NATIONAL ENGINEERING POLICY CENTRE

ROYAL ACADEMY OF ENGINEERING

Safety and ethics of autonomous systems

Project overview

June 2020

Executive Summary

Autonomous systems **make informed decisions for themselves in complex environments**. As they become increasingly used in all aspects of our daily lives, new questions will arise about: the role of the public, engineers and regulators in ensuring safe and ethical deployment; what we expect of them; and the conditions under which we can and should trust them.

The Future of Mobility Urban Strategy or the Robots for a Safer World Industrial Strategy Challenge Fund are live examples of how the UK is investing in the research and development of autonomous systems. These have **potential to deliver wide economic and social benefits**, including substantial export and productivity gains, and the removal of dull, dirty and dangerous tasks. Realising these benefits without causing direct or indirect harm requires a coordinated effort that reaches across sociotechnical systems.

The National Engineering Policy Centre (NEPC) is exploring opportunities that arise across different sectors through a new project on *Safety and ethics of autonomous systems*. This will examine how autonomous systems should **be ethically and safely designed, developed and deployed to ensure benefits are widely distributed and no one is disadvantaged**.

To explore these themes the Royal Academy of Engineering, alongside NEPC partners, hosted an event on the safety and ethics of autonomous systems. The aim of this workshop was to better understand the **challenges and opportunities that autonomous systems present and how the professions represented can ensure responsible innovation**.

The event findings that follow will be explored and tested through deep dives in specific sectors, including transport and healthcare. These will consider: **what is unique about how autonomous systems are developing in each sector; the specific challenges to safe and ethical deployment; and identification of emerging good practice**. The NEPC has sought input from different

bcs The Chartered Institute for IT · Engineering Council · IAgrE · IET The Institution of Engineering and Technology · Institute of Marine Engineering, Science & Technology · Institution of Mechanical Engineers · Royal Aeronautical Society

audiences at the Driverless Science Museum Late and the international perspectives at the 2019 Global Grand Challenges Summit. The NEPC will publish the proceedings and insights as they emerge.

This document summarises the workshop discussions and presents challenges that arise from ensuring safe and ethical deployment of autonomous systems. It also emphasises some critical trade-offs.

There are **many challenges that need to be overcome in order to make informed decisions about these trade-offs and to minimise harm**. For the purposes of this project, the challenges are categorised into technical, ethical, professional responsibility, regulatory, public acceptability and oversight (see Figure 1). They are considered in more detail throughout the document.

Trade-offs

There are a series of trade-off decisions to be made:

- **Safety** is a critical factor but must be considered along with cost and performance.
- **Ethical** trade-offs between degree of transparency, the protection of intellectual property and the accuracy of the system.
- **Social** trade-offs in how and where these systems are deployed.

Many of these trade-offs are built into the way the system is designed. Deciding where the appropriate balance lies should take into account public perceptions, requiring a conversation beyond engineers and regulators.

Figure 1. Visual summary of the event discussion

Introduction

We are surrounded by automated systems in our day to day lives from the closing train doors to voice activated virtual assistants and building management systems that adjust to maintain a constant temperature regardless of the weather. Technology capability is advancing and new systems are requiring less human intervention and moving from automated to autonomous, making informed decisions for themselves.

These autonomous systems ask new questions of the public, engineers and regulators. These include what we expect of them and the conditions under which we can and should trust them. For the UK to become a leader in the development and deployment of autonomous systems, it needs to develop a shared answer to these questions to realise the economic and social benefits, including substantial export and productivity gains.

To explore these themes the Royal Academy of Engineering, alongside NEPC partners, hosted an event on the safety and ethics of autonomous systems. It considered the new challenges posed by autonomous systems and whether a regulatory step-change is required.

The day brought together cross-sector expertise from industry, academia, regulators and government through a series of panel discussions curated around:

- The risks and benefits of autonomous systems.
- The regulatory step-change required to mitigate such risks.
- The range of non-regulatory mechanisms used to support development of autonomous systems.

This event summary includes a definition of autonomous systems provided for attendees. It also captures the highlights arising from the panel discussions and wider audience Q&As. which start to answer the following questions:

- What benefits do autonomous systems bring?
- What are some of the challenges?
- Where does ethical risk emerge?
- How can public acceptance be built?
- What are the trade-offs?
- What are the challenges to assuring safety?
- What are the non-regulatory support mechanisms?
- What is the role for regulation?
- How is the regulatory system planning to adapt?

These insights present a snapshot of many of the issues associated with autonomous systems. This summary will be used to support future discussions and inform the next stages of the project.

