UNIVERSITY OF CAMBRIDGE

CSaP | centre for science and policy

Centre for Science and Policy

Policy Workshop series

Robotics, Autonomous Systems and AI in Safety-Critical Applications

ASSURING AUTONOMY
INTERNATIONAL PROGRAMME

Summary report of the discussions held on 17 May & 11 July 2023

Newnham College, Cambridge; and Institute for Safe Autonomy, University of York

Prepared by Nick Cosstick, Policy Researcher, CSaP; and Dr Magda Osman, Head of Research and Analysis, CSaP (and Visiting Professor, Leeds Business School)

1. Executive Summary

The Centre for Science and Policy at the University of Cambridge, has been working to support the University of York's Assuring Autonomy International Programme (AAIP), part of the Institute for Safe Autonomy, through a series of workshops held between mid-May and July 2023. The topic of the workshops was how academics, policy makers, and regulators can work together to solve some of the policy and regulatory issues caused by the deployment of robotics and autonomous systems (RAS) in safety-critical systems applications.

Each workshop presented one side of the same coin. The first workshop focused on the problems faced by policy makers and regulators (regarding the deployment of RAS in safety-critical systems). Academic experts then related these problems to their own research, and the wider literature. Working together, the first workshop's participants formulated research questions which addressed evidence gaps that hinder the establishment of policy. The second workshop considered the other side of the coin: the issues with the deployment of RAS in safety-critical systems which pose research challenges, especially for assurance.

Through sharing insights into each other's interests (along with room to tie these interests together), this approach builds a foundation from which future collaboration and co-production on these issues can develop (with a lower risk of misalignment in understanding the core needs of the two sides: policy and academia/research).

The research questions generated in the first workshop are categorised using The Taxonomy of Policy Questions.[1] The taxonomy categorises questions by the structure of the information requested (by 'type'). For example, Instrumental/Procedural questions seek information regarding how a policy/tool/lever can be used to achieve a goal. The questions are also categorised by 'theme': by the type of expertise required to answer them. The most frequent (36%) *type* of research questions requested answers that are practical in nature (i.e Instrumental/Procedural). The most frequent (43%) research-question *theme* concerned addressing the specific needs of policy/regulation/legislation. Purely technical (pure basic science) research questions (23%) were more frequently generated than purely social scientific research questions (11%). The analysis of the generated research questions by *type* and *theme* reveals that, by creating a co-productive environment via the workshops, the type and theme of research questions can be oriented more directly towards supporting the needs of policy, and in turn offer a route to direct the impact of research.

The anticipated policy and regulatory issues generated in the second workshop are grouped into four broad topics — which were later considered in the plenary discussion — judged the most pressing areas to target. The policy and regulatory issues of RAS in safety-critical systems concern: (1) validation, (2) quantification, (3) data sharing and incentives, and (4) regulation. The plenary discussion also identified some solutions that could target the issues grouped under these topics — however, these in turn reveal further evidence gaps. For other issues, potential solutions exist but barriers to implementation were also identified.

Combing the insights and the outputs (research questions and policy/regulatory issues) from the two workshops there are seven high-level recommendations. The recommendations are grouped into those that require initiatives where the research community would be responsible for implementation, those where the policy and regulatory community would be responsible, and one where there would be joint responsibility.

[1] https://doi.org/10.1038/s41598-022-21830-z

Recommendations

Research *and* policy and regulatory communities		
1	**Working group**	Create a working group for academics, policy makers, and regulators (and perhaps key consultants) focused on the deployment of RAS in safety-critical systems. The working group would be used to enable co-productive activities that mutually benefit research and policy/regulatory agendas. Terms and conditions of the working group would be established to ensure the demarcation of roles and responsibilities of its members.
Research community		
2	**Research directory**	Establish a comprehensive and publicly accessible research directory. This directory should include a curated list of research questions organised according to the type of research expertise required to address each question — either broadly (purely technical, purely social scientific, etc.) or by subject-specific expertise (e.g. behavioural scientists, engineers, computer scientists).
3	**Interdisciplinarity**	Interdisciplinarity should be upheld as a key principle in research projects tackling the issues generated by the deployment of RAS in safety-critical systems — be it across different academic disciplines or between academics, consultants, policy makers, and regulators.
4	**Expert advice**	The wider academic community should make themselves available for consultation regarding policy and regulatory issues relevant to their expertise. A network of knowledge brokerages in, or across, universities could facilitate efficient access to relevant expertise and high-quality evidence/advice as well as support training for academics to enable them to engage with policy more confidently.
Policy and regulatory community		
5	**Policy and regulation directory**	Establish a comprehensive and publicly accessible policy and regulation directory. A list of clearly articulated policy and regulatory issues, grouped by topic and (where relevant) type of 'expertise' (defined broadly) required to address each issue. It might also include a list of policy makers and regulators who are interested on collaborating on each issue.
6	**Intersectoral and multisectoral collaboration**	Establish a mechanism to enable policy makers and regulators to prioritise the solution of *common issues* — duplicated across many parts of government and which are the most resource draining — and focus on *common goals.* Given that RAS in safety-critical systems engage areas that cut across sectors, this requires a new approach for policymaking and regulation that focuses on the *systems* rather than the specific sectors in which they are implemented in.
7	**Research support**	The policy makers and regulators should provide support to the academics in the working group (and broader networks of academics and industry experts) in terms of (i) supporting the grant applications for their research projects, (ii) citing any work that is of relevance, (iii) data accessibility, (iv) communicate current and forthcoming policy and regulatory issues that could inform research projects along with the pathways to impact.

