The Case for an Ethical Black Box

Alan F.T. Winfield[1]* and Marina Jirotka[2]

[1] Bristol Robotics Lab, University of the West of England, Bristol
[2] Department of Computer Science, University of Oxford

Abstract. This paper proposes that robots and autonomous systems should be equipped with the equivalent of a Flight Data Recorder to continuously record sensor and relevant internal status data. We call this an ethical black box. We argue that an ethical black box will be critical to the process of discovering why and how a robot caused an accident, and thus an essential part of establishing accountability and responsibility. We also argue that without the transparency afforded by an ethical black box, robots and autonomous systems are unlikely to win public trust.

Keywords: robot ethics, ethical governance, traceability, transparency, responsible robotics, trust

1 Introduction

Driverless car accidents are headline news. The fatal Tesla accident of May 2016 [22] resulted in considerable press and media speculation on the causes of the accident. But there is a worrying perception that transparency is lacking in how these events are disclosed. System developers speak reassuringly but we may suspect they have a vested interest in giving events a positive gloss. This raises the crucial question of how the transparency of robot AIs can be guaranteed so as to avoid publics becoming fearful and to ensure that robots gain high levels of trust and acceptance in society [10]. Whilst our existing concepts of accountability and liability are being stretched by semi-autonomous machines, new heights of machine autonomy are likely to shatter them completely [27], giving urgency to the search for suitable revisions.

In this paper we propose that robots and autonomous systems should be equipped with the equivalent of an aircraft Flight Data Recorder to continuously record sensor and relevant internal status data. We call this an ethical black box. We argue that an ethical black box will play a key role in the processes of discovering why and how a robot caused an accident, and thus an essential part of establishing accountability and responsibility. We also argue that without the transparency afforded by an ethical black box, robots and autonomous systems are unlikely to win public trust.

This paper is structured as follows. In section 2 we outline the development and practice of flight data recorders. Then in section 3 we make the link between

* alan.winfield@uwe.ac.uk

ethical governance, transparency and trust. Section 4 discusses concerns over transparency in safety-critical AI to propose an ethical black box, and suggest a generic specification. In section 5 we consider the human processes of robot accident investigation.

2 Black box flight data recorders

Black box or flight data recorders were introduced in 1958, for larger aircraft, and since then have vastly expanded in scope in what flight data they record. Initial flight data recorders included time navigation data position of surfaces and pilots movement of controls; latterly sensor data about the internal and external environment as well as the functioning of components and systems are also included, alongside intentional autopilot setting such as selected headings, speeds, altitudes and so on [16]. The first black boxes recorded 5 flight parameters, data recorders on modern aircraft record more than 1000 parameters [8]. A significant innovation was the introduction of the Cockpit Voice Recorder (CVR) to capture the conversation of the pilots to aid interpretation of flight data. Although initially resisted [16], conversations captured by the CVR have proved an invaluable tool in reconstructing the circumstances of an accident by providing an intentional context from which to interpret flight data. Air accident investigations help to rule out systematic failure modes, help to preserve trust in aviation, provide accountability and produce lessons that contribute to overall levels of safety.

Factors that typically contribute to air accidents have parallels to incidents that may in the future arise through our increased dependence on robots. Air accidents are typically the result of unpredicted interactions between the intentions and actions of the aircrew, the integrity and behaviour of the aircraft systems, and the environmental circumstances of the flight. Somewhat analogously we anticipate future robots to be subject to a similar mix of factors. As we expect robots to do more for us, they will necessarily become more sophisticated, operate more autonomously, and do so within open, unconstrained settings. These greater freedoms come with the increased risk of unanticipated combinations of unforeseen factors leading to hazardous situations or actually resulting in harm. This is not to make a judgement about how frequently such events might occur, as with air disasters, significant harm may be rare. But we need to acknowledge hazardous events will inevitability take place. So our very ambition for robots to contribute to human activities in new ways implies giving robots new capabilities and freedoms, which in turn creates the potential for robots to be implicated in significant hazard and injury. This may prompt an overall assessment of the benefits versus the harms of expanding our dependency on robots, but our view is that more than this is needed and that a further analogy with aviation is also relevant. We suggest that acceptance of air travel, despite its catastrophes, is in part bound up with aviation governance, which has cultural and symbolic importance as well as practical outcomes. A crucial aspect of the former is ren-

dering the tragedy of disaster comprehensible through process of investigation and reconstruction.