Definition

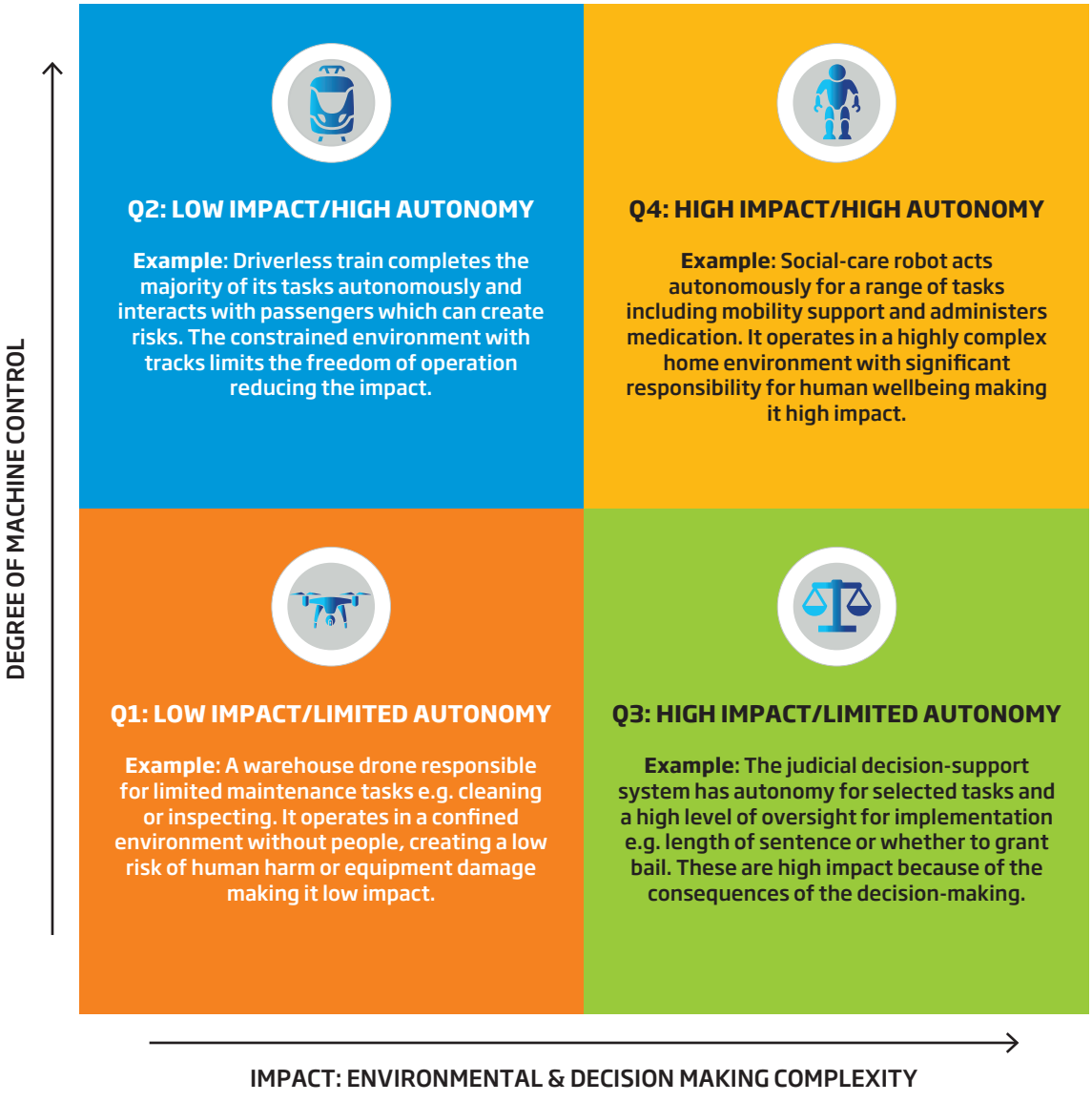
For this event, the following definition of autonomous systems was presented alongside Figure 2:

> Autonomous systems **make informed decisions for themselves in complex environments**. The systems can have a physical manifestation such as a vehicle or a robot or be purely software based, such as a financial trading algorithm or a medical decision-support system.

These differ from automated systems such as a lift or an automated braking system, which carry out fixed functions in simple environments without the intervention of an operator.

Further down the scale, and also excluded from the project, are controlled and supervised systems (such as a remote-controlled drone or a programmed lathe) that have their decisions made for them by an operator.

Figure 2. The spectrum of autonomous systems within the scope of this project with some illustrated examples. The machine control represents the increasing autonomy and the impact considers the risk of potential harm based on the complexity of the operational environment and decision making context. The degree of human collaboration will vary across the quadrants.



What benefits do autonomous systems bring?