2. Introduction: Aims and Structure

Safety-critical systems are systems whose failure could result in loss of life, serious injury, equipment/ property damage, or harm to the environment.[2] Given the potentially catastrophic effects of these systems, it is essential to rigorously examine how effective policy and regulation can be used to help minimise their associated risks. Safety-critical systems are ubiquitous in modern society[3] — from pacemakers, to avionics, and nuclear-reactor control systems. Furthermore, the number of safety-critical systems has increased in recent years[4], as has the extent to which RAS — some of which use technologies such as machine learning and robust and adaptive control — are deployed. This shift is bringing novel challenges for regulation and policymaking. Radical changes in standard operating procedures will need to be made. Understanding the insights of researchers, and integrating them into policy and regulation, will be essential. This means that policy makers and regulators need to understand the cutting-edge research, and researchers will need to understand the requirements that policy makers and regulators have — in order to target the most useful applications.

Against this backdrop, CSaP partnered with the AAIP, part of the University of York's Institute for Safe Autonomy, to organise a series of two policy workshops on the policy and regulatory issues stemming from the deployment of RAS in safety-critical systems. The aim was to advance policy and regulatory decision-making through an understanding of the latest research in this area. This was done by increasing the exposure to the AAIP's important work on safety-critical issues — an area which has not received appropriate attention from other research institutes.

The workshops took place on 17 May 2023 in Cambridge and 11 July 2023 in York. The first workshop was attended by 20 participants — 60% of whom work in a policy or regulatory capacity, 35% in the academic sector, and 5% in consultancy. The second workshop was attended by 16 participants, 44% of whom work in a policy or regulatory capacity, 50% in the academic sector, and 6% in consultancy. Across the two workshops, there were 26 unique participants, with some individuals attending both sessions — 54% of whom work in a policy or regulatory capacity, 42% in the academic sector, and 4% in consultancy.

The two workshops were designed to complement each other. The first workshop aimed to help academics understand the problems (with the deployment of RAS in safety-critical systems) which matter to *policy makers and regulators*. Each group was presented with a case study from an attending policy maker or regulator. The themes of the case studies were: (1) the issues associated with semi- and fully autonomous airborne systems, (2) self-driving vehicles and their 'behaviour', (3) using AI to support/automate risk analysis, and (4) using AI/digital twins to map the regulatory landscape. The academic participants then related the problems represented in the case studies to their own research interests, and the wider literature. They worked with the policy makers and regulators to formulate research questions which addressed the evidence gaps related to the policy problems and consider their evidential status (already solved; under investigation, with solutions in the pipeline; and fiendishly difficult).

The second workshop considered the other side of the coin: the problems (with the deployment of RAS in safety-critical systems) which pose significant research challenges. Each group was presented with a case study from an academic participant regarding the broader issues in research on the horizon that will potentially have implications for policy and regulation. The topics of the case studies were: (1) how to craft a framework for assurance of autonomy that has longevity; (2) the problems concerning design, construction, and operation encountered in the deployment of RAS in autonomous chemical plants; and (3) the testing of machine-learning/AI-based systems. The attending policy makers and regulators then helped

[2] https://doi.org/10.1145/581339.581406

[3] https://link.springer.com/chapter/10.1007/978-3-319-47166-2_10

[4] https://www.emerald.com/insight/content/doi/10.1108/IMDS-07-2021-0419/full/html

to shape the discussion by asking the academics questions regarding up-and-coming challenges and emerging themes in the RAS research. Together, each group formulated a set of *anticipated* policy and regulatory issues that reveal new evidence gaps.

The practical outputs of each workshop were generated through cooperation and engagement between two communities with different goals: academics and policy makers/regulators.[5] The practical outputs of each workshop — research questions and anticipated policy and regulatory issues — were generated through co-production. In the sense relevant to this project, 'co-production' refers to (potentially iterative) interaction between the producers and users of scientific research, producing practical outcomes such as research questions or policy agendas — or, more weakly mutual knowledge and understanding, which can be leveraged for the purpose of shaping the work of each party.[6,7] The successful generation of the categorised research questions required that academics listen to the needs of policy makers. The successful generation of the anticipated policy and regulatory issues required that policy makers and regulators listen to the research interests, and expositions, of academics. The insights that each side provided to the other regarding their needs and interests, and the outputs from these exercises, constitutes a foundation from which future collaboration and co-productive projects can develop. The risk of misalignment between policy makers and researchers is still significant— given the different goals of the individuals who comprise these two communities. Yet, the outputs presented in this report provide a resource which can be used to mitigate this risk, and this knowledge base can be built upon in future collaborative work.

This report presents syntheses of the outputs from each workshop — with an effort made to synthesise the insights from across both workshops, where possible. The first workshop's research questions are categorised by 'question type': the structure of the information sought to gain a particular type of research response. (Two such examples are Instrumental/Procedural and Causal Analysis questions.) The questions are also categorised by the theme: the type of research expertise required to address them. In this way, it is possible to expose the complement of disciplinary expertise needed, given that no single research question could be sufficiently addressed in isolation of other disciplines — illustrative examples are presented to show the interconnected nature of addressing RAS in safety-critical systems. The second workshop's anticipated policy and regulatory issues are grouped into four broad topics, those which were identified as the most pressing. The synthesis of both will help to reveal some key recommendations for academics and policy makers and regulators; both individually and together.

3. Presentation and Synthesis of the Outputs

Introduction to organisation of the questions from the first workshop

A total of 47 research-oriented questions[8] generated from the first workshop are classified in two ways. First, by question type, based on Osman and Cosstick's Taxonomy of Policy Questions[9], which groups questions according to the types of answers that they are designed to elicit. Second, by theme: the type of discipline that could best respond — for example is the answer going to be purely technical, or oriented towards solving practical regulatory/policy/legislative problems?

