsecops/#:~:text=DevSecOps%20is%20the%20practice%20of,is%20both%20efficient%20and%20secure.

An AI-enabled system can be defined as referring to an AI model in conjunction with its deployment parameters, interface design and (often but not always) the input of a human decision maker.[30] For the purposes of ensuring that the cross-sectoral principles outlined in the AI white paper can be readily applied by regulators to use-specific contexts (particularly the principles of accountability, safety and fairness) it is important that a distinction between model and system be made. This is because the metrics by which the effectiveness of a model are evaluated should be different from those by which a system is judged, particularly where the system can pose a threat to life, as in maritime systems or self-driving vehicles.

While an AI model can be assessed as functioning well using metrics such as accuracy or mean squared error, an AI-enabled system must be evaluated differently.[31] If an AI-enabled system is utilising a high-functioning model but is presenting information in a manner that is not clear to its end user or it is not trained on sufficiently inclusive data, then the system will not perform well – regardless of the strength of the model (or models) it is utilising.[32] Further, metrics such as mean squared error are not appropriate at system level for, say, self-driving vehicles or autonomous trains – here the metrics will be in terms of accidents and data on those killed or seriously injured, analysed to assess the causal role of the AI elements.

While the white paper proposes focusing on those AI-enabled systems which have the capacity to work adaptively and autonomously, it is important to note that there are models that support systems commonly referred to as AI which are deterministic. However, the focus of NEPC work has been on autonomous systems, as we agree that adaptivity and autonomy pose particular challenges that require good governance and clear evidence of safety in order to be implemented. It is essential that any regulatory system is clear on the distinctions between AI, automated systems and autonomous systems, and their relative risks.

We agree that innovation and regulation should go hand in hand. Regulation is essential for creating safeguards to protect society; clarity for developers; and can build trustworthiness. AI and autonomy are developing rapidly – ensuring good regulation, that can adapt at pace, will be central to unlocking its benefits and guarding against risks.

Clarifying the UK's Role in Developing International Standards

The UK's role in defining and implementing standards for the development and deployment of AI models and AI-enabled systems must be further clarified. The fast-paced and transnational nature of model and system development essentially precludes the UK from being able to develop its own unique standards. However, as a sovereign nation, the UK certainly has the ability to develop a tailored approach to how these standards will be applied.

The white paper, as such, must make it clear that while the UK certainly has a role to play in the development of standards, that role involves collaborating with other governments to develop a sensible, safe and pro-innovation approach. To eschew such an approach in favour of developing entirely unique standards would cut the UK off from the rest of the world, and forestall the development and deployment of cutting-edge models and systems in the UK. The white paper, accordingly, should clarify its purpose of enabling and accelerating existing collaborations between UK-based and international standards bodies.

The white paper, however, should also make it clear that the UK can take a tailored approach to the application of standards within the UK. Devising approaches to the application of

---

[30] https://www.microsoft.com/en-us/research/blog/ai-models-vs-ai-systems-understanding-units-of-performance-assessment/#:~:text=For%20example%2C%20a%20radiology%20scan,clinical%20decision%20and%20treatment%20plan

[31] https://neptune.ai/blog/performance-metrics-in-machine-learning-complete-guide

[32] https://www.microsoft.com/en-us/research/blog/assessing-ai-system-performance-thinking-beyond-models-to-deployment-contexts/

internationally developed standards to respond to the specific needs of the UK's people, institutions and economy should be a point of priority for all stakeholders. As such, the white paper should more clearly outline how UK-based institutions can contribute to priority-setting, evaluation and other processes that are integral to the effective application of standards.