A major transition will be required to meet the future aspirations for autonomous systems. However, **the requirement for the development of an autonomous system should be clear from the outset, rather than deploying a technology just because it exists**.

To ensure that widespread benefits are realised, the extent that an autonomous system will add value relative to the existing approach with human oversight must be clearly identified and articulated.

The UK will have opportunities to derive economic benefit from developing autonomous systems through increased productivity and trade. An international marketplace will be required and it will be important to have **international agreement for global standards and regulation to support decision-making and risk management**.

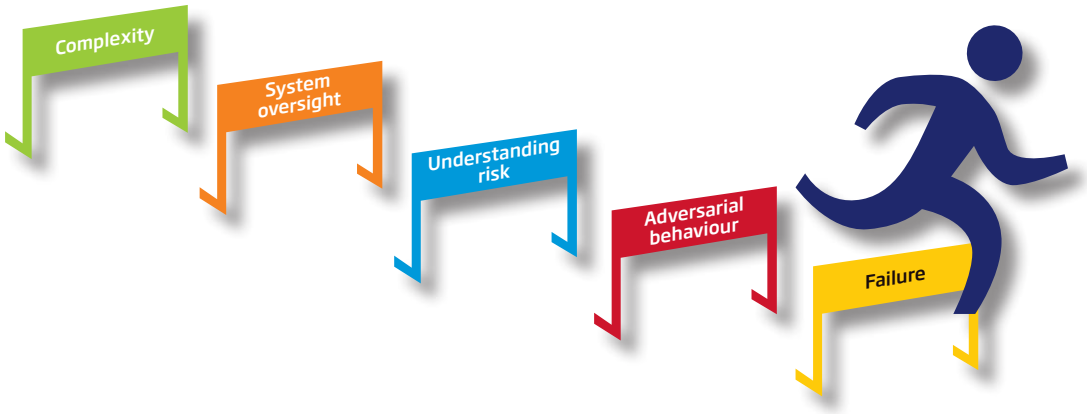
Many of the wider benefits vary depending on the application but they are primarily focused on the **safety improvements derived from deploying autonomous systems for dull, dirty and dangerous tasks**. The following are additional examples of sector-specific opportunities (Table 1).

Table 1. Examples of sector-specific opportunities

Sector	Opportunity
Nuclear	Allowing humans to be removed from hazardous environments will create significant benefits to safety, security, improving control, protection and monitoring of plant operations. However, these must be balanced with potential external risks resulting from reliance on an autonomous system.
Medicine	Medical error is the third most common cause of death. Autonomous systems used in drug delivery or diagnostics may allow accuracy improvements, reducing the prevalence of human errors.
Shipping	With marine autonomous surface ships, remote operators with system oversight could be located shore-side rather than having crews on ships. During a 24-hour period, this could move between European, American and Asian bases to allow everyone to work within a typical working day, rather than through the night onboard ships or in a single country.
Defence	Autonomous systems will be designed to operate in a range of defence applications and scenarios such as accurate identification and neutralisation of malicious cyber-attacks. There might be an opportunity to explore specific lessons around authority and responsibility, which cross over into civil applications.

What are some of the challenges?

Figure 3. Illustration of some of the challenges relating to autonomous systems



There are several overarching challenges facing autonomous systems beyond safety and ethics (Figure 3). These stem from the complexity of the systems and the subsequent challenges to provide oversight, understand the risk, manage the adversarial behaviour and get to the root cause in instances of failure.

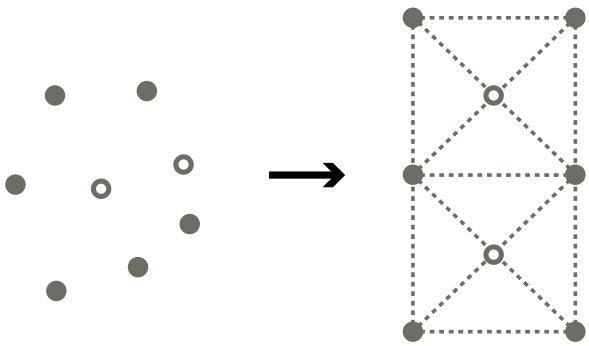
Complexity

The emergent behaviours that result from interactions within the system are the pervasive challenge of any complex system. Understanding and regulating each of the individual components of the system does not ensure the safe performance of the system as a whole and the **unanticipated emergent properties increase the risk to safe operation**.

For complex systems that operate autonomously these emergent properties (Figure 4) can have implications for the system's ethical behaviour and public acceptance. To cope with this, risk management practices will need to consider what could go wrong, and the subsequent consequences, to understand the mechanisms for prevention and mitigation.

