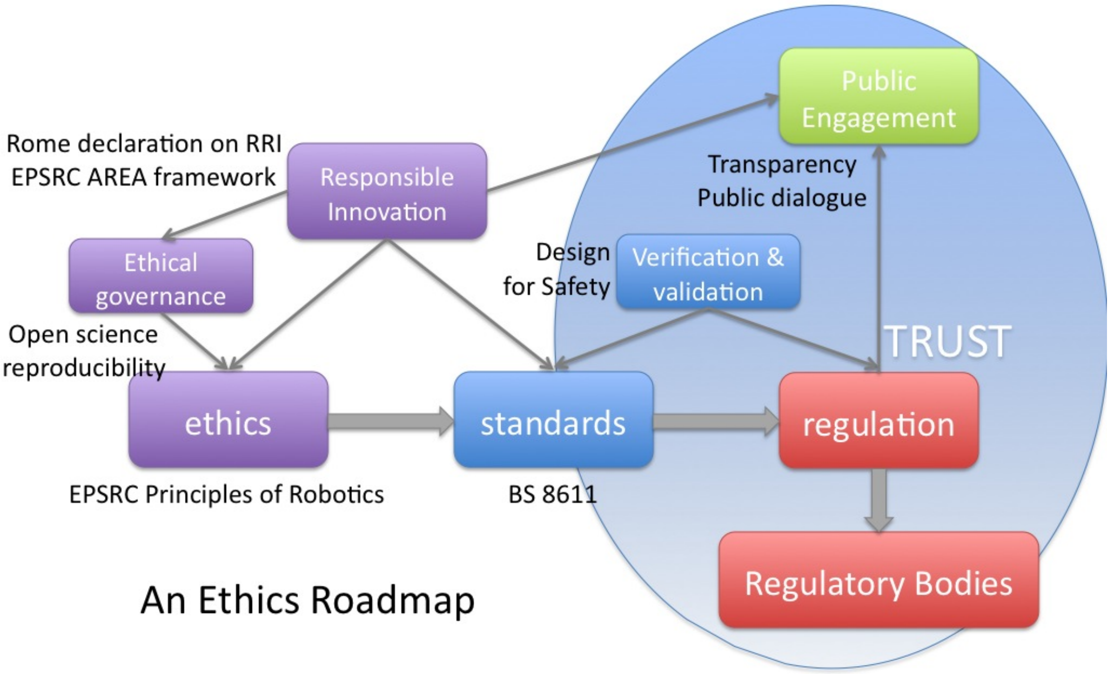
Written evidence submitted by Professor Alan Winfield (ROB0070)

From Ethics to Regulation and Governance

The following text was drafted in response to question 4 of the Parliamentary Science and Technology Committee Inquiry on Robotics and Artificial Intelligence on *The social, legal and ethical issues raised by developments in robotics and artificial intelligence technologies, and how they should be addressed.*

1. **Public attitudes**. It is well understood that there are public fears around robotics and artificial intelligence. Many of these fears are undoubtedly misplaced, fuelled perhaps by press and media hype, but some are grounded in genuine worries over how the technology might impact, for instance, jobs or privacy. The most recent Eurobarometer survey on autonomous systems showed that the proportion of respondents with an overall positive attitude has declined from 70% in the 2012 [1] survey to 64% in 2014 [2]. Notably the 2014 survey showed that the more personal experience people have with robots, the more favourably they tend to think of them; 82% of respondents have a positive view of robots if they have experience with them, where as only 60% of respondents have a positive view if they lack robot experience. Also important is that a significant majority (89%) believe that autonomous systems are a form of technology that requires careful management.

2. **Building trust** in robotics and artificial intelligence requires a multi-faceted approach. The ethics roadmap below illustrates the key elements that contribute to building public trust. The core idea of the roadmap is that ethics inform standards, which in turn underpin regulation.



3. **Ethics** are the foundation of trust, and underpin good practice. Principles of good practice can be found in **Responsible Research and Innovation** (RRI). Examples include the 2014 Rome Declaration on RRI [3]; the six pillars of the Rome declaration on RRI are: Engagement, Gender equality, Education, Ethics, Open Access and Governance. The EPSRC framework for responsible innovation [4] incorporates the AREA (Anticipate, Reflect, Engage and Act) approach.

