Workshop Report: From Ethical Principles to the Ethically-Informed Engineering of Autonomous Systems

Zoë Porter

26 February 2021

Contents

1. Introduction

The many sets of high-level ethical principles for AI and autonomous systems indicate a widespread societal concern about the ethical implications of these technologies. But, despite some notable examples, operationally-specific practical guidance on how to address this concern, during the design and engineering of the systems, is not yet at a mature stage.

On 22 January 2021, the University of York's Assuring Autonomy International Programme (AAIP) held an Ethics Workshop to contribute to the maturing of, and critical reflection upon, that practical guidance.

We invited participants from three communities: technical; ethics; and regulatory policy. There were 19 discussants present, and the workshop ran for 2.5 hours. No formal presentations were given. The format was a discursive, multi-disciplinary knowledge exchange on four questions:

1. What are the best ways to translate ethical principles into ethically-informed engineering practice?
2. Should the same core ethical principles be included in the decision-making process of every development team?
3. How should engineers approach ethics for, and under, uncertainty?
4. Is an 'ethical kitemark' a good idea for systems developed by teams who have done their best to consider ethical impact?

This report details the answers proposed in response to the four questions, abbreviating the discussion questions to:

1. Best ways to translate ethical principles into engineering practice
2. Standardisation of ethical principles
3. Ethics for, and under, conditions of uncertainty
4. Certification of ethical assurance practices

The report also gives details on some further themes that emerged throughout the course of the discussions, and it outlines next steps for AAIP research activity.

2. Executive summary

Here is an overview of the main findings and discussion points.

A high-level summary of answers given to each of the four discussion questions is as follows. These answers are expanded in more detail in **Part 3**.

1. Best ways to translate ethical principles into engineering practice

Engagement with this question primarily took the form of proposing different methodological starting points:

1. Start by trying to establish reflective, system-specific societal consensus on goals;
2. Start with foundational conceptual clarity, and a delineation of which properties (whether of the human organisation, or of the system itself) are being assured by which type of assurance process;
3. Start by leveraging existing traditions and frameworks within engineering practice;
4. Start by engaging with technical processes at the granular level;
5. Start with case-based reasoning, looking at concrete use cases to understand the ethical implications of different design and engineering choices.

The importance of dynamic ethical assurance was also discussed.

2. Standardisation of ethical principles

Systems are being developed for a diverse range of purposes. Participants agreed this means that ethical principles will apply differently in each case. Trade-offs are also to be expected. But there was an informal consensus in favour of a broadly standardised framework of ethical principles, which does not foreclose the possibility of new ethical principles and allows for flexibility at the implementation level.

3. Ethics for, and under, conditions of uncertainty

The computer scientists and engineers in the workshop provided details about the breadth, complexity, and many dimensions of uncertainty faced during the process of building autonomous systems.

Different solutions were proposed. Technical solutions focused on system architecture or on engineering methodology. Ethicists and philosophers looked at the problem from the angle of other domains – such as the environment, medicine and the pharmaceutical industry – where uncertainty has been regulated effectively for some time.

It was also recognised that the uncertainties faced by designers and engineers of autonomous systems, particularly ML-based systems, are such that they cannot be quantified *ex ante*.

4. Certification of ethical assurance practices

An ambiguity in the concept of an 'ethical kitemark' – between assurance of the system and assurance of the process – makes it unlikely that this specific idea is a good one in practice.

But it was agreed that some sort of certification was both likely to happen in the real world, and that this would be a good thing.

Throughout the course of workshop discussion, three further themes emerged, and appear to be foundational:

1. The tension between qualitative and quantitative.
2. The accuracy and adequacy of current language.
3. The question of whether, to what degree, and in what respect autonomous systems present an unprecedented regulatory challenge.

These are discussed in more detail in **Part 4.**

Part 5 outlines **next steps** for AAIP research on these issues. One of AAIP's objectives is to build on best practice and the wider literature to write implementable guidance for developers and others addressing the ethical implications of autonomous systems. This workshop has been instrumental to us in thinking about how to frame that guidance. We will also explore more systematically the intersection of discussion questions 1 and 3.

Part 6 gives a **list of participants**.

Part 7 provides **references** of publications mentioned within this report.

Throughout the report, speaker initials are given next to the participant. These may not be expressions of firm positions; the tone of the workshop was open, with people speaking off-the-cuff as well as sharing their considered insights and professional experiences.