Safety assurance practices must be developed as part of the risk mitigation process. **A range of regulatory and non-regulatory measures should be deployed to manage the risk**. However, there are no simple solutions and instead trade-offs will be required.

Figure 4. Illustration of emergence showing the unexpected higher-level properties that can arise from the interaction of components from Boehnert J. (2018) ***The visual representation of complexity.***



System oversight

A system architect has proved beneficial for previous generations of sociotechnical systems, which takes responsibility for the oversight of system design and risk management. This model is in place for automated systems such as the Docklands Light Railway. As autonomous systems are deployed in increasingly complex environments, the number of actors within the system will undoubtedly increase. This could include businesses developing autonomous systems and the enabling technologies, organisations involved in maintaining the environment in which they operate, and the individuals who interact either passively or actively with the system. **For autonomous systems the number of actors that require overseeing is likely to be much broader than previous systems that have taken this approach**. This will create challenges of liability and authority making it difficult to deliver in practice.

Understanding risk

It can be useful for decision-makers to allocate a numerical value to the degree of risk for these scenarios, helping to quantify the trade-offs. However, risk and benefit are not just statistics to be calculated, **risk needs to be understood socially, and managed in its cultural context across the whole lifecycle of the system**. Human factors research suggests that use of autonomous drug administration devices in Intensive Care Units would risk removing an important regular patient engagement point because while administering drugs the nurse is making many other assessments about the patients wellbeing. Who the risk and benefits will affect is an important consideration; any resulting asymmetries will define the relationships that people have with that technology and whether it is considered a success.

Further challenges to assessing risk arise when the autonomous system decision-making is supported by machine learning. These result from the learning systems' unpredictable or emergent behaviours, unforeseeable interactions within a complex environment and the opacity of the systems themselves.

Adversarial behaviour

Individuals may behave in subversive or adversarial ways towards autonomous systems, especially given their inherent facelessness. This was demonstrated with Microsoft's machine learning chatbot Tay that was gamed by Twitter users to post inflammatory and offensive tweets. People look for ways to subvert an impersonal system or potentially one that is representative of an unpopular power structure. This will have to be considered for deployment but t**here will be opportunities to learn from previous disruptive technologies**.

Failure

There are inevitably going to be incidents and failures. Engineers and regulators have a duty to learn from those failures. Aviation is a sector that does this well. With every accident, a detailed investigation is carried out and freely shared internationally so that all can learn from it. This process is currently ongoing following the most recent Boeing 737 Max incident. Such an investigation concentrates on **finding out the cause of the accident rather than attributing blame** and includes legal protection for the witnesses. This culture arose from regulatory pressure and represents a shift to considering the role of system-induced error rather than assuming fault to be human error.

It is important that as autonomous systems become more prevalent, they are accompanied by a **culture of institutional humility in which system failures are learned from openly and fairly**; identifying the cause, the liability and, most importantly, how it could have been prevented.

Where does ethical risk emerge?

While a risk of harm is a problem with many automated systems, autonomous systems add an extra dimension to the problem due to the reduction in human oversight in complex, safety-critical and operational domains.

Moral responsibility

Normally, **human agents who build and use machines have a moral responsibility for any associated consequences**. As the information about any incident will always be incomplete, in practice, responsibility considers aspects of recklessness and negligence. The ability to ascribe responsibility is integral to our ability to trust machines, and to enable interpersonal relations and social practices. **Autonomous systems put pressure on the ability to attribute moral responsibility**.

While human agents are still causally involved in the development of autonomous systems, they may not have robust control or understanding of the system as it encounters new scenarios or learns in practice. Control and understanding are considered necessary conditions of being morally responsible for an action. The system itself is not a genuine moral agent, as it is not capable of acting with reference to right and wrong, so it would be reckless to just transfer moral responsibility to the system. Instead **there may be a need to readjust and revise existing practices of moral responsibility to encompass delegation to autonomous systems**.

Semantic gap

A semantic gap can emerge as a system develops from its intended design to real world specification. This results when the instructions that have been codified into the system do not translate directly to the desired computational behaviours. For some applications certain limits might be built in to avoid certain unethical behaviours but this prompts the question of whose ethics should be used, as principles will vary across individuals and cultures.

Resolving uncertainty

In some scenarios there can be general moral uncertainty about what would be the right thing for the system to do. Rather than relying on the autonomous system's decision, **resolving moral uncertainty is best achieved through collaborative, collective, reflective decision-making. This requires the involvement of many different perspectives and establishment of a reflective equilibrium or coherence among them**.

Understanding behaviour

The difference between intent and reality needs to be understood to enable design teams to choose the right metrics to measure the actual change that an autonomous system is making in the world. This difference can also create challenges when it comes to failure assessment. For example, incorrect control and oversight decisions can be made because the system is not behaving as expected: it can be analysed as doing one thing while actually doing a second, while the operator thinks it is doing a third.