The transfer of the black box concept into settings other than aviation is not novel. It has been mooted for software [9] and micro-controllers [7]. The largest deployment of black box technology outside aviation is probably within the automobile and road haulage industries [25, 20, 33].

3 Ethical Governance and Trust

Ethics, standards and regulation are connected. Standards formalise ethical principles (i.e. [4]) into a structure which could be used either to evaluate the level of compliance or, more usefully perhaps for ethical standards, to provide guidelines for designers on how to conduct an ethical risk assessment for a given robot and mitigate the risks so identified [6]. Ethics therefore underpin standards. But standards also sometimes need teeth, i.e. regulation which mandates that systems are certified as compliant with standards, or parts of standards. Thus ethics (or ethical principles) are linked to standards.

Although much existing law applies to robots and autonomous systems [23], there is little doubt that rapidly emerging and highly disruptive technologies such as drones, driverless cars and assistive robotics require – if not new law – then regulation and regulatory bodies at the very least.

Ethics and Standards both fit within a wider framework of Responsible Research and Innovation (RRI) [24, 14]. RRI provides a scaffold for ethics and standards, as shown in Figure 1. Responsible Innovation typically requires that research is conducted ethically, so ethical governance connects RRI with ethics. RRI also connects directly with ethics through principles of, for instance, public engagement, open science and inclusivity. Another key principle of RRI is the ability to systematically and transparently measure and compare system capabilities, typically with standardised tests or benchmarks [13].

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, for example, 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 and publicly visible processes of air accident investigation.

Regulation requires regulatory bodies, linked with public engagement [29, 28] to provide transparency and confidence in the robustness of regulatory processes. All of which supports the process of building public trust, as shown in Figure 1.

Trust does not, however, always follow from (suggested) regulation. A recent survey of decision making in driverless cars reveals ambivalent attitudes to public trust in driverless cars [5] "... participants approved of utilitarian Autonomous Vehicles (AVs) (that is, AVs that sacrifice their passengers for the greater good) and would like others to buy them, but they would themselves prefer to ride in AVs that protect their passengers at all costs. The study participants disapprove

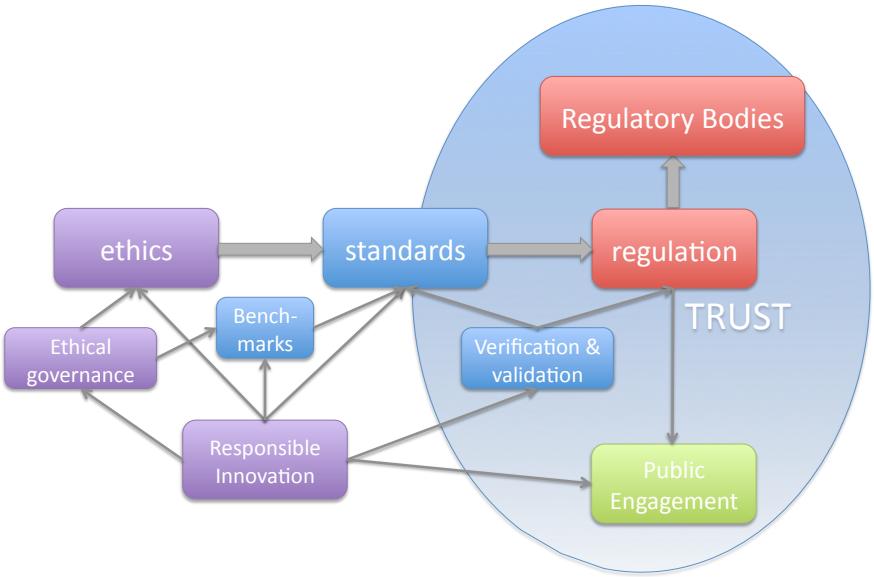


Figure 1. Building Public Trust, from [32]

of enforcing utilitarian regulations for AVs and would be less willing to buy such an AV".