We are grateful to all of the participants for their time, contributions, and collegiality.

----

3. Responses to the discussion questions

This part of the workshop report provides fuller descriptions of the discussions on the four central workshop questions.

Speaker initials are given next to the relevant participant (for reference, see participant list in Part 6).

3.1. Question 1: Best ways to translate ethical principles into engineering practice

In response to the first question, participants proposed different methodological starting points. These are ordered below:

1. ESTABLISH REFLECTIVE, SOCIETAL CONSENSUS ON GOAL FOR THE SYSTEM (CB)

Start by trying to establish a societal consensus on the system's goals. These goals should be treated as hierarchically decomposing, and then tracked throughout the specification and implementation process. In some cases, this tracking may lead to the decision that the system should not be developed or deployed at all. Consensus on goals precedes, and sets the constraints for, an ethical risk analysis.

Responses to 1:

- Mirrors deliberations in healthcare, where the social license is key. The salient questions are: do we want to develop the system and, if so, which principles apply and who decides? (FM)
- How precise do we need to be about the goals before proceeding with design and engineering decisions? (SB)
- Goals and user-context will change when systems become adaptive to environments. (GR)

2. START WITH FOUNDATIONAL CONCEPTUAL CLARITY, AND A DELINEATION OF WHICH PROPERTIES ARE ASSURED BY WHICH TYPE OF ASSURANCE PROCESS (DL)

Distinguish between high-level ethical values and more practical principles. Ethical values (which include dignity, solidarity, autonomy, and sustainability) are action-motivating, and set the direction of travel for innovation. The practical principles (which include bias mitigation, safety, and explainability) are incorporated into the actual practices of building the systems. Ethical values can be assured by an ethical impact assessment. The practical principles require more specific, evidence-based assurance arguments.

Responses to 2:

- Distinction between assurance of the human organisation and assurance of the system itself. (JMcD) Distinction between values that attach the business model and those that attach to the system. (AS)
  - *Response*: precisely that distinction between process-based and outcome-based explanation is crucial. There is the scrutiny of those qualities that

come from the organisation, and the scrutiny of the technical system. This was central to the ICO-Turing guidance on Explainable AI. (DL)

3. LEVERAGE EXISTING TRADITIONS AND FRAMEWORKS WITHIN ENGINEERING PRACTICE (AW)

Start with ethical standards, such as British Standard 8611. It provides a methodology for ethical risk assessment, building on established safety engineering practices: horizon-scanning; risk assessment; failure modes and effects analysis. It involves broadening the envelope on what is already done by the design and product team. That team should ideally be as diverse as possible. Standards are only one of a raft of methods, but they are a good place to start.

Responses to 3:

- What's the difference between an ethical risk assessment and an ethical impact assessment? (CK)
  - *Response*: ethical impact assessment is broader. (AW)
- The design and product team may not appreciate the nuances of ethical impact when interpreting the standards; an ethical risk is different to a technical or safety risk. (ZP)
- In practice, only a few standards are significant, at least in safety. And a great deal comes down to how standards are interpreted. (IH)
- We're moving towards having a more self-reflective awareness of standards, with IEEE P7000 and the AI groups working in ISO. (DL)
- Engineers need training in Responsible Research and Innovation, both in academia and in industry. We are not upskilling the workforce and training people early enough. (GR)
- Ethnographic studies reveal AI ethicists come up against significant barriers within companies; could that extend to the implementation of ethical standards? (AB)
  - *Response*: we need a lot of advocacy from professional bodies, as well as levers such as public procurement (AW).

4. ENGAGE WITH THE TECHNICAL PROCESSES AT THE GRANULAR LEVEL (EK)

Sometimes an ethical principle, such as fairness, is already being implicitly addressed by the technical processes, and may be being addressed sufficiently. This can be a good place to start when considering ethically-informed governance practices.

Responses to 4:

- Some things, such as data preparation and training function, can be measured at a deep technical level, but this does not extend to an evaluation of ethical principles, nor how to assure the governance processes. We should be cautious of trying to make everything as measurable as possible. (JMcD)

5. REASON ABOUT USE CASES TO UNDERSTAND THE ETHICAL (AND LEGAL) IMPLICATIONS OF DIFFERENT DESIGN AND ENGINEERING CHOICES (GR)

Use cases provide concrete examples that can ground understanding and deliberations during the design and engineering process. Contextual issues will foreground some of the ethical challenges that we face. Decisions to deploy systems turn on thinking about the situations in which the systems will be embedded. We need multi-disciplinary engagement to help engineers construct and consider use cases where a decision has been delegated to the system, and there is a risk it will misalign with ethical values or ethical principles.