Furthermore, there is the risk of inadequate justification of a system's actions based on its opacity. **Transparency is an important factor when considering whether an autonomous system is ethical**, but often these systems are designed to be a 'black box'. There is ongoing research into the concept of an 'ethical black box', designing systems with a degree of transparency that allows people from outside the system or its development team to learn what happened and why. **Enabling understanding of the decision-making process may increase user confidence in the system**.

How can we build public acceptance?

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. **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**.

Case study one

Public dialogue for self-driving vehicles

From public dialogue on autonomous vehicles, it seems that people are excited by the potential for self-driving technology in mobility.

However, there is wide awareness that self-driving vehicles will be **transformative technologies that place demands on the places in which they operate, changing the environment around them**. This will require behavioural change from other road users and pedestrians.

People are concerned about their own autonomy since they associate car ownership with freedom and worry that some of the perceived benefits won't be equally distributed.

To create wider trust in autonomous systems, it is important that public concerns are heard and accounted for via two-way public dialogue throughout both design and deployment.

Case study two

Clinical engagement for medication management systems

The University of York's Assuring Autonomy programme is currently running a pilot at the Royal Derby Hospital's intensive care unit to trial an autonomous intravenous medication preparation, administration and management system. This project is engaging the clinicians in their willingness to accept the risks associated with autonomous systems to inform the degree of safety assurance required. **Active collaboration between engineers developing healthcare autonomous systems and the users is critical to understanding the needs of the system and those working with it**.

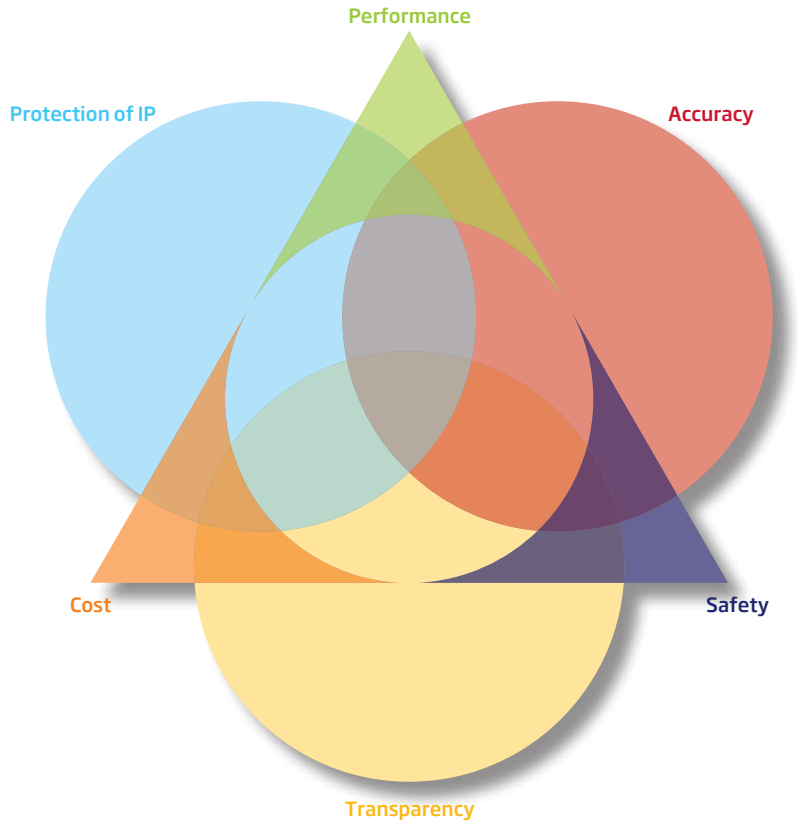
Where are the trade-offs?

Safety is a critical factor for autonomous systems but it must be considered along with cost and performance. As a result, there will be trade-offs between safety and benefit (Figure 5). For autonomous systems much of that trade-off is built into the way the system is designed. However, **deciding where the appropriate balance point lies requires guidance from the public, with a conversation beyond just engineers and regulators**.

There are further potential trade-offs between transparency, the protection of intellectual property and the accuracy of the system. There is a risk that because data is seen as a source of competitive advantage, there are incentives to hold onto that information, which will impede social learning in the rest of the system.

Uncertainties will make it difficult to judge whether the benefits will outweigh the risks with many autonomous systems. How decisions are made to put an autonomous system into operation, or to continue its operation following incidents will involve trade-offs and needs to have appropriate oversight. **The right governance will be needed to make a judgement about whether the benefits should be realised despite uncertainty about the risks**.