[5] https://doi.org/10.3389/fmars.2020.00409

[6] https://doi.org/10.1016/j.gloenvcha.2004.09.004

[7] https://doi.org/10.1002/wcc.482

[8] All 47 questions classified by type and theme are presented in the Appendix.

[9] https://doi.org/10.1038/s41598-022-21830-z

Classification 1: Some brief details on the Taxonomy of Policy Questions

The Taxonomy of Policy Questions categorises questions via two overarching categories and seven specific question types. The two overarching categories are Bounded and Unbounded: those questions specified in ways which provide more constraints for relevant answers (i.e. closed-type questions) versus those specified in ways which provide fewer constraints (i.e. open-type questions).

The seven specific question types are: Verification, Forecasting, Comparison, Causal Analysis, Explanation/Example, Instrumental/Procedural, and Asserting Value Judgments. Verification, Forecasting, and Comparison fall under the Bounded category, because these question types more heavily constrain what constitutes a relevant answer. For instance, a Verification question requests information that comes in the form of a yes or no answer. (See Table 1 for an example of a Verification question generated in the first workshop.) Causal Analysis, Explanation/Example, Instrumental/Procedural, and Asserting Value Judgments fall under the Unbounded category, because these question types are less constrained regarding what constitutes a relevant answer. For instance, an Instrumental/Procedural question asks how a specific goal can be achieved. (See Table 1 for an example of an Instrumental/Procedural question generated in the first workshops.)

Instrumental/Procedural questions are by far the most frequently generated by policy makers in seeking evidence and expert advice from academia. Understanding what types of questions are generated is important, because it gives a clue as to the general range of answers that are being invited. For instance, we might expect that, to answer a practical question, we will require answers to other practical questions. For example, consider the question ‘how can we increase the ‘explainability’ of the deployment of RAS in safety-critical systems?’. The answer to this question might require an understanding of psychological implications. For instance, it might require an answer to the question ‘what level of expertise should this explainability be tailored to?’. Moreover, the issue may also require an understanding of the underlying causal mechanisms. Thus, it might require an answer to the question ‘how can we demonstrate that increasing the explainability of RAS deployed in safety-critical systems contributes to better regulation?’. Therefore, a variety of inquiries can be made on a single issue that (when directed at the right experts) will bring about a relevant response, and those responses can be utilised in aid of answering other inquiries concerning that issue.

Table 1. The Taxonomy of Policy Questions: Examples from the first workshop

Question Type	Example	Frequency (approximate %)
Verification	Is there a solution to hallucinations in large language models (LLMs)?	5 (11%)
Forecasting	N/A	0
Comparison	What are the benefits and the costs of a digital twin (or, twins) that can be ‘plugged’ into a network (i.e. the real system) itself?	2 (4%)
Causal Analysis	There might be recorded data from a drone, but how does this relate to the ground truth?	8 (17%)
Explanation / Example	What are the thresholds needed to regulate autonomous air systems?	12 (26%)
Instrumental / Procedural	How do we build public and government capacity for working with LLMs (and other emerging technology)?	17 (36%)
Asserting Value Judgments	How sophisticated and complex should a simulation be?	3 (6%)

Classification 2: Some brief details on the thematic classification

One way to approach a thematic classification of the research questions generated from the first workshop, via the insights of the second workshop, is to proceed from the *high-level* insights that were made. One overarching insight from the second workshop — that was shared by the academics, policy makers, and regulators — is that RAS safety-critical systems might be better construed as *processes* to avoid treating the issues that are generated from them into specific sectors, and to avoiding siloed thinking. Thus, a productive and practical way to approach RAS safety-critical systems is that any single issue that arises needs to be addressed by a range of governmental departments (i.e. any issue is in fact intersectoral) and from a range of academic disciplines (i.e. any issue is in fact interdisciplinary). Relatedly, another observation made was that RAS safety-critical systems present issues that could broadly be construed as multi-level — for instance, technical, individual-technical, and socio-technical. The thematic classification of the research question integrated these high-level insights.

While all the research questions generated from the first workshop were geared towards addressing evidence gaps for policy/regulation in some way, they could be differentiated according to whether they did this directly or indirectly. For instance, some were specifically technical in nature, and so indirectly address the needs of policy/regulation, some were behavioural in nature, and so considered the implications at individual and societal level, and of course there were those that were directly considering specific policy/regulatory needs.

The scheme presented in Table 2 shows the mapping of the themes classified into 6 categories, colour coded to show how the questions were grouped.

Table 2 Question themes from the first workshop.

Colour code	Theme: type of research expertise required	Frequency (approximate %)
⬜	Purely technical	11 (23%)
🟨	Technical applications for regulation/policy/legislation	16 (34%)
🟧	Technical analysis of impact of regulation/policy/legislation	4 (9%)
🟦	Purely social scientific	5 (11%)
🟦	Social scientific aspect of technical applications for regulation/policy/legislation	4 (9%)
⬛	Technical and social scientific factors for primarily human-centred issues	7 (15%)

Summary of insights from the analysis of the research questions

1. The most frequent (36%) research questions requests answers which show how a practical goal can be achieved — perhaps via a specific knowledge base, technology, etc. — i.e. Instrumental/Procedural questions.
    - In fact, Instrumental/Procedural questions were most common (44% — 7/16) in the theme concerning technical applications for regulation/policy/legislation needs.
2. The most frequent (43%) research-question theme requests answers that addressing specific needs of policy/regulation/legislation.
3. Pure basic science questions of a technical nature (23%) were more frequently generated than pure basic science questions of a social scientific nature (11%).
4. At least one research question on ethics was raised for most of topics that were discussed.
5. At least one research question from every topic that was covered invited an interdisciplinary response to current and future societal issues that could be solved by autonomous systems.