In addition to discussing where to start operationalising ethics during the design and engineering process, participants discussed the need for ongoing assurance of the systems.

It was widely agreed that ML-based, robotic and autonomous systems need dynamic assurance processes, to enable human actors to target their interventions on the system over time in response to unexpected events in the operating environment.

6. ETHICAL ASSURANCE PRACTICES NEED TO BE DYNAMIC (AZ, GR, DL, JH, IH)

- Strict criteria can be specified and implemented, but the system could still do something that no one foresaw. We can construct rules for responding to surprising events - whether to investigate the system further or to advise against deployment in its current state or for this purpose. That would be a practically, context-specific rule that supports dynamic assurance. (AZ)
    - *Response*: Yes, impact assessment has to involve continuous assessment, perhaps with a revision rule as you suggest. Reflective, continuous assessment is needed to keep up with concept drift, data drift, and unpredictable events. (DL)
- There are moments when we need to ask the difficult, contextually-based value questions – when scanning the horizon, setting the target variable, formulating the problem – but it doesn't end there. That's where sustainability comes in. (DL)
- The Law Commission is looking at these sorts of issues with automated vehicles: an in-use safety monitoring function over the system's lifecycle, with a responsibility on users to ensure safety and security updates are done. (JH)
- Data-rich systems can generate representation of their operational history; this can inform continual review and support sustainability. (JMcD)
- Engineers need more awareness on how users perceive failure. Earlier exposure and discussion of failure cases can help to understand how users perceive those failures, and contribute to adoption of the systems in the long-term. (GR)
    - *Response*: safety engineering already has good processes for understanding failures and their consequences, but these do need to include impact on users and wider society. (JMcD)
- This will become particularly pressing when the systems become more adaptive. (GR)

3.2. Question 2: Standardisation of ethical principles

In response to the second question, a broad consensus emerged that some form of standardisation of the ethical principles in the assurance framework would be a good idea in principle, with caveats that this must allow for flexibility according to context.

Some specific responses to this question are below:

- Transparency is key to all of this. Transparency of the system, of the designer's intent, and in the processes around the system. *That* should be a common principle or value. (MF)
- I think that having an overarching set of principles of ethics that everyone tries to adhere to, which are then implemented and amended as appropriate to context, would be the way forward. (CP)
- Common ethical principles and values can provide a starting point for an ongoing conversation. This does not have to mean we have to be so strict about them that we might foreclose possibilities of new principles and values emerging as the technologies develop. (DL) Technologies open up new horizons that invariably require us to sharpen our ethical principles or to create new ones entirely. (CB)
- The relevant ethical principles are application-specific - even different applications within the same domain, such as healthcare, will have different principles that apply. (FM)
- The same ethical principle may play out differently depending on context (e.g. informed consent will play out differently). We will also need to evaluate trade-offs between principles. (CB)
- A process of reflective equilibrium can help us to evaluate which particular principles arise and are action-guiding in which cases, and how principles should be operationalised. (CB, SB)

3.3. Question 3: Ethics for, and under, uncertainty

The computer scientists and engineers in the workshop provided details about the breadth, complexity, and many dimensions of uncertainty they encounter in the development process.

Different solutions were proposed. Technical solutions focused on system architecture or on engineering methodology. Solutions from ethicists and philosophers considered the problem from the angle of other areas – such as the environment, medicine, and the pharmaceutical industry – where there are established governance mechanisms for dealing with uncertainty.

Responses to the third question are organised here into three distinct parts.

1. Descriptions of the uncertainties faced by developers;
2. Technical approaches to uncertainty;
3. Regulatory suggestions for dealing with uncertainty.

3.3.1. UNCERTAINTIES FACED BY DEVELOPERS

The uncertainties described were striking in both depth and scope. The sources included: system uncertainty; the epistemic uncertainty of developers; uncertainties about the operating environment, and the interaction of multiple actors in *"a living system"* (DL's phrase). In addition, there are complex interactions between these uncertainties.

