The Ethical Roboticist: a journey from robot ethics to ethical robots

Alan F T Winfield
University of the West of England, Bristol
184 PUBLICATIONS **2,260** CITATIONS

ALAN F. T. WINFIELD

The Ethical Roboticist: a Journey from Robot Ethics to Ethical Robots

Abstract

This aim of this essay is to address the question *what does an ethical roboticist do?* by explaining the author's work in both *robot ethics* and *ethical robots*. Robot ethics is concerned with the *human* problem of the ethical design and application of robots and robotic systems, whereas ethical robots describes the *technical* problem of how to design robots that are capable of choosing actions on the basis of ethical considerations. Taking a somewhat autobiographical approach this essay outlines the author's journey from robot ethics to ethical robots.

1. Introduction

In the last 10 years machine ethics has changed from a niche concern of a small group of academics to a subject of intense societal, political and media interest, with multiple initiatives since 2016, notably from the US White House, EU and UK parliaments, and the Japanese Government. An informal survey reveals that at least 10 sets of ethical principles for robotics and artificial intelligence (AI) have been proposed to date, 7 of which were published in 2017 [1].

What has driven this explosion of interest in machine ethics? The recent high profile successes of Deep Learning have no doubt played a key role, exemplified by the dramatic success of DeepMind's *AlfaGo* AI in defeating the world's best Go player in 2016 (Silver et al, 2016). Advances in AI alongside real-world trials of driverless cars – and a good deal of media hype – are driving significant investment in AI and robotics companies while also raising public and political concerns over the societal and economic implications of a *fourth industrial revolution* (Schwab, 2017).

This essay will be somewhat autobiographical; I will outline robot and AI ethics, and ethical robots from the perspective of someone who has been – and still is – deeply involved in a number of ethics initiatives, including new emerging ethical standards. I will describe my journey into robot ethics while outlining the various initiatives in which I have been (or am) involved. This essay does not therefore provide a comprehensive review of the field, but instead the personal perspective of an ethical roboticist.

2. Robot Ethics

Robot ethics has been defined as *a new field of robotics concerned with both the positive and negative implications of robots to society*, which also aims to inspire *the moral design, development and use of robots* (Tzafestas, 2016). Robot ethics is often described as concern for the ethical, legal and societal impact of robotics; more broadly, we should add economic and environmental impacts to these concerns. A useful way of framing ethical concerns is to think about the unintended harms that might arise from the use of robots. These include physical, psychological or socio/ economic harms and the ethical roboticist is concerned with both articulating and raising awareness of these harms and recommending ways to minimise or mitigate them. Harms can come about at many levels, from the individual, to groups or populations (societal harms) or global (environmental harms).

In order to illustrate robot ethics let us consider a number of ethical problems in robotics. The first and undoubtedly the one that is the greatest concern to most people is the ethical problem of robots that displace jobs. This is of course not a new phenomenon; robots have been replacing humans on car assembly lines since the 1960s and the trend has continued into domains as diverse as warehouses and milking parlours. Near future fears revolve around driverless cars or trucks displacing taxi and truck drivers.

Many people share a second major ethical concern over the development of robot weapons and – more generally – the weaponisation of AI. So-called Lethal Autonomous Weapon Systems (LAWS) are the subject of international efforts to secure a treaty ban [2], while weaponised AI has now been recognised as a threat in the form of both cyber warfare and – in the political sphere – as a worrying means of influencing voter opinion in election campaigns.

I believe there are two very high level ethical concerns that reflect wider societal problems: the first is the significant under-representation of women in robotics and AI – in the UK for instance only 9% of professional engineers are women; and the second is widening wealth inequality. These two issues are of course not intrinsic to robotics and AI, but few would disagree that in order for robotics and AI to benefit all in society, they must be addressed.