Introduction to organisation of research themes of the horizon scanning from the second workshop

The idea behind having academics present case studies to policy makers and regulators was for the latter to get a snapshot of the types of emerging issues that are occupying the interests of researchers in the RAS safety-critical domain. In this way, the insights from the academic world are a way for policy and regulation to anticipate and proactively address safety-critical issues on the horizon. Moreover, the plenary discussion was used as a vehicle to identify where there are common topics in the case studies, and to consider where potential solutions might target multiple common issues at once. The three case studies which were presented by academics: 1) Assurance of autonomy that is fit for future purpose, 2) Challenges of automating the validation process of RAS safety-critical systems, 3) Robust and agile approaches to validation and standard setting.

Case study 1: To date, and into the future, there will be continuing issues regarding how to craft a framework for assurance of autonomy that has longevity. Machine learning is a key enabler of autonomy, but while it can achieve significant improvements in performance, the problem is that, when the system meets reality, safety standards need to be met, and typically a 'safety case' needs to be produced. (A safety case is a well-reasoned argument (supported by evidence), provided by developers of RAS, showing that their systems meet a satisfactory threshold of safety.[10]) There are inefficiencies that will continue if the research community doesn't find a way to solve the problem of setting safety standards that take into account that autonomous systems are highly susceptible to change, which comes from the environments in which they operate (because it is dynamic) and because of continual improvements made to their performance through machine learning.

Case study 2: The application of RAS in developing autonomous chemical plants brings challenges for design, construction, and operation. How should a plant controlled by an autonomous system be validated? Certain procedures exist for ensuring the safety of standard design — are they robust enough to cover autonomous control? Digital twins could potentially be used as a proof of concept (concerning part of the system) at the design stage. This issue becomes even harder when machine learning is used, since it makes changes which haven't necessarily been validated. Can we make a machine-learning-based autonomous system learn validation and safety procedures? This point is also relevant to the use of autonomous systems in construction: has the implementation of this process been in line with the design, or has the system made changes which need to be validated? Autonomous operation brings novel challenges. In short, how do we know the plant is functioning as it is supposed to? In particular, monitoring procedures are essential. Furthermore, cyber security will be an absolute requirement for avoiding safety-critical failures. Regulation will be needed to handle both of these challenges.

Case Study 3: Machine-learning and artificial-intelligence systems present unique challenges in safety-critical applications. Unlike traditional non-machine-learning systems, they can't be verified using standard tests and are prone to unpredictable and abrupt failures. Simulation has become the primary method of testing, allowing the evaluation of numerous scenarios, but is it sufficient to assuring the safety of these systems? This question stems from the complexity of machine-learning/artificial-intelligence systems and the heavily skewed statistical distribution of potential outcomes — with long tails of low-probability high-impact events. Thus, minor changes can cause unexpected impacts. The expectation that machine learning/artificial intelligence must be error free and perfectly safe — contrasting with the fallibility of human operators — opens a debate on risk tolerability and public expectations. Strategies like enhancing simulations, employing introspection, red teaming, and real-world "sandbox" testing are being explored, but the risk of overlooking critical features or failures remains. Tactical approaches include adding physical guard rails and adopting standards like those applied to human military personnel. Balancing safety assurance with the recognition of ML/AI complexity is a nuanced task requiring continuous innovation and evaluation.

[10] https://www.jstor.org/stable/44699541

The group discussions and plenary discussion built upon the specific examples provided by these case studies, to the consideration of *anticipated* policy and regulatory issues that reveal new evidence gaps. Four (non-mutually exclusive) topics emerged from the discussion of these issues.

Topic 1: Validation

Observation 1: Our validation techniques are not well fitted to the outputs of RAS safety-critical systems.

Machine-learning-based autonomous systems have the potential to make changes to pre-validated designs. However, this raises the question: is it possible for search systems to learn validation and safety procedures? Moreover, a subsidiary question is what regression-testing protocol (against the most critical scenarios) can be put in place to test for degradation in the system? *A practical response that was considered was whether, as a requirement for certification, the learning function of a RAS safety-critical system might be disabled — and a data-gathering function enabled to give the option to make supervised changes in future.*

The use of a hazards-based approach to validation seems questionable in the case of safety-critical systems. This approach ranks the potential hazards and uses the lowest-ranked hazard as the first test case. However, the problem is that, often, the lowest-ranked hazard will still be safety-critical. This means that creative alternatives are needed.

Case-based validation is the standard at the moment, but, with RAS, humans often are not able to *see* which 'decisions' were made. *Potential solution: digital twins allow for the possibility of maintaining a connection between the real world and the twin. Sensors (with human-like abilities: 'sight', 'hearing', 'smell', and 'touch') can help with real-time monitoring, predictive maintenance, and real-time non-destructive testing.* Discussions also focused on needing to do more to understand the relationship between safety cases and the real world. In particular, what data do we need to establish this link? Moreover, this raised the question: what theoretical assumptions are being made, and how can we justify them (especially given the dangers of poor analogues)? *Another possible solution to case-based validation issues is 'red teaming' — whereby teams work competitively to force a system failure, which is then learned from. Whilst the accuracy of every single 'decision' that a machine-learning-based system makes cannot be individually validated (due to their ability to change), the robustness of that system's decisions can be tested through red teaming. In general, this will involve warping the environment in which these systems operate as a means of trying to force a failure and evaluating its robustness in the face of such tests.*

Topic 2: Quantification

Observation 2: Our quantification techniques may not be adequate in fully capturing RAS safety-critical systems.