System uncertainty

- Uncertainty is inherent to the technology itself, specifically to ML-based systems, which have a technical imprecision running through them, from the sensors, to the algorithms that are used to interpret sensor data, to the decision algorithms themselves (e.g. system may be 80% sure object x is a car in one video frame, 50% x is a car in the very next frame, or 100% x is a tree, and nevertheless be 100% wrong). (SB, GR)
- There is nothing new here. Normal cyber-physical systems have lots of uncertainty in them. Reinforcement learning algorithms are just feedback systems. All the techniques that have been developed for uncertainty management and uncertainty modelling in normal cyber-physical systems can be used to manage autonomous systems. In particular, system architectures that protect key decisions from uncertainty will be vital. (MF)

Epistemic uncertainty

- We face significant uncertainty about whether we have defined and specified the right set of technical system requirements (often called the 'semantic gap'). Most accidents come down to not getting the requirements on the system right; this epistemic uncertainty is critical. (SB)
- There is uncertainty in terms of behavioural and cognitive science not yet having evidence on new harms, such as psychological harms, and their impact. (CB)

Uncertainty from operating environment

- Uncertainty from the synchronically (at any given point in time) and diachronically (over time) complex environments. (GR, AZ)

Uncertainty about behaviour of actors in the environment

- Uncertainty arises due to the sheer number of actors. There are different combinations of human-machine partnership and differential behaviour patterns that accompany these combinations. There is a diversity of types of human behaviour and reactions. There is the possibility of malicious actors, as well as actors (whether human or artificial) acting rationally to maximise their own self-interest or utility. (GR)
- Cognitive bias of users. (CB)

Complex interactions of uncertainties

- When the environment changes, the original system goal may no longer be suitable. (FM)
- There is a danger that we may optimise for one type of uncertainty at the expense of another. (AZ)

3.3.2. TECHNICAL APPROACHES TO UNCERTAINTY

- Deal with it through a **hybrid system architecture**. Separate the decision-making component from the ML-components. Keep the decision-making component a rule-based system, which can be verified, can self-evaluate, and can evaluate the reliability of the ML-algorithms. (MF)
    - *Response*: does this mean we are going to exclude ML techniques, like reinforcement learning, from decision-making tasks, and restrict the use of this technology in the future? (SB)
    - *Response:* absolutely – if strong evidence cannot be provided for a decision-making component in critical areas then we should not be afraid to disallow this. It is irresponsible to allow just any technology into such situations. (MF)
- **Evolve the methodology**. To design and manage human-machine partnerships we need to model the user, user preferences, the environment, mathematically engineer that into the system, and optimise for efficiency. But we also need to include more intuitive and qualitative human factors, reactions, and context. We do the latter through user studies, perception studies, and field trials. But there is still a fundamental tension, because the prevailing methodology guides you towards working for efficiency, avoiding errors, avoiding redundancy, and ensuring human actors achieve their objectives. This can help us answer questions such as when to delegate to the machine and when to the human, and which level of control gives the best outcome, but it doesn't account for the wider ethical and legal implications of those decisions. (GR)

3.3.3. REGULATORY SUGGESTIONS FOR DEALING WITH UNCERTAINTY

- We should focus on **societal deliberation about which benchmarks are appropriate**. E.g. how reliable should we deem particular systems given particular measures of uncertainty? Established uncertainty measures (e.g. Brier scores) can get us quite far. We shouldn't fall into a trap of "autonomous systems exceptionalism" where we throw out everything we already know about decision-making under uncertainty. From a regulatory perspective, dealing with uncertainty in this domain need not necessarily be dramatically different to other inherently uncertain, high stakes domains, such as the environment or medicine. (AZ)
- Regulators could develop different **"tolerance of uncertainty" levels**, which are increasingly morally strict for highly critical systems. (CP)

- *RESPONSES*:
- With ML-based systems, the probabilistic models are not just wrong, we don't even know how wrong they are. We would not know if a threshold or benchmark of uncertainty had been met. (MF)
- We need to find ways not just to approach ethics under uncertainty, but also to communicate that uncertainty itself. (SB)
    - *Response*: Communicating uncertainty can be dangerous. Trust, cognitive bias, and the interpretation we give to statistical interpretation have to be incorporated as well. (CB)
- Users find it difficult to interpret things like Brier scores, and to integrate that into their own decision-making. People need to be trained in probabilistic reasoning skills to accompany the deployment of systems. (AB) Is there a role for regulators to translate Brier scores into something the public can understand? (JH)
- There is domain-specificity in human judgements about how much uncertainty is appropriate. (AB)
    - *Response*: Yes, that's an important point. This suggests the need for iterative feedback from stakeholders on uncertainty. (AZ)

3.4. Question 4: Certification of ethical assurance practices

The fourth question was initially answered by putting it in context and clarifying what was being certified. The notion of a 'kitemark' was seen to be too ambiguous to be helpful. While participants agreed some form of certification was both on the horizon and necessary, there were caveats that certification alone was not sufficient and should be a part of a whole system of approaches.

