Assuring Autonomy International Programme
Critical Barriers to Assurance and Regulation
Concept and Examples – July 2018

Introduction

The Assuring Autonomy International Programme is developing a Body of Knowledge (BoK) intended, in time, to become the definitive reference source on assurance and regulation of Robotics and Autonomous Systems (RAS). The scope of such a BoK is potentially vast, so the Programme is using the concept of 'critical barriers' that may preclude assurance or regulation of RAS as a way of guiding the development of the BoK. They will also be used as a way of focusing calls for 'demonstrator projects' – the Programme's primary means of interacting with the RAS development and regulatory community.

The purpose of this document is to give:

- A clear definition for the concept of 'critical barriers';
- Motivating examples of 'critical barriers' that will be used to guide the next call for demonstrator projects;
- A worked example of a summary for a demonstrator project, showing how the 'critical barrier' might be addressed, by developing principles and practices to overcome the barrier.

The first two points are addressed below; the worked example is included in Annex A.

Critical Barriers: Concept

A Critical Barrier to Assurance and Regulation (C-BAR) is a problem that must be solved for a particular system or domain, in order to avoid one or more of the following risks:

- A safe system cannot be deployed (losing the benefit of the technology);
- An unsafe system is deployed (lack of clear evidence to assure operation);
- The adoption of safe technology is slow;
- There is a lack of progress in adoption in a particular domain;
- The level of accidents and incidents leads to a backlash.

In practice, many C-BARs will be well-known problems, *e.g.* explanation of decisions made by Artificial Intelligence (AI). Some of them will have tentative, but currently unproven, solutions, and that makes them prime candidates for demonstrator projects.

Critical Barriers: Examples

The Programme will define and evolve a set of C-BARs. The intent is that the Programme's core staff will do this, with assistance from visiting Programme Fellows and in conjunction with other initiatives. At this stage, the set reflects expert judgement although it is intended that it will be put on a more systematic basis in future.

The current working set of C-BARs is shown below, noting that they may not be entirely disjoint. For ease of understanding the C-BARs are presented as questions[1]:

- **Adaptation** – where RAS adapt their behaviour in operation, *e.g.* through machine learning, how can it be assured that they are/continue to be safe, and what regulatory frameworks are necessary to enable risk to be accepted?
- **Bounding Behaviour** – where RAS can operate safely within known bounds, *e.g.* of visibility or adhesion on a road, how are these limits identified in design and in operation, and a safe transition achieved before reaching the limits, and assured?

[1] There is no significance to the order – it is simply alphabetical.

- **Cross-Domain Usage** – where a RAS (capability) is known to be effective in one domain, *e.g.* sense and avoid for mobile robots in a factory, how can it be assessed for adequacy in another environment, *e.g*. in a hospital (NB: this has similarities with the problem of safe reuse)?
- **Explanations**[2] – what decisions made by a RAS, *e.g.* object classification and path planning, need to be explained to users or regulators (as part of an acceptance or regulatory process), and how can this be done effectively given that the systems will make many decisions a second, in operation?
- **Handover** – if (semi-) autonomous systems have to hand (back) control to a human operator, how can it be ensured and assured that the operator has sufficient situational awareness to be able to take over control safely and effectively[3]?
- **Human-Robot Interaction**[4] – where humans and RAS unavoidably interact physically, and the RAS is sufficiently powerful/capable to cause harm, how can it be ensured and assured that the RAS does not injure humans it interacts with?
- **Incident and Accident Investigation** – if a potentially harmful incident or an accident occurs, what information needs to be provided to support investigation, and how is this achieved and enforced in regulatory frameworks, noting that it may require gathering information from systems not directly involved in the incident or accident?
- **Monitoring**[5] – where operators are required to monitor RAS to ensure that the system is operating as expected and/or safely, how can it be ensured and assured that they retain sufficient levels of attention and concentration, or what bounds can be put on the monitoring function to ensure that it will be undertaken effectively?
- **Risk Acceptance** – as RAS are likely to significantly modify the risk-benefit balance across many domains, and the effects of autonomy (especially some aspects of AI) exacerbate the intrinsic uncertainty in assessing risk, how can risk be estimated, communicated and accepted by both the regulatory community and the public?
- **Role of Simulation** – as many RAS cannot be tested in real operational environments prior to their use, how can simulation be used to greatest effect to enable assurance and regulation, and when does simulation provide sufficient evidence (in itself or in combination with other means of verification and validation (V&V)) to allow controlled use of the RAS?
- **Systems of Systems** – where RAS are elements of systems of systems (either with other RAS or 'manually' controlled systems) that are known to be 'individually safe' how can safe interaction be assured, in their intended operational environment?
- **Training and Testing AI** – when RAS use machine learning how can it be shown that the training sets (and test sets) give enough coverage of the environment to provide sufficient evidence (in itself or in combination with other means of V&V) to allow controlled use of the RAS?
- **Validation** – how can we identify effective means of validating RAS especially their AI components, *e.g.* using simulation, hazard analysis, *etc.*, and are there effective coverage measures of the environment to allow controlled use of the RAS?
- **Verification** - how can we identify effective means of verifying RAS especially their AI components, *e.g.* using testing, formal verification, *etc.*, and are there effective coverage measures of the learnt decision space to allow controlled use of the RAS?

Note that the way 'critical barriers' are defined above assumes that, in some cases, RAS are not approved for operation because of lack of 'solutions' for the barriers and, in other cases, systems are approved because

[2] The General Data Protection Regulations give people affected by decisions made using AI a "right to explanation; the Information Commissioner's Office indicate that this right would apply to RAS.
[3] This C-BAR is perhaps most apparent with autonomous vehicles, where times of circa 20 seconds are quoted for regaining situational awareness, but it is a more generic problem.
[4] The example in Annex A is a slightly more specific version of this issue
[5] The accident in Tempe Arizona illustrates this problem, however it is more generic and some analyses suggest that the limit on effective monitoring is around 10 minutes.

although there is no agreed way of assessing the systems, there are no grounds for rejecting them within the regulatory framework.

Annex A: Worked Example – Addressing a C-BAR for Assurance of Human-Robot-Interaction in Social Care

A demonstrator uses humanoid robots to care for the elderly and infirm in a domestic environment. These robots will interact physically with people, enabling them to stand and walk, or to carry them, see Figure 1. The robots are capable of lifting weights up to 100kg, thus they can pose a risk of harm to the individual.

Figure 1: Robot lifting a teenager

The C-BAR is:

- **Human-Robot Interaction**: How can we demonstrate that a robot cannot interact with the human they are supporting in a way that will cause harm, including bruising?

The principles to be adopted are:

- The robot assesses manoeuvres, and doesn't attempt them if unsafe, *e.g.* a lift would interact with a bandaged limb;
- The force applied is distributed, so pressure never exceeds a given threshold;
- Impact velocity is always below a specified threshold.

The practices to be explored and validated are:

- Use of simulation to compute maximum forces in challenging manoeuvres, for representative physiological types;
- Validation of simulation through use of dummies with pressure and impact sensors;
- Trials with humans, including "emergency" response, *e.g.* tripping, near fall.

The demonstrator would deliver a method for assessing safe physical interaction with humans, and a data set from the trials that might assist other projects.