2.1 Asimov's laws of robotics

Without doubt the first fully articulated principles of robotics are Isaac Asimov's now famous three laws of robotics, which first appeared in his short story *Runaround* (Asimov, 1942). Asimov's laws were remarkable prescient; implicit in the three laws is an assumption that robots are autonomous, self-aware and (although the term artificial intelligence didn't exist at the time), intelligent. Indeed, Asimov himself came to regard his laws as a basis for governing real robot behaviour, writing in 1981: *I have my answer ready whenever someone asks me if I think that my Three Laws of Robotics will actually be used to govern the behavior of robots, once they become versatile and flexible enough to able to choose among different courses of behavior. My answer is, "Yes, the Three Laws are the only way in which rational human beings can deal with robots – or with anything else"* (Asimov, 1981 [3]).

It is perhaps a testament to the enduring influence of Asimov’s laws of robotics that a number of scholars have felt the need to argue that they are unsuitable as a basis for governing real world robots (Clarke 1993, 1994 [4]; Anderson 2008; Murphy and Woods 2009), although others have argued that – in the absence of any agreed framework of computational ethics – they can at least serve as a starting point for researching ethical robots (Etzioni and Weld 1994 [5]; Vanderelst and Winfield 2017 [6]).

What is however incontrovertible is that Asimov’s laws established the *principle* that robotics (and by extension intelligent systems) should be governed by principles.

2.2 The EPSRC Principles of Robotics

Although Asimov’s laws of robotics are designed to govern the behaviour of robots, not roboticists, Clarke observed that while *many facets of Asimov's fiction are clearly inapplicable ... the substantive content of the laws could be used as a set of guidelines to be applied during the conception, design, development, testing, implementation, use, and maintenance of robotic systems* (Clarke, 1994).

My own journey into robot ethics began with public engagement. Many years of public lectures, panel debates and dialogue events had sensitised me to public fears (as well as fascination) with robot futures – thus, when asked by the UK Engineering and Physical Sciences Research Council (EPSRC) to co-organise a workshop on robot ethics, I jumped at the chance.

First published online in 2011, the EPSRC principles of robotics were not just inspired by Asimov’s laws of robotics, but decidedly revised to refocus those three laws from robots to roboticists. That revision resulted in not three but five general ethical principles for roboticists (Winfield, 2011; Boden et al, 2017 [7]). They are:

1. 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.
2. Humans, not robots, are responsible agents. Robots should be designed and operated as far as is practicable to comply with existing laws and fundamental rights & freedoms, including privacy.
3. Robots are products. They should be designed using processes, which assure their safety and security.
4. Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent.
5. The person with legal responsibility for a robot should be attributed.

Importantly, these principles downplay the specialness of robots, treating them as tools and products to be designed and operated within legal and technical standards. Or, as Bryson writes ... *robots are not responsible parties under the law, and that users should not be deceived about their capacities*. In the same essay Bryson argues that the principles are *de facto policy: the EPSRC principles are of value because* *they represent a policy … their purpose is to provide consumer and citizen confidence in robotics as a trustworthy technology fit to become pervasive in our society* (Bryson, 2017).

2.3 From Principles to Standards

The diagram in Fig. 1 was developed as part of written evidence submitted to the 2016 UK Parliamentary Select Committee on Science and Technology inquiry on Robotics and Artificial Intelligence. The diagram attempts to answer the question *how do we trust our technology?* by showing that ethical principles lead to standards, which in turn lead to regulation. Of course not all ethical principles find expression as standards, but one which notably does is the principle that systems must be *safe*, and most standards are voluntary – only a subset of standards, often because they relate to safety-critical systems, are mandated by regulation.

When regulation is transparent – through effective public engagement – and regulatory bodies are seen to be effective, then trust is engendered. A good example is passenger airliners. We trust them because they have an outstanding safety record and – when rare accidents do happen – processes of accident investigation are swift, thorough and transparent. As Fig. 1 also tries to show the key elements of ethics, standards and regulation are underpinned by the human processes of