Figure 5. Illustration of some of the layers of trade-offs that must be made in the development of autonomous systems. Decisions made on safety, cost, performance, transparency, accuracy and protection of intellectual property are all interconnected.



What are the challenges to assuring safety?

Autonomous systems are developing at pace while good engineering practice regarding them is still evolving. That uncertainty creates challenges for the regulator. As there will be no individual who can be held to account, **the regulator not only has to be concerned with the decisions a system makes, but why that decision was made**.

One of the biggest challenges facing industry is the current gap in assurance models, which provide the justified confidence or certainty in a system's capabilities. Closing this gap will require validating the system safety and verifying it meets its requirements through a variety of mechanisms.

There can be many uncertainties associated with new technology. The people who are going to use the technology must put forward a safety case, verified by a technology demonstration, for a regulator to approve the use of it. **The safety case requires consideration of the systems' function and operational environment with evidence that the risk has been reduced to as low as is reasonably practicable**. Provision of this evidence often informs a legal requirement for operation. Creating a safety case is proving very challenging as a result of:

Validation

Similar to other technologies, **validation of autonomous systems requires a combination of real-world data and trials, alongside the use of simulation**. However, given the complex environments in which they operate, all the possible situations that an autonomous system might experience cannot be foreseen. To combat some of this uncertainty **the range of situations tested should be risk based**. This involves giving much denser coverage in those situations of potentially very high risk, although they may be statistically very unlikely to occur.

Regulation for safety-critical applications typically assumes that you can predict all possibilities. However, this will not be possible for autonomous systems. In unforeseen situations, a system has two options. The first is for the system to stop to allow human intervention, while the second is for the system to make its own decision based on the information available at that point. Human operators' ability to monitor and take over control of autonomous systems when those systems reach the limits of their capability or when problems arise must also be validated.

Verification

System verification typically involves a combination of testing it repeatedly in different ways. There is simulation of the environment it is going to work in, simulation of the system, or mathematical proofs known as formal methods to show that the system matches its requirements. Each of the different ways of doing verification vary in difficulty and provide different levels of confidence or guarantees about what the system will do.

Verification of autonomous systems can be challenging, especially for those systems that learn and adapt based on their environment. This takes the decision-making beyond the initially designed and verified system. **This can be further complicated by the opacity of the process, which means that the software cannot be verified by standard approaches**.

Future methods

Ongoing academic work is establishing methods to carry out mathematical proofs of properties of machine learning systems. These methods will have limits for the high-impact, high autonomy applications. As such, **there will need to be new, transparent approaches to verification of machine learning**. To manage the continued learning from both that individual system and those connected to it may require a move towards operational verification of systems. This would involve identifying where the decision-making occurs and applying focused verification techniques to those elements. **These methodologies are still in**

development. If this is considered to be critical to safety of certain autonomous systems, decisions will need to be made on the deployment timeline.

If we analyse the decision-making process in a formal way, exposing the system intentions, it may be possible to ensure that the system never tries to be unsafe, intends to hurt someone or does anything that is considered harmful.

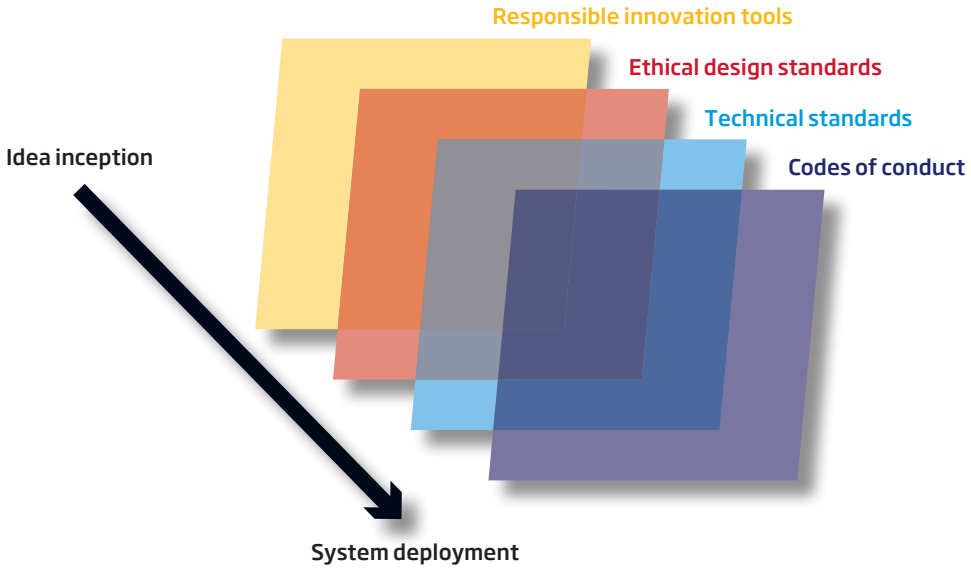
Building systems in the right way with a modular system, with proper architectures and transparency, allows more robust validation and verification and creates more trustworthy systems.