4 Safety Critical Artificial Intelligence and Transparency

All machines, including robots, have the potential to cause harm. Responsibly designed present day robots are designed to be safe and to avoid unintended harms, either accidentally or from deliberate misuse, see i.e. [18] for personal care robots. Such robots are implicit ethical agents [21].

The primary focus of this paper is robotics and autonomous systems, and not software artificial intelligence. However, a reasonable definition of a modern robot is 'an embodied AI' [30], thus in considering the safety of robots we must also concern ourselves with the AI controlling the robot. Three important classes of robot are drones, driverless cars and assistive robots (including care or workplace assistant robots); all will be controlled by an embedded AI[3], of some appropriate degree of sophistication. Yet these are all *safety critical* systems, the safety of which is fundamentally dependent on those embedded AIs; decisions made by which have real consequences to human safety or well being. Let us consider the issue of transparency, in particular two questions:

[3] Perhaps augmented by a cloud based AI

1. 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? In [32] we argue that ethical governance is a necessary, but (probably) not sufficient, element of building public trust.
2. 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? This question essentially underlies the case for an ethical black box.

Transparency will necessarily mean different things to different stakeholders – the transparency required by a safety certification agency or an accident investigator will clearly need to be different to that required by the system's user or operator. But an important underlying principle is that it should always be possible to find out why an autonomous systems made a particular decision (most especially if that decision has caused harm). A technology that would provide such transparency, especially to accident investigators, would be the equivalent of an aircraft flight data recorder (FDR). We call this an *ethical black box*, both because aircraft FDRs are commonly referred to as black boxes, and because such a device would be an integral and essential physical component supporting the ethical governance of robots and robotic systems. Like its aviation counterpart the ethical black box would continuously record sensor and relevant internal status data so as to greatly facilitate (although not guarantee) the discovery of why a robot made a particular decision or series of decisions – especially those leading up to an accident. Ethical black boxes would need to be designed and certified according to standard industry-wide specifications, although it is most likely that each class of robot would have a different standard; one type for driverless vehicles, one type for drones and so on.

4.1 An outline specification for an Ethical Black Box

All robots collect sense data, then – on the basis of that sense data and some internal decision making process (AI) – send commands to actuators. This is of course a simplification of what in practice will be a complex set of connected systems and processes but, at an abstract level, all intelligent robots will have the three major subsystems shown in blue, in Fig. 2. If we consider a driverless car, its sensors typically consist of a Light Detection and Ranging sensor (LIDAR), camera, and a number of short range collision sensors together with GPS, environmental sensors for rain, ambient temperature etc, and internal system sensors for fuel level, engine temperature, etc. The actuators include the car's steering, accelerator and braking systems.

The Ethical Black Box (EBB) and its data flows, shown in red in Fig. 2 will need to collect and store data from all three robot subsystems. From the sensors the EBB will need to collect either sampled or compressed raw data, or perhaps features that have been extracted by the sensor subsystem's post-processing. From the actuator system the EBB will need to collect actuator demands (i.e. steer left 10 degrees). From the AI system the EBB will need as a minimum high level 'state' data, such as 'braking', 'steering left', 'parking' etc and, ideally, also