4. The first European work to articulate ethical considerations for robotics was the EURON Roboethics Roadmap [5].

5. In 2010 a joint AHRC/EPSRC workshop drafted and published a set of five **Principles of Robotics** for designers, builders and users of robots [6]. The principles are:

(i) Robots are multi-use tools. Robots should not be designed solely or primarily to kill or harm humans, except in the interests of national security;

(ii) Humans, not robots, are responsible agents. Robots should be designed; operated as far as is practicable to comply with existing laws & fundamental rights & freedoms, including privacy.

(iii) Robots are products. They should be designed using processes which assure their safety and security.

(iv) Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent.

(v) The person with legal responsibility for a robot should be attributed.

6. Work by the British Standards Institute technical subcommittee on Robots and Robotic Devices led to publication – in April 2016 – of **BS 8611**: *Guide to the ethical design and application of robots and robotic systems* [7]. BS8611 is not a code of practice; instead it gives "guidance on the identification of potential ethical harm and provides guidelines on safe design, protective measures and information for the design and application of robots". BS8611 articulates a broad range of ethical hazards and their mitigation, including societal, application, commercial/financial and environment risks, and provides designers with guidance on how to assess then reduce the risks associated with these ethical hazards. The societal hazards include, for example, loss of trust, deception, privacy & confidentiality, addiction and employment.

7. Significant recent work towards **regulation** was undertaken by the EU project RoboLaw. The primary output of that project is a thorough report entitled *Guidelines on Regulating Robotics* [8]. That report reviews both ethical and legal aspects; the legal analysis covers rights, liability & insurance, privacy and legal capacity. The report focuses on driverless cars, surgical robots, robot prostheses and care robots and concludes by stating: "The field of robotics is too broad, and the range of legislative domains affected by robotics too wide, to be able to say that robotics by and large can be accommodated within existing legal frameworks or rather require a *lex robotica*. For some types of applications and some regulatory domains, it might be useful to consider creating new, fine-grained rules that are specifically tailored to the robotics at issue, while for types of robotics, and for many regulatory fields, robotics can likely be regulated well by smart adaptation of existing laws".

8. In general technology is **trusted** if it brings benefits while also safe, well regulated and, when accidents happen, subject to robust investigation. One of the reasons we trust airliners is that we know they are part of a highly regulated industry with an excellent safety record. The reason commercial aircraft are so safe is not just good design, it is also the tough safety certification processes and, when things do go wrong, robust processes of air accident investigation. Should driverless cars, for instance, be regulated through a body similar to the Civil Aviation Authority (CAA), with a driverless car equivalent of the Air Accident Investigation Branch?

9. The primary focus of paragraphs 1 – 8 above is robotics and autonomous systems, and not **software artificial intelligence**. This reflects the fact that most work toward ethics and regulation has focussed on robotics. Because robots are physical artefacts (which embody AI) they are undoubtedly more readily defined and hence regulated than distributed or cloud-based AIs. This and the already pervasive applications of AI (in search engines, machine translation systems or intelligent personal assistant AIs, for example) strongly suggest that greater urgency needs to be directed toward considering the societal and ethical impact of AI.

The IEEE has recently launched a global initiative on Ethical Considerations in the Design of Autonomous Systems, to encompass all intelligent technologies including AI, computational intelligence and deep learning [10].

10. AI systems raise serious questions over **trust** and **transparency**:

- How can we trust the decisions made by AI systems, and – more generally – how can the public have confidence in the use of AI systems in decision making?
- If an AI system makes a decision that turns out to be disastrously wrong, how do we investigate the logic by which the decision was made?

Of course much depends of the consequences of those decisions. Consider decisions that have real consequences to human safety or well being, such as medical diagnosis AIs, or driverless car autopilots. Systems that make such decisions are *critical* systems.

11. Existing critical software systems are not AI systems, nor do they incorporate AI systems. The reason is that AI systems (and more generally machine learning systems) are generally regarded as impossible to verify for safety critical applications - the reasons for this need to be understood.