There is a further need for a safety process for autonomous systems that accurately addresses relevant failure causes, including those arising from machine learning, sensor limitations, problems of human system interaction, and other issues. **These safety processes may need to be quite different to traditional ones**. As decision-making is moved to the machine, safety processes that start with the decisions and possible unsafe decision errors in an operational context will be needed.

As these assurance processes develop, there are additional questions about **how safe and resilient these new systems need to be before they are trialled and then deployed at scale**. There are wider considerations within this about what the system is intended to do or what its unplanned effects may be. **During the higher risk test phases, the developers and regulators must ensure that the burden placed upon the people at risk in that testing regime is fair**.

What non-regulatory mechanisms can support this?

Figure 6. Layers of non-regulatory support mechanisms that should be adhered to where relevant throughout the lifetime of a system.



To fully understand the potential sociotechnical impact of autonomous systems, there is a need to anticipate the regulatory needs, employ levers other than regulation and collaborate with others. Non-regulatory mechanisms to support the development of autonomous systems include technical standards, ethical design standards, codes of conduct and responsible innovation tools (Figure 6).

Technical standards

Current **regulatory systems are supported by technical engineering standards that encourage the use of good practice** and define the conditions that systems must be tested under. While an important component of the regulatory system, 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. Finally, many organisations are creating their own standards, which can result in too many standards to be meaningful. How the unhelpful proliferation of these standards can be limited needs to be considered.

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 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.

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.

Responsible innovation tools

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 non-regulatory 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**.

Case study three.

Methodology for consequence scanning

Doteveryone, a responsible technology think tank has developed a methodology that adjunct teams can use to map and understand the intended, and unintended, consequences of the technology that innovators are creating. These TechTransformed resources provide a practical prompts to consider how the innovation could impact different types of user, communities and the environment so these risks can be managed. Such tools embed ethical thinking and collaborative working into the design process, which will be important for autonomous systems and a range of digital technologies

What is the role for regulation?

While standards and Codes of Practice are helpful components of developing responsible technologies, they only really set a minimum quality, outlining what's needed. **Regulations and regulators that are outcome-focused, risk-informed and vulnerability-aware** are required to manage the risks of autonomous systems.

It is often believed that there is an inevitable tension between regulating markets and encouraging innovation, but this is not necessarily the case. **Regulation can stimulate innovation**. FinTech is a good example of thriving innovation in a very constrained environment, both through regulation and industry standards that must be adhered to. However, disruptive innovation also needs disruptive assurance. **The mechanisms to enable regulation to adopt such a concept need to be better understood**.

For autonomous systems that completely remove human oversight, there will need to be **more robust regulatory requirements, with better scrutiny of system design and opportunities for the public to input**. Some emerging applications have caused regulators to assess their existing regulations.

There may be **competition-related regulatory challenges** from autonomous systems and other data-driven technologies that reduce the prevalence of technology acquisitions and takeovers. As concentrations of power shift and industries restructure, the result could be just one supplier and owner of the data that trades these systems or has the necessary computing power. That may change how these systems are regulated and the ability to extract data for the public good.

There may be a **role for a new regulator that looks across typical sectoral silos**, sharing learning from transport to health or education and beyond. This regulator would need to enable innovation while understanding the potential consequences, and **have the expertise to understand the technology and ensure compliance**.

Case study four.

Regulatory review for maritime autonomous surface ships

In shipping, the International Maritime Organization is undertaking a regulatory scoping exercise to look at the applicability of the existing regulatory instruments. This process will determine which will need to be amended to ensure that the safe, secure and environmentally sound use of marine autonomous surface ships. Gaining international agreement on these amendments may prove challenging.

How is the regulatory system planning to adapt?

The existing regulatory model is already being challenged by the speed of innovation, which increasingly exceeds the rate at which deliberative regulatory systems can adapt. There is also a convergence, as **innovations are increasingly 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 or compromise safety when they get there.

Many regulators are now trying to take an enabling approach, with early engagement to try and understand the issues, agree solutions with the industry and enable joint research. **Increasing the agility of the UK's regulatory systems is important to seize the opportunities ahead**.

The goal must be **a responsive regulatory system that connects across the many silos**, that learns from what is happening in one area such as transport and takes that into others such as healthcare and education. This will involve defining what is meant for an autonomous system to be safe or to have the right behaviours in place.