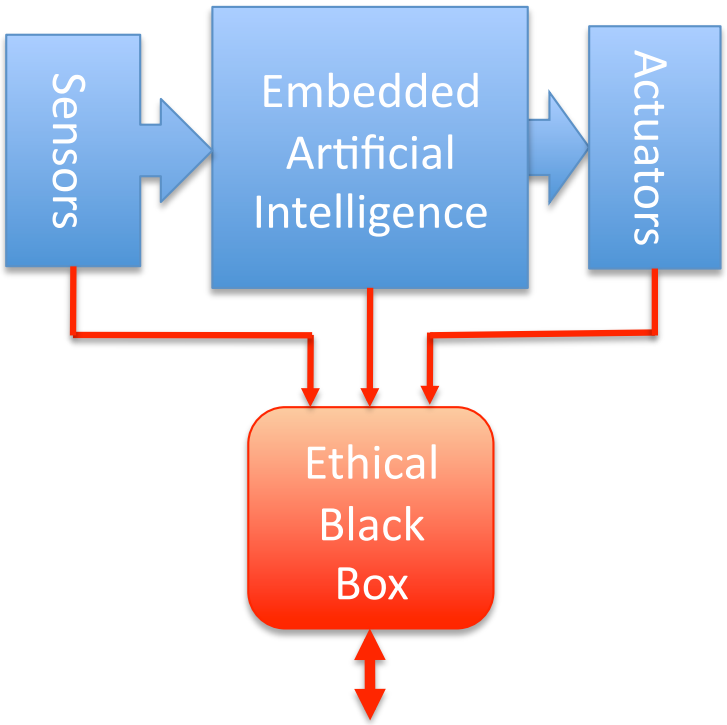


Figure 2. Robot with Ethical Black Box and key dataflows

high level goals such as 'turning left at junction ...' and alerts such as 'cyclist detected front left - taking avoiding action'. All of these data will need to be date and time stamped, alongside location data from the GPS.

How much data could be stored in the EBB? Some reports have suggested that Google's driverless car generates 1GByte of raw data per second[4]. If we (reasonably) assume that we can sample and/or compress this to say 100MByte per second, then an EBB equipped with a 1TByte solid-state drive would allow it to continuously log data for about 3 hours of operation. The EBB would, like an aircraft flight data recorder, continuously overwrite the oldest data logs, so that at any one time the EBB stores the most recent 3 hours.

It is beyond the scope of this paper to specify exactly which data will need to be recorded in the EBB. We can however be clear about what those data

[4] `http://www.kurzweilai.net/googles-self-driving-car-gathers-nearly-1-gbsec`

are for: the key principle is that, from the data recorded in the EBB, it must be possible to *reconstruct* the timeline leading up to and during an accident; a timeline annotated with the key sensory inputs, actuator demands, and the high-level goals, alerts and decisions that drove those actuator demands. A full EBB specification will need to set out which data is to be recorded and, equally important, the specification of the interface(s) between the robot subsystems and the EBB. The interface specification must include the hardware (connectors), signaling and protocols Note that the dataflows between robot subsystems and the EBB are one way only - the EBB must, as far as the robot is concerned, be an entirely passive subsystem.

The EBB specification must also cover the physical and data interface which allows an accident investigator to access data logged by the EBB (shown at the bottom of Fig, 2). There are many other aspects of the EBB that will need to be specified. These include the physical form of the EBB noting that, like its aviation counterpart, it will need to be rugged enough to survive a very serious accident. Some would argue that the EBB does not have to be a physical component at all (see, for instance [19]) and that instead all data should be streamed to secure cloud storage. We would however argue that the high volumes of data (especially in an environment with a large number of, for instance, driverless cars) could overwhelm the local wireless infrastructure. The EBB will of course also need to be secure and tamper-proof.

Note that as specified here the EBB does not record the internal low-level decision making processes of the embedded AI (just as an aircraft FDR does not record the low-level processes within the vehicle's autopilot). The reason for this is that different robot AIs are likely to have very different internal architectures: some may be algorithmic, others based on artificial neural networks. Finding a common specification for the low-level decision making processes for all AIs is therefore likely to be impossible. Instead the transparency argued for in this paper is not achieved by the EBB alone but through the processes of accident investigation, as outlined in section 5 below.