- First is the problem of **verification of systems that learn**. Current verification approaches typically assume that the system being verified will never change its behaviour, but a system that learns does – by definition – change its behaviour, so any verification is rendered invalid after the system has learned.
- Second is **the black box problem**. Modern AI systems, and especially the ones receiving the greatest attention, so called Deep Learning systems, are based on Artificial Neural Networks (ANNs). A characteristic of ANNs is that after the ANN has been trained with data sets (which may be very large, so called "big data" sets – which itself poses another problem for verification), any attempt to examine the internal structure of the ANN in order to understand why and how the ANN makes a particular decision is impossible. The decision making process of an ANN is not transparent.

The problem of **verification and validation** of systems that learn may not be intractable, but is the subject of current research, see for example [11]. The black box problem may be intractable for ANNs, but could be avoided by using algorithmic approaches to AI (i.e. that do not use ANNs).

Recommendations

12. It is vital that we address public fears around robotics and artificial intelligence, through **renewed public engagement and consultation**.

13. Work is required to **identify the kind of governance framework(s)** and regulatory bodies needed to support Robotics and Artificial Intelligence in the UK. A group should be set up and charged with this work; perhaps a **Royal Commission**, as recently suggest by Tom Watson MP [9]

References:

[1] Special Eurobarometer 427, Autonomous Systems, June 2015.
http://ec.europa.eu/public_opinion/archives/ebs/ebs_427_en.pdf See also analysis summary http://robohub.org/study-shows-public-perception-of-robotics-generally-positive-in-eu-but-declining/

[2] Special Eurobarometer 382, Public Attitudes towards Robots, Sept 2012.
http://ec.europa.eu/COMMFrontOffice/PublicOpinion/index.cfm/ResultDoc/download/DocumentKy/56814

[3] The Rome Declaration on Responsible Research and Innovation (2014): http://www.science-and-you.com/en/sis-rri-conference-recommendations-rome-declaration-responsible-research-and-innovation and http://ec.europa.eu/research/science-society/document_library/pdf_06/responsible-research-and-innovation-leaflet_en.pdf

[4] The EPSRC framework for responsible Innovation: https://www.epsrc.ac.uk/research/framework/

[5] Veruggio G (2006), EURON Roboethics Roadmap: http://www.roboethics.org/atelier2006/docs/ROBOETHICS%20ROADMAP%20Rel2.1.1.pdf

[6] Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., Newman, P., Parry, V., Pegman, G., Rodden, T., Sorrell, T., Wallis, M., Whitby, B. and Winfield, A. Principles of Robotics, EPSRC, May 2011. https://www.epsrc.ac.uk/research/ourportfolio/themes/engineering/activities/principlesofrobotics/

[7] British Standards Institute BS8611:2016 Robots and robotic devices, Guide to the ethical design and application of robots and robotic systems, ISBN 978 0 580 89530 2.

[8] Palmerini, E., Azzarri, F., Battaglia, A., Bertolini, A., Carnevale, A., Carpaneto, J., Cavallo, F., Di Carlo, A., Cempini, M., Controzzi, M., Koops, B.J., Lucivero, F., Mukerji, N., Nocco, L., Pirni, A., Shah, H., Salvini, P., Schellekens, M. and Warwick, K. D6.2 – Guidelines on regulating robotics. Pisa, Italy: Robolaw project, 2014. http://www.robolaw.eu/RoboLaw_files/documents/robolaw_d6.2_guidelinesregulatingrobotics_20140922.pdf

[9] Stewart H (2016), Government urged to investigate impact of robots on UK workforce, 8 March 2016. http://www.theguardian.com/commentisfree/2016/mar/08/robots-technology-industrial-strategy

[10] The global initiative on Ethical Considerations in the Design of Autonomous Systems http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html
Note, Winfield co-chairs the General Principles committee.

[11] EPSRC project Verifiable Autonomy: http://gow.epsrc.ac.uk/NGBOViewGrant.aspx?GrantRef=EP/L024845/1 and the EPSRC Network on the Verification and Validation of Autonomous Systems: http://gow.epsrc.ac.uk/NGBOViewGrant.aspx?GrantRef=EP/M027309/1

May 2016