A new category was suggested – *"ethically-critical systems"* – in addition to the category of safety-critical systems. (AW)

- It is important to clarify the following distinction: it is one thing to institute standards for ethical impact assessment (which involves human decision-makers weighing ethical values) and it is another thing to say these are the ethical principles that operationalised and technically manifested within the system itself. The direction of travel hopefully points to the former. (DL)
- Taking aviation as an example, you need to look at both the organisational processes and the system itself. A certification or marking of systems would be worthwhile, it would provide a minimum threshold. But this sort of compliance is not sufficient to show that the system is good enough to deploy for a particular purpose. (JMcD)
- The idea of an 'ethical kitemark' is potentially misleading – you need to separate out what the kitemark is giving assurance of – and it may not be possible to define. But some kind of assurance by a third-party assurance is essential. We should be clear on what we are trying to achieve with this assurance process: regulatory compliance; public trust; technical rigour. (CK)

- Systems have components from different suppliers and may be used in different contexts to the ones designers intended. A kitemark would need many caveats, which would make it difficult to implement. (AS)
- When it comes to safety-critical systems like AVs, an ethical standard is likely to be wrapped up in general safety assurance. We also often talk of the behaviour of the AV in a holistic manner. I don't know how or to what extent you could ethically kitemark individual components of the AV. Aside from training data, I'm not sure what input ethics could be kitemarked. Rather, you could only judge the output behaviour. Doing so is also highly dependent on the skills set of regulators. (JH)
- Ethical issues broaden the scope of regulation, but regulatory frameworks don't always encourage this wider thinking. The GDPR (data protection law), for example, does not give much scope for thinking about broader ethical considerations such as fairness and accountability. (CK)
- The important role of professional bodies (such as IEEE, British Computer Society, Royal Academy of Engineering, Nesta) was emphasised as part of the wider assurance and certification framework, to advocate for and to encourage compliance. (AW)

The conversation around certification turned to the question of the relation between standards and law, and to the importance of ethical governance.

- We should be clearer that standards and statutory law are underpinned by existing norms. (DL)
- We've reached a point where we need to make explicit what the ethical values are because we are now confronted with the question of a system making decisions in place of a human, whom we would have held personally accountable. (DL)
- Imposing ethical values on people could move away from creating trust. The law probably has the more important role to play here. (EK)
    - *Response*: Yes, I share trepidation about imposing ethical values. But we can distinguish between procedural and substantive ethics. The latter might involve imposing an ethical value. But procedural ethics can be part of good governance processes. (DL)
- Ethical governance can give significant confidence and build trust with the public. Google is an example of an organisation that is eroding public trust because the quality of the ethical governance is being called into question. Internal whistleblowing mechanisms are crucial to good governance. (AW)
- Perhaps people want a public body that they can place confidence in for certification, and that satisfies public demand for the "theatre of the process" as well as the actual assurance. When the MHRA signed off the Covid vaccine, this became a big landmark event, even though a lot of the work was probably done in back room processes. (CK)
    - *Response*: I think having a body which makes explicit decisions and says "yes, this is good enough to be deployed" or "no, you should stop" is right. I would probably say "bodies" not "body" because understanding of the specific domain is really important. That may mean that all the existing regulators have to take on these additional skills. (JMcD)

- From being involved in CAHAI's feasibility study of a legal instrument on AI, it is interesting that the perspective there is that human rights, such as human autonomy and equality, are being fixed at a high legal level. The approach is that soft law mechanisms sit beneath that higher level. (DL)

4. Recurring themes

Throughout the breakout and plenary discussions, several underlying themes emerged. These signal issues that look foundational.