Many of the processes underlying the deployment of RAS in safety-critical systems assume that we can quantify things, but often this simply isn't the case. For example, the risk-based approach is the standard strategy. Once decreasing returns on investment in improving the risk-reduction rate are hit, we stop. However, how do we carry out this process for an autonomous system? Relatedly, much of the reasoning which goes into risk analysis is dubious. This moment of reflection provides the opportunity for honesty regarding this, so that improvements can be made. *Potential solution: switch to more qualitative arguments. For example, look at every deployed metric (e.g. risk analysis) and make a case for swapping it for something new. In addition, another solution proposed was to use monitoring data for the purpose of quantification — assuming that the problems with monitoring do not block this. Finally, another solution discussed was a need to emphasise the importance of white-boxed models and systems, so that we can trace the 'reasoning' back to the root cause of a problem*.

Topic 3: Data sharing and incentives

Observation 3: Our data sharing practices are not adequate for regulating RAS safety-critical systems.

Most industries which incorporate safety-critical systems do not share data effectively. *Potential solution: look to the aviation and pharmaceutical industries as case studies concerning best practice. Competitors share operational data which can be used to update designs effectively when failures occur. Government regulation requires this behaviour, thereby incentivising it. This approach could be laterally applied to other relevant industries.*

A real problem here is the trade-off between having all relevant data — and so being swamped — and having a workable amount of data — and running the risk of dishonest sharing strategies from some companies (or simply accidentally overlooking pertinent data). *Potential solution: the strategy used by financial auditors might be the correct model to follow in managing this trade-off. The companies carry out some of the data sifting, so that targeted work can be done with it.* However, we don't know what standards should be used in auditing. This is potentially more of a problem for management schools — what constitutes due diligence in such a case? Yet, given the enormity of this task, the incentives for gaming any regulatory requirements will be considerable. Perhaps reducing the burden down to sharing data related to safety-critical relevant events (such as near misses) might provide more of an incentive to avoid gamification. (Such data could then be used to produce validation test cases.)

Accessing data very quickly eats up the time that regulators have to do their work. Just dealing with 'easy' cases, such as requesting CCTV footage, is time consuming. Such data comes in different formats, requiring different codecs, and different proprietary software. *A potential solution that was considered was to move towards standardisation in data sharing.* However, in many cases, data sharing won't work, because different companies would use different sensor setups and learning algorithms. Standardisation might work against innovation if regulation concerning this is too heavy handed.

Topic 4: Regulation

Observation 4: There is a need to increase capacity and capability of regulation of RAS safety-critical systems.

Regulators struggle with capacity issues. First, they struggle with accessing the expertise they need to solve the complex problems they face. *It was felt that, often, regulators need access to panels of experts, not just lists of expert contacts.* Second, they struggle with demand pressures; for example, in helping industry partners to validate the virtual testing environments they are using. *A solution here might be to have a standardised test environment would help, so that regulators would just need to check the learning models within the environments.*

Regulatory rules are one thing, compliance is another. For instance, recent research has shown the extent to which internet of things data transfer is not compliant with legal standards.[11] *The solution here is to properly fund regulators to deal with compliance.*

Another regulatory issue concerns the expectations we have of RAS in comparison to humans. Our tolerability of risk in the human case is greater, so we must do more to understand what it is *reasonable* to ask of RAS. *A solution would be to have a regularly updated register to capture public perceptions of risk and tolerances of risks in relation to a range of RAS safety-critical systems that are engaged with. This, in turn, can be used as an indirect metric to determine the impact of regulation on public tolerances — if regulation is effective and is observed to be effective then perceptions of risk should reduce.*

[11] https://www.cst.cam.ac.uk/news/internet-stings

A question raised was whether regulatory standards applied to humans, such as certification, could also be applied to RAS safety-critical systems? *In response, one consideration was the UK's relationship with national and international regulatory bodies which might be leveraged for mutual sharing of best practice.*

Combining the insights from Workshop 1 and 2

One way to combine the insights of the two workshop is to look at question type and 'topic' (the subject the question concerns) together. For example, we might look at Instrumental/Procedural questions (type) which concern self-driving vehicles (topic). Even when these two dimensions are kept constant, by considering different questions generated by theme (the type of expertise required), we see the clear variability of different research approaches to practical inquiries regarding this topic. Thus, two questions of the same type, on the same topic, but with different themes, are framed such that their answers target different needs. (Recall that, in the second workshop, discussions consistently revolved around how to approach thinking about autonomous systems.)

Table 3 Examples of generated research questions coded by topic, type, and theme

Topic	Questions	Question Type
Self-driving vehicles	How can we make sure that a simulation is a perfect (or, adequate) representation of reality?	Instrumental/Procedural
Self-driving vehicles	How do we prove that self-driving vehicles are safe?	Instrumental/Procedural
Self-driving vehicles	How do we show the impact of safety standards of self-driving vehicles on improved safety of self-driving vehicles?	Instrumental/Procedural
Self-driving vehicles	How do we get evidence on impact (e.g. societal) of AI?	Instrumental/Procedural
Self-driving vehicles	How do we assure ministers and the public that self-driving vehicles are safe?	Instrumental/Procedural
Self-driving vehicles	Human presence is reassuring (i.e. having 'bus captains'[12] and safety drivers on self-driving buses during trials), how can we find a way to provide these functions?	Instrumental/Procedural

This illustration also helps to showcase how important it is to take a policy-/regulation-relevant topic — such as self-driving vehicles — and consider the different research approaches to examining it which, in turn, will have direct as well as indirect implications for regulation/policy/legislation. More to the point, this shows how the same type of question (Instrumental/Procedural) can be applied in for the pursuit of multiple research approaches.