4.2 Ethical Black Boxes for Moral Machines

This paper makes the case that all safety critical robots should be equipped with an ethical black box. Consider now robots that are explicit moral agents. It is clear that some near future autonomous systems, most notably driverless cars are *by default* moral agents. Both driverless cars and assistive (i.e. care) robots make decisions with ethical consequences, even if those robots have not been designed to explicitly embed ethical values and moderate their choices according to those values. Arguably all autonomous system implicitly reflect the values of their designers or, even more worryingly, training data sets (as dramatically shown in AI systems that demonstrate human biases [11]).

There is a growing consensus that near future robots will, as a minimum, need to be designed to explicitly reflect the ethical and cultural norms of their users and societies [6, 17]. Beyond reflecting values in their design a logical (but difficult) next step is to provide robots with an *ethical governor*. That is, a

process which allows a robot to evaluate the consequences of its (or others') actions and modify its own actions according to a set of ethical rules. Developing practical ethical governors remains the subject of basic research and presents two high level challenges: (1) the philosophical problem of the formalisation of ethics in a format that lends itself to machine implementation and (2) the engineering problem of the implementation of moral reasoning in autonomous systems [15].

There are two approaches to addressing the second of these challenges [1]:

1. a constraint-based approach – explicitly constraining the actions of an AI system in accordance with moral norms; and
2. a training approach – training the AI system to recognise and correctly respond to morally challenging situations.

The training approach is developed for an assistive robot in [2], while examples of constraint-based approaches are explored in [3, 31]. One advantage of the constraint-based approach is that it lends itself to verification [12].

Extending the EBB outlined in this paper for a robot equipped with an ethical governor would in principle be straightforward: the EBB would need to additionally log the decisions made by the ethical governor so that an accident investigator could take those into account in building a picture of the processes that led to the accident.

5 The Processes of Robot Accident Investigation

In aviation it is the investigation, not the black box data per se, which concludes why an air accident occurred. We anticipate this will also be true for accidents involving robots, where an investigation will draw upon EBB data amongst other information to determine the reason for an accident. So alongside the technical parameters of what to record within the EBB, we have also to consider how the interpretation of those data fits into the process of an investigation. Air accident investigations are social processes of reconstruction that need to be perceived as impartial and robust, and which (we argue) serve as a form of closure so that aviation does not acquire an enduring taint in the publics consciousness. We anticipate very similar roles for investigations into robot accidents.

Taking the example of driverless cars, an accident investigation would bring together data and evidence from a variety of sources. The extant local context will undoubtedly play a part in tracing the causes of the accident. While the EBB is witness to events that often would otherwise remain unwitnessed, the activities of driverless cars are likely to take place in populated spaces where there may be many witnesses, some of whom will record, share and publish details of the event via mobile phones and other devices, creating multiple perspectives on what may have happened. Thus, there may be bi-standers, pedestrians, passengers or other drivers who will have particular views on the event and may have captured the accident as it unfolds in a variety of ways. Traditional police forensics will investigate the scene in the conventional way examining evidence such as, skid marks in the road or the impact of the crash on other objects (walls, cars) or

people (for an example of the analysis of crash data see [33]). This raises the question of how the interpretation of EBB data sit alongside the interpretation of evidence from other witnesses and brings to the fore consideration of the epistemological status of different types of witnessing.

As a further key part of the interdependent network of responsibilities in this case, the car's manufacturer and possibly also maintainer are also likely to be called to provide input. This will be particularly important if initial investigation points to some incorrect internal low-level decision making process in the car's AI as the likely cause of the accident. Indeed companies may be required to release the data sets used to train the algorithms that are driving these systems when accidents occur; a requirement that may conflict with corporations desire to gain competitive advantage from such data.

An obvious concern is how conclusions might be reached where the data from these different types of witnessing is in conflict. It is clear that a specialist team with deep expertise in driverless car technology, perhaps within the umbrella of the road traffic investigation agency, will be needed to carefully weigh these data and reach conclusions about why the driverless car behaved the way it did, and make recommendations. But without doubt the data provided by the EBB sits at the very centre of the process of accident investigation, providing the crucial and objective timeline against which all other witness accounts can be superposed.