4.1. The tension between the qualitative and quantitative dimensions of assurance

- Ethics is largely qualitative. Engineering is inherently quantitative. A tension between them is to be expected. There are lots of deep challenges to operationalising qualitative goals. (AZ) Monitoring of socio-technical systems cannot be equated with measuring in the technical sense. (DL) It is very hard to quantify context and use.
- Different approaches to reconciling the qualitative and quantitative dimensions of assurance can be abstracted from the participants' remarks. Broadly, they can be summarised in the following structural terms:
    - Reduce the qualitative to the quantitative (the 'official engineering' perspective)
    - Reduce some but not all of the qualitative to the quantitative (a modular or hybrid approach)
    - Tackle them in parallel, with differently-oriented assurance processes (a multi-track approach)
- One participant, drawing on first-hand experience in safety assurance, pointed out the real-world difficulties of maintaining focus on the qualitative dimension of assurance arguments: *"once a quantitative measure is out, the qualitative argument is lost."* (IH)

4.2. Accuracy and adequacy of language

- The importance of clear, unambiguous language was raised at several points. This serves both user safety and regulatory progress. To paraphrase J. L. Austin: words are our tools, and we should use clean tools.
- There were three strands to this recurring theme:
    - Accuracy of descriptive language;
    - Adequacy of normative language;
    - The need for a common language.

4.2.1. ACCURACY OF DESCRIPTIVE LANGUAGE

- Ambiguous and exaggerated descriptions of **systems** can lead to incorrect expectations of, or over-reliance on, the system's perceived capabilities. This is a safety risk. (GR)
- Regulators could lead the way by enforcing accuracy of terminology. Unfortunately, they sometimes contribute to the problem. It was noted that a German Court (Munich) has banned Tesla Germany from using advertising claims like "Autopilot" or "full potential for autonomous driving", on the grounds that they could give the impression the cars can drive without human intervention. (GR, ZP).
- The **assurance guidance** should also be accurately stated, so that people understand precisely what properties are being assured by any given process and how. (GR, DL)
- The importance of descriptive clarity on **ethical concepts** was also emphasised. For example, a delineation of ethical goals, ethical values, and ethical principles. (DL, CB, CK)
- We also need to state clearly which concepts apply to the human organisation and which to the system itself. (CK, AS, JMcD, DL)
- Much greater accuracy and clarity on the **different properties of different systems** was also highlighted as important: "*Autonomous systems are different to AI systems, which are different to ML systems*." (MF)

4.2.2. ADEQUACY OF NORMATIVE LANGUAGE

(i.e. language expressing what people 'should' do – rationally, ethically, legally, or prudentially)

- We may be facing a "poverty of language" around ethics in this space, particularly in the distinction between ethical discussions at the regulatory level and ethical discussions at the technical level. As an umbrella term, "ethics" perhaps obscures more than it helps. (CK)
- Communication with users should respond to the fact that users and operators often impute an inaccurate 'mental model' to the system. (IH)
- Even the use of the term "standards" can be misleading. It suggests something technically precise. Healthcare regulators tend not to use "standards" because they acknowledge the difficulty of quantifying some of the dimensions of evaluation. (IH)
- "Ethics washing" is a real danger, and could occur when developers and organisations make statements about their ethical procedures. (SB)

4.2.3. THE NEED FOR A COMMON LANGUAGE

- We need a common language. (SB)
- We need a common vocabulary for anticipatory reflective discussions between engineers and policy makers. A common normative vocabulary will help to create the conversations that will guide our responsible practices. (DL)
- Engineers would benefit from this mutual understanding when trying to define system requirements. (SB)
- Regulators would benefit when considering appropriate assurance processes for these specific technologies. (MF)
- A common language could emerge from engagement in user case studies. (GR)

4.3. Autonomy exceptionalism?

There was a divide amongst participants as to whether autonomous systems are unprecedented, and whether regulating organisational processes around the systems, as well as the systems themselves, requires "reinventing the wheel". (CP) Some questioned the value of *"autonomy exceptionalism".* (AZ's phrase) Others highlighted the distinctiveness of autonomous technology, and the disruption it presents to our moral, legal, and conceptual frameworks. (DL, JH)

WHAT IS NEW OR UNPRECEDENTED ABOUT AUTONOMOUS SYSTEMS?