The aim here is to help show how co-productive activities can generate research questions that would be of practical value, but also recognise how they serve different needs, at different times, for different audiences. It is likely that purely-technical or purely-social-scientific research questions will be addressed as ongoing matters, whereas those that are more solution oriented towards applications for regulation/policy/legislation will be addressed in a shorter time frame because regulation/policy/legislation is time sensitive. All of these factors matter for researchers engaged in research that cuts across both disciplines the needs of different policy/regulation holders.

In fact, this is a point that was raised in the second workshop. A typical approach is to think of autonomous safety-critical systems in isolation, such as separately thinking about drones and their safety-critical implication, and separately thinking about semi-autonomous vehicles and their safety-critical implications. This in turn has implications for how government departments and regulators think about the development of policy/regulation. However, a better way to go is to consider autonomous safety-critical systems as a

[12] There to assist passengers with buying tickets, boarding, disembarking, and other general issues.

process, in which case it is also possible to reduce replication of effort, increase sharing of best practice, and solutions to problems that are domain general, which in turn help to identify those problems that are uniquely domain specific.

4. Recommendations

Research *and* policy and regulatory community

The workshop series on the deployment of RAS in safety-critical systems was made possible by the contributions of the academics, consultants, policy makers, and regulators who attended the sessions (in Cambridge and in York), and together shaped the workshops' outputs. These outputs represent an initial knowledge base which can be added to over time, as a result of further interaction between these groups. Not only can further co-productive activity inform the expansion of this knowledge base, the base can also form the basis of such co-production — since it highlights key knowledge gaps, areas of common interest, and barriers. However, such co-productive activity cannot proceed without an effort to carry on the momentum created in these workshops, and organise it efficiently. Therefore, our first recommendation is:

Research and policy and regulatory community		
1	**Working group**	Create a working group for academics, policy makers, and regulators (and perhaps key consultants) focused on the deployment of RAS in safety-critical systems. The working group would be used to enable co-productive activities that mutually benefit research and policy/regulatory agendas. Terms and conditions of the working group would be established to ensure the demarcation of roles and responsibilities of members of its members.

The **working group** could be used to spearhead collaborative research projects and policy/regulatory reform efforts.

Research community

For the success of research projects concerning the policy/regulatory problems brought about by the deployment of RAS in safety-critical systems, researchers require an agreed-upon record regarding *what the problems and priorities are*. This is the case whether such research projects are organised under the auspices of a **working group** or not. Therefore, our second recommendation is:

Research community		
2	**Research directory**	Establish a comprehensive and publicly accessible research directory. This directory should include a curated list of research questions organised according to the type of research expertise required to address each question — either broadly (purely technical, purely social scientific, etc.) or by subject-specific expertise (e.g. behavioural scientists, engineers, computer scientists).

From the perspective of the research community, the establishment of a **research directory** will be of enormous importance. An easily accessible directory of important research questions will provide them with ideas for research programmes which are of *practical* value. Furthermore, by coding these questions by the type of research expertise needed to address the question (see tables 2 and 3 and the appendix), several opportunities are created. First, the opportunity to signpost the relevance of these issues to academics who might not know that their skillset is required. For example, social scientists and policy-studies researchers might not immediately see the relevance of their expertise to the deployment of RAS in safety-critical systems. In fact, many relevant experts will have very little understanding of this area. This brings us to the second opportunity created: the opportunity for interdisciplinary research projects to tackle the list of problems. The coding of the research questions generated in the first workshop revealed that the most popular kind of expertise required to answer the question was technical applications for regulation/policy/legislation. This alone suggests that genuine interdisciplinarity is required (and this is backed up by the other findings on this topic). Technical researchers might work together with (for example) management experts, policy-studies researchers, and legal experts — importantly, along with policy makers and regulators — to co-produce the answers needed to such multifaceted research questions.

The **working group** could be used to convert these opportunities into further outputs. Therefore, our third recommendation is:

Research community		
3	**Interdisciplinarity**	Interdisciplinarity should be upheld as a key principle in research projects tackling the issues generated by the deployment of RAS in safety-critical systems — be it across different academic disciplines or between researchers, consultants, policy makers, and regulators.

We must also consider what *duty* researchers have to the policy and regulatory community. This community works to different timescales than the academic community. Meaning that they often cannot wait until the end of the publication process to hear the latest insights. This means that researchers need provide support to policy makers and regulators. However, it is not just up to the individual academic. Academic institutions must get involved to broker the knowledge produced under their auspices. Therefore, our fourth recommendation is:

Research community		
4	**Expert advice**	The wider academic community should make themselves available for consultation regarding policy and regulatory issues relevant to their expertise. A network of knowledge brokerages in, or across, universities could facilitate efficient access to relevant expertise and high-quality evidence/advice as well as support training to academics to enable them to engage with policy more confidently.

The recommendation that the research community make themselves available to provide **expert advice** arguably signposts an important role for consultants interested in this area. By incorporating consultants into the **working group** — or, more broadly, the relevant networks — they will be able to help with consultancy work needed by policy makers and regulators. It might be realistic to ask academic researchers to attend consultation meetings and workshops. However, providing rapid evidence reviews and reports might be beyond their capacity. Consultants would be well placed to provide such services.