The recent white paper, *Regulation for the Fourth Industrial Revolution*, identified six goals:

- **Future facing** – continuously identifying new opportunities and driving regulatory reform, developing regulatory guidance for innovators and establishing the right governance to address emerging ethical issues including a Regulatory Horizons Council to advise government on priorities for regulatory reform.
- **Informed by society and industry** – creating a wider dialogue about the opportunities and risks from emerging technologies and building confidence in regulation of innovation.
- **Flexible and outcome-focused** – creating a more resilient, flexible regulatory framework to encourage new technology solutions and business models underpinned by eight principles, including safety for people and the environment and competition.
- **Experimentation and testing of innovations** – finding ways to allow innovations to be trialled and inform how regulatory systems need to adapt, learning from the Financial Conduct Authority's regulatory sandbox and the 15 regulatory experiments funded by the Regulators' Pioneer Fund.
- **Support for innovators to navigate the regulatory landscape** – making it easier for innovators to access the system to obtain rapid regulatory advice, reduce time to market, and increase investor confidence in new proposals while giving regulators an insight into what is ahead.
- **Global outlook** – working with partners across the world to shape regulations so that innovations can be freely traded across markets while supporting the sharing of intelligence and the testing of innovations across administrations.

Conclusions

Autonomous systems present many opportunities and risks that vary depending on the sector and the specific application.

Many parallels can be drawn between autonomous systems and existing complex systems as many of the same challenges exist without the addition of autonomy. There will also be opportunities to learn from what has been effective for processes that are already highly automated such as flight. However, for autonomous systems the combination of use in safety-critical applications and complex environments, potential scale of deployment, lack of human oversight and limited transparency on decision-making, increases the risk of harm. The scale of this risk could be viewed as a catalyst for resolving many of the issues that technology is already presenting, especially with many sectors grappling with similar issues with autonomous systems.

As regulatory policy focuses on innovation and experimentation, there will be new opportunities for autonomous systems to be deployed to improve processes and services. However, ensuring that this is done safely will require careful consideration of planning and implementation timelines while new safety assurance methods for autonomous systems are researched.

There are a range of ethical considerations throughout the phases of inception, design and deployment of autonomous systems. At the inception stage, the benefits from the autonomous system must be clearly articulated. In the design phase, decisions will need to be made collaboratively to resolve moral ambiguity and inform the degree to which the system is transparent, unbiased and respects privacy. When it comes to deployment, the public will need to be consulted and the right governance will need to be put in place to make a judgement about whether or not the benefits should be realised despite uncertainty about the risk of harm or asymmetries in benefits.

Non-regulatory measures will help support the decision-making process throughout these phases. However, the extent to which the industry is responsible for defining the baseline for trust in autonomous systems will require continual review by policymakers and decision-makers.

There are many opportunities to learn lessons across application domains, for example military learnings on authority, responsibility and oversight. Cross-sector learning will be helpful to ensure harmful mistakes aren't repeated.

Engineers play a critical role in the development and deployment of autonomous systems. Through a considered approach to design that enables robust verification, validation and evaluation, careful consideration of what could possibly go wrong and working collaboratively many of the risks will be overcome enabling the benefits to be widely realised.

Next steps

The themes presented in this document will be explored and tested through deep dives in specific sectors, , including transport, healthcare and social media. These will consider: **what is unique about how autonomous systems are developing in each sector; the specific challenges to safe and ethical deployment; and identification of emerging good practice**. With an understanding of these different sector contexts mechanisms to ensure safe and ethical development and deployment of autonomous systems will be considered. The NEPC will publish the proceedings and insights as they emerge.

Given the importance of public dialogue the NEPC has sought input from different audiences at the Driverless Science Museum Late and the international perspectives at the 2019 Global Grand Challenges Summit. The results and workshop resources are available at: **www.raeng.org.uk/policy/safety-and-ethics-of-autonomous-systems**

For further information please contact **NEPC@raeng.org.uk**

NATIONAL
ENGINEERING
POLICY CENTRE

ROYAL
ACADEMY OF
ENGINEERING

The **National Engineering Policy Centre** connects policymakers with critical engineering expertise to inform and respond to policy issues of national importance, giving policymakers a route to advice from across the whole profession, and the profession a unified voice on shared challenges.

The Centre is an ambitious partnership, led by the Royal Academy of Engineering, between 43 different UK engineering organisations representing 450,000 engineers.

Our ambition is that the National Engineering Policy Centre will be a trusted partner for policymakers, enabling them to access excellent engineering expertise, for social and economic benefit.

Royal Academy of Engineering
Prince Philip House
3 Carlton House Terrace
London SW1Y 5DG

Tel 020 7766 0600
www.raeng.org.uk
@RAEngNews

Registered charity number 293074