6 Concluding discussion

The recent Tesla crash report blames human error, often cited as a primary cause or contributing factor in disasters and accidents as diverse as nuclear power, space exploration and medicine. But in this case, we echo Jack Stilgoe's statement in the Guardian [26] that this conclusion and the process through which it was reached is a missed opportunity both to learn from such incidents about the systemic properties of such autonomous systems in the wild, and also to initiate a more transparent and accountable process for accident investigation. As reported, after the US National Highways and Transport Safety Agency (NHTSA) published their report, much of the media coverage concentrated upon the fact that no products were recalled thereby prioritising the business value for Tesla. But as Stilgoe recounts, "As new technologies emerge into the world, it is vital for governments to open them up, look at their moving parts and decide how to realise their potential while guarding against their risks." Human error is a common refrain and yet, "To blame individual users of technology is to overlook the systemic issues that, with aeroplanes, have forced huge safety improvements. When an aeroplane crashes, regulators' priority must be to gather data and learn from it. User error should be seen as a function of poorly-designed systems rather than human idiocy."

Responsibility within RRI is framed as a collective accomplishment which seems well matched to the interdependent systemic properties in the design, development and use of robots and autonomous systems. Whilst the EBB we

propose in this paper seems to focus on determining accountability once an accident has happened, it is actually an outcome of reflections on applying anticipatory governance to these robot technologies. Accidents will be inevitable. Within that, it is vital to consider what bodies, processes and stakeholders need to be in place to determine a just account of the reasons for the accident, drawing upon the EBB as one piece of evidence amongst others. Whilst the process itself may seem simple, it will be complicated by local contexts and the nature of the accident, political climates, legal actions, international differences and corporate concerns.

Acknowledgments

This work has, in part, been supported by EPSRC grant ref EP/L024861/1.

References

1. Allen, C., Smit, I., and Wallach, W., (2005) Artificial morality: Top-down, bottom-up, and hybrid approaches. Ethics and Information Technology 7: 149-155, Elsevier.
2. Anderson, M. and Anderson, S.L. (2014) GenEth: A General Ethical Dilemma Analyzer, in Proc. AAAI, pp.253-261.
3. Arkin, R.C., Ulam, P., and Wagner, A.R. (2012) Moral Decision Making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust, and Deception. Proc. IEEE 100 (3): 571589.
4. 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.F. (2017), Principles of Robotics, Connection Science, In Press.
5. Bonnefon, J-F., Shariff, A., and Rahwam, I. (2016) The social dilemma of autonomous vehicles, Science Vol 352, Issue 6293, pp. 1573-1576.
6. British Standards Institute (2016), BS8611:2016 Robots and robotic devices: guide to the ethical design and application of robots and robotic systems, ISBN 9780580895302, BSI. London.
7. Choudhuri, S. and Givargis, T. (2009). FlashBox: a system for logging non-deterministic events in deployed embedded systems. In Proc. of 2009 ACM symposium on Applied Computing (SAC '09). ACM, New York, NY, USA, 1676-1682.
8. Campbell, N. (2007), The Evolution of Flight Data Analysis, ASASI Proceedings, 2007. `http://asasi.org/papers/2007/The_Evolution_of_Flight_Data_Analysis_Neil_Campbell.pdf`
9. Elbaum, S.; Munson, J.C. (2000), Software Black Box: an alternative mechanism for failure analysis, Proc. 11th Int. Symposium on Software Reliability Engineering, ISSRE 2000, pp.365 - 376.
10. Hibbard, B. (2014). Ethical Artificial Intelligence. `http://arxiv.org/abs/1411.1373`
11. Caliskan-Islam, A., Bryson, J. and Narayanan, A (2016) Semantics derived automatically from language corpora necessarily contain human biases, arXiv:1608.07187v2.
12. Dennis, L.A., Fisher, M., Slavkovik, M., and Webster, M. (2016): Formal Verification of Ethical Choices in Autonomous Systems. Robotics and Autonomous Systems 77: 1-14, Elsevier.