Policy and regulatory community

For the success of policy/regulatory reform efforts (aimed at solving the problems caused by the deployment of RAS in safety-critical systems), policy makers and regulators require an agreed-upon record regarding *what the problems and priorities are.* This is the case whether such reform efforts are organised under the auspices of a **working group** or not. Therefore, our fifth recommendation is:

Policy and regulatory community		
5	**Policy and regulation directory**	Establish a comprehensive and publicly accessible policy and regulation directory. A list of clearly articulated policy and regulatory issues, grouped by topic and (where relevant) type of 'expertise' (defined broadly) required to address each issue. It might also include a list of policy makers and regulators who are interested on collaborating on each issue.

From the perspective of policy makers and regulators — who work to shorter timescales and face more significant issues with capacity — prioritisation will be key for the effectiveness of further collaboration. If a **working group** is created, then they will determine the prioritisation of the policy and regulatory issues within the directory. However, an important talking point — in the plenary discussion of the second workshop — provides a suggestion for prioritisation. When the discussion turned to what to focus on in the aftermath of the workshop series, it was felt that the policy makers and regulators need some way of solving the common issues which they are facing. In light of this, our sixth recommendation is:

Policy and regulatory community		
6	**Intersectoral and multisectoral collaboration**	Establish a mechanism to enable policy makers and regulators to prioritise the solution of *common issues* — duplicated across many parts of government and which are the most resource draining — and focus on *common goals.* Given that RAS in safety-critical systems engage areas that cut across sectors, this requires a new approach for policymaking and regulation that focuses on the *systems* rather than the specific sectors in which they are implemented in.

The example discussed in the second workshop was the *lack of standardisation* — both in data-sharing formats and virtual-testing environments. The main problem inherent in such lack of standardisation is the enormous amount of extra time taken up dealing with the 'package' that the information of interest comes in — be it a data format or virtual-testing environment. In the case of virtual-testing environments, work with industry partners is ineffective because the greater length of time it takes to validate a testing environment means fewer validation checks are completed — and, therefore, either backlogs or turning away industry partners. Market forces have, thus far, failed to bring about the standardisation required to boost the efficiency of industry-government interaction. Perhaps industry partners, regulators and policy makers, and researchers could work together to co-produce a standardisation policy which would work both for industry and government — without having problematic secondary consequences on either party.

Finally, we have already suggested the duty that researchers have to provide **expert advice** to policy makers and regulators, given the shorter timescales that the latter work to. We suggest that policy makers and regulators should see themselves as having duties towards researchers — in order to incentivise this collaborative activity. The duties concern certain goals which researchers have: funding, citations/recognition, data access, and inspiration. Therefore, our seventh recommendation is:

Policy and regulatory community		
7	**Research support**	The policy makers and regulators should provide support to the academics in the working group (and broader networks of academics and industry experts) in terms of (i) supporting the grant applications for their research projects, (ii) citing any work that is of relevance, (iii) data accessibility, (iv) communicating current and forthcoming policy and regulatory issues that could inform research projects along with the pathways to impact.

Duties (i)-(iv) might be easier to sell to policy makers and regulators who are members of an interdisciplinary **working group** — with common goals and projects, as well as personal relationships across the different professional communities.

5. Appendix: Coded Research Questions

Research questions that invite answers that are technical that could be exclusively answers by machine learning/computational modelling/robotics

Topic of group discussions	Questions	Question Type
Semi- and fully autonomous air systems	If the semi/fully autonomous air system is learning, then what is the process for retesting (for example, should it be centralised or delegated)?	Explanation/Example
Semi- and fully autonomous air systems	How do we determine what level of evidence is required to achieve technical goals regarding semi/fully autonomous air system?	Explanation/Example
Semi- and fully autonomous air systems	There might be recorded data from a drone, but how does this relate to the ground truth?	Causal Analysis
Semi- and fully autonomous air systems	How do we manage a systems of systems (further complicated by coalitions) with an understanding of emergent behaviours of the system?	Instrumental/Procedural
Semi- and fully autonomous air systems	How do we introduce variability to a swarm safely in order to ensure the swarm performs effectively and robustly?	Instrumental/Procedural
Self-driving vehicles	How can we make sure that a simulation is a perfect (or, adequate) representation of reality?	Instrumental/Procedural
Self-driving vehicles	How sophisticated and complex should a simulation be?	Asserting Value Judgments
Automation of product safety risk assessment	Is there a solution to hallucinations in large language models (LLMs)?	Verification
AI (digital twins) and the regulatory landscape	What are the key considerations if a digital twin (or, twins) can be ‘plugged’ into a network (i.e. the real system) itself?	Explanation/Example
AI (digital twins) and the regulatory landscape	What are the benefits and the costs of a digital twin (or, twins) that can be ‘plugged’ into a network (i.e. the real system) itself?	Comparison
AI (digital twins) and the regulatory landscape	What kind of conceptual analysis is needed for the next generation risk analysis (including fault diagnosis) when using AI to simulate threat scenarios?	Explanation/Example

Research questions that invite answers that are technical solutions oriented towards addressing policy/regulatory/legislation

Topic of group discussions	Questions	Question Type
Semi- and fully autonomous air systems	At what point do we need to regulate autonomous air systems using special measures?	Asserting Value Judgments
Semi- and fully autonomous air systems	Who decides what the special measures are that are needed to regulate autonomous air systems?	Asserting Value Judgments
Semi- and fully autonomous air systems	What are the thresholds needed to regulate autonomous air systems?	Explanation/Example