- Fundamentally, what is distinctive about autonomous systems is that the machine substitutes the decision-making human actor. This has unprecedented practical, ethical, and regulatory implications. (expressed variously by MF, DL, ZP)
- What's become exceptional with AI, ML-based, and autonomous systems is that the technology itself is standing in for human judgement in ways that other technologies do not; they're now serving human cognitive functions. This calls on us to re-think traditional approaches to standards and regulation. (DL)
- The difficulties in holding anyone criminally responsible for accidents caused by highly automated systems is also unprecedented. Civil responsibility mechanisms exist, but that may not satisfy the public in the event of accidents. (JH)
- Traditional safety approaches do not work well for autonomous systems in unknown environments. You cannot possibly work out all the things that could go wrong. The step-change is that autonomous systems make decisions in the place of humans. This shifts the question from "what could go wrong?" to "how do we get the system to decide what to do in an unknown situation?" (MF)
- Much discussion on standards presupposes that autonomous systems are like ordinary cyber-physical systems. Current approaches to assurance don't fully appreciate what 'autonomy' means. For that reason, we're in the middle ground. Deployed systems are not fully autonomous. (MF)

- We need a different process. Autonomy means that safety engineers are now coming up against questions/barriers that traditional safety approaches are not equipped to deal with: How safe is safe enough? How accurate do we need to be? (SB)

WHERE IS THERE A PRECEDENT?

- The mechanisms that are in place in the pharmaceutical industry could be very helpful, because it is another domain in which we face highly uncertain scenarios where the short-term and long-term consequences are often unclear. There are effective governance mechanisms for the pharmaceutical industry that could offer insights to the regulation and assurance of autonomous systems. (CP)
- The environment and medicine are highly uncertain, complex, dynamic domains where there are established benchmarks for dealing with uncertainty. What value is there to thinking that ML-systems, AI-systems, and autonomous systems are dramatically different in this respect? (AZ)
- The uncertainty is not what is unique. Normal cyber-physical systems have lots of uncertainty and there is nothing new here. Many learning algorithms are just feedback or adaptive systems. All the techniques that have been developed to manage and model uncertainty can be used to manage autonomous systems. Something else is unique: transfer of decision-making function to the system from the human. (MF)
- There are great frameworks for Responsible Research and Innovation already. There is an established precedent here in terms of broadening participatory and stakeholder engagement. (AW)

5. Next steps

Part of AAIP's global research strategy is to produce free, implementable, practical guidance for developers, assessors, policy makers, and regulators of robotics and autonomous systems. One strand of this will focus on the ethical dimensions of assurance. The workshop has been instrumental to us in our thinking about how best to frame and structure that guidance.

We would also like to consider practical ethical implications at the intersection of discussion question 1 and discussion question 3 in a more structured way.

We would welcome continuing the conversation.

6. Annex 1

Participant list and initials

(AB) Aleksandra Berditchevskaia, Nesta

(CB) Christopher Burr, Alan Turing Institute/Oxford Internet Institute

(SB) Simon Burton, Fraunhofer IKS

(MF) Michael Fisher, University of Manchester

(EK) Emre Kazim, University College London

(CK) Carly Kind, Ada Lovelace Institute

(IH) Ibrahim Habli, AAIP and University of York

(JH) Jamie Hodsdon, Centre for Connected and Autonomous Vehicles

(DL) David Leslie, Alan Turing Institute

(AM) Ana MacIntosh, AAIP and University of York

(FM) Farah Magrabi, Macquarie University

(JMcD) John McDermid, AAIP and University of York

(ZP) Zoë Porter, AAIP and University of York

(CP) Carina Prunkl, University of Oxford

(GR) Gopal Ramchurn, University of Southampton

(AR) Arizona Rodriguez, Royal Academy of Engineering

(AS) Alex Smyth, Royal Academy of Engineering

(AW) Alan Winfield, University of the West of England

(AZ) Annette Zimmermann, University of York/Harvard University

With thanks to Sarah Heathwood (AAIP) for her support in organising the workshop and producing this report.

7. References

British Standard - BS8611:2016 *Robots and Robotic Devices: Guide to The Ethical Design and Application Of Robots And Robotic Systems*

ICO/Alan Turing Institute co-badged guidance. 2020. *Explaining decisions made with artificial intelligence*

https://ico.org.uk/media/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-artificial-intelligence-1-0.pdf

CAHAI – Council of Europe's Ad hoc Committee on Artificial Intelligence. Feasibility study on a legal framework on AI design, development, and application, based on CoE standards, was adopted by the CAHAI on 17 December 2020.

https://rm.coe.int/cahai-2020-23-final-eng-feasibility-study-/1680a0c6da

Austin, J.L., 1975. *How to do things with words*. Oxford University Press. (p. 181)