13. Dillmann, R. (2004) Benchmarks for Robotics Research, EURON, April 2004. http://www.cas.kth.se/euron/euron-deliverables/ka1-10-benchmarking.pdf
14. EPSRC (2016), The EPSRC framework for responsible Innovation. website https://www.epsrc.ac.uk/research/framework/.
15. Fisher, M., List, C., Slavkovik, M., Winfield, A.F. (2016), Engineering Moral Machines, Informatik-Spektrum, Springer.
16. Grossi, D.R. (2006). Aviation recorder overview [Electronic Version]. National Transportation Safety Board [NTSB] Journal of Accident Investigation, 2 (1), 31-42.
17. IEEE (2016), Global initiative on Ethical Considerations in the Design of Artificial Intelligence and Autonomous Systems, website http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html, visited 8 December 2016.
18. ISO 13482:2014 Robots and robotic devices Safety requirements for personal care robots, http://www.iso.org/iso/catalogue_detail.htm?csnumber=53820
19. Kavi, K.M. (2010). Beyond the Black Box. IEEE Spectrum. Posted 30 Jul 2010, http://spectrum.ieee.org/aerospace/aviation/beyond-the-black-box
20. Menig, P., et al (2010), Transportation Recorders on Commercial Vehicles, Int. Symposium on Transportation, retrieved from http://www.ntsb.gov/events/symprec/proceedings/authors/menig.pdf on Nov. 19, 2010 (14 pages).
21. Moor, J.H. (2006), The Nature, Importance and Difficulty of Machine Ethics, IEEE Intelligent Systems, 21(4), 18-21.
22. Vlasic, B. and Boudette, N.E. (2016) Self-Driving Tesla Was Involved in Fatal Crash, U.S. Says, New York Times, 30 June 2016.
23. 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, website http://www.robolaw.eu/RoboLaw_files/documents/robolaw_d6.2_guidelinesregulatingrobotics_20140922.pdf, visited 8 December 2016.
24. The Rome Declaration on Responsible Research and Innovation (2014), website: http://www.science-and-you.com/en/sis-rri-conference-recommendations-rome-declaration-responsible-research-and-innovation
25. Thom, P.R. and MacCarley, C.A. (2008). A Spy Under the Hood: Controlling Risk and Automotive EDR, Risk Management Magazine 55.2: 22-26. Available at: http://works.bepress.com/amaccarl/9
26. Stilgoe, J. (2017), Tesla crash report blames human error - this is a missed opportunity, The Guardian, 21 January 2017. https://www.theguardian.com/science/political-science/2017/jan/21/tesla-crash-report-blames-human-error-this-is-a-missed-opportunity
27. Vladeck, D.C. (2014) Machines without Principals: Liability Rules and Artificial Intelligence. In: Wash. L. Rev. 89 (2014), p. 117.
28. Wilkinson, C. and Weitkamp, E. (2016) Creative research communication: Theory and practice. Manchester: Manchester University Press.
29. Wilsdon, J. and Willis, R. (2004) See-through science: Why public engagement needs to move upstream, Demos.
30. Winfield, A. F. (2012) Robotics: A very short introduction, Oxford University Press.
31. Winfield, A. F., Blum, C. and Liu, W. (2014) Towards an ethical robot: Internal models, consequences and ethical action selection. In: Mistry, M., Leonardis, A., Witkowski, M. and Melhuish, C., eds. Advances in Autonomous Robotics Systems, LNCS Vol 8717, pp. 85-96, Springer.

32. Winfield, A.F. (2017) Ethical Governance is essential to building Trust in Robotic Systems, Ethics and Information Technology, under review.
33. Worrell, M. (2016) Analysis of Bruntingthorpe crash test data, Impact: The Journal of The Institute of Traffic Accident Investigators, Vol 21, No 1 pp 4-10.