Semi- and fully autonomous air systems	How are the thresholds defined that are needed to regulate autonomous air systems	Explanation/Example
Self-driving vehicles	How do we prove that self-driving vehicles are safe?	Instrumental/Procedural
Self-driving vehicles	How do we prioritise areas for deployment of AI?	Explaining/Example
Automation of product safety risk assessment	How do we build public and government capacity for working with LLMs (and other emerging technology)?	Instrumental/Procedural
Automation of product safety risk assessment	How do we build public and government capacity preparing for LLMs (and other emerging technology)?	Instrumental/Procedural
Automation of product safety risk assessment	Is there a solution to the copyright and policy issues associated with using LLMs within government?	Verification
AI (digital twins) and the regulatory landscape	How could a digital twin (or, twins) map the regulatory landscape to identify efficiencies to achieve net zero?	Instrumental/Procedural
AI (digital twins) and the regulatory landscape	How could a digital twin (or, twins) map the regulatory landscape to identify where to introduce interventions to generate price efficiencies for the consumer?	Causal Analysis
AI (digital twins) and the regulatory landscape	Is it that using digital twins for the goal of increasing efficiencies to achieve net zero is a trade off against the goal of identifying ways to generate price efficiencies, or are they related goals?	Comparison
AI (digital twins) and the regulatory landscape	How can risk analysis be integrated into a digital twin (or, twins) to reveal vulnerabilities in a system (e.g. the domain of regulation—e.g. transport, health care, and energy)?	Instrumental/Procedural
AI (digital twins) and the regulatory landscape	How can digital twins be used to examine the different value systems of agents (e.g. regulator, business, consumers, and special interest groups) operating in a regulatory environment?	Instrumental/Procedural
AI (digital twins) and the regulatory landscape	How can the trade-offs be optimised to ensure the safest system that benefits the consumer the most when using digital twins to map the regulatory environment?	Instrumental/Procedural
AI (digital twins) and the regulatory landscape	What is the level of 'explainability' of explainable AI that is needed given the audience **that uses** the application of AI in the safety-critical domain (i.e. level of explainability for legal purposes, for technicians, for civil service analysists, for civil service policy makers)?	Explanation/Example

Research questions that invite answers examine the technical outcomes resulting from policy/regulatory/legislation

Topic of group discussions	Questions	Question Type
Self-driving vehicles	How do we show the impact of safety standards of self-driving vehicles on improved safety of self-driving vehicles?	Instrumental/Procedural
Semi- and fully autonomous air systems	What would be the effects of setting standards (or reaching agreements or signing treaties) for semi/fully autonomous air systems?	Causal Analysis
Semi- and fully autonomous air systems	How would the effects of setting standards (or reaching agreements or signing treaties) for semi/fully autonomous air systems affect proliferation?	Causal Analysis

Semi- and fully autonomous air systems	How would the effects of setting standards (or reaching agreements or signing treaties) for semi/fully autonomous air systems affect competitive advantage?	Causal Analysis

Research questions that invite answers that are primarily inviting responses from social scientific disciplines

Topic of group discussions	Questions	Question Type
Semi- and fully autonomous air systems	How do we determine what level of evidence is required to gain human/societal trust regarding semi/fully autonomous air system?	Explanation/Example
Self-driving vehicles	How do we weigh the (social) impacts of AI?	Causal Analysis
Self-driving vehicles	How do we get evidence on (social) impact of AI?	Instrumental/Procedural
Self-driving vehicles	Are we focusing AI on the right areas (of society/societal needs)?	Verification
AI (digital twins) and the regulatory landscape	What ethical issues do we need to consider when using self-learning AI systems (i.e. machine learning that learns from other machine learning) given the domain and data that the systems are trained on?	Explanation/Example

Research questions that invite answers that prioritise social/psychological factors that require solutions that support policy/regulation

Topic of group discussions	Questions	Question Type
Self-driving vehicles	How do we assure ministers and the public that self-driving vehicles are safe?	Instrumental/Procedural
Self-driving vehicles	How do we engage the public on new technologies?	Instrumental/Procedural
Self-driving vehicles	Could we link new technologies to societal change and new opportunities it offers (i.e. supporting those with disabilities—increased mobility, can work from anywhere)?	Causal Analysis
Self-driving vehicles	How do we communicate the benefits of this technology, acknowledging that society is not a homogeneous group and the impact of AI will vary among individuals (some lose jobs, others gain opportunities)?	Instrumental/Procedural

Research questions that invite answers that involve integration between computational and social science oriented towards addressing social/psychological factors

Topic of group discussions	Questions	Question Type
Self-driving vehicles	Human presence is reassuring (i.e. having bus captains and safety drivers on self-driving buses during trials), what functions are linked to this human presence?	Explanation/Example
Self-driving vehicles	Human presence is reassuring (i.e. having bus captains and safety drivers on self-driving buses during trials), how can we find a way to provide these functions?	Instrumental/Procedural
Self-driving vehicles	How do we link advancement of AI with social change to increase acceptability of self-driving vehicles?	Causal Analysis

Automation of product safety risk assessment	How can we combat mis- and disinformation brought about via LLMs?	Instrumental/Procedural
AI (digital twins) and the regulatory landscape	Can AI be used to develop the next generation of nudges?	Verification
AI (digital twins) and the regulatory landscape	Can we help to achieve even more effective nudges through nudge 2.0 through the use of AI?	Verification
AI (digital twins) and the regulatory landscape	What is the level of 'explainability' of explainable AI that is needed given the audience that has **oversight** over the uses of the application of AI in safety-critical domain (I.e. level of explainability for legal purposes, for technicians, for civil service analysists, for civil service policy makers)?	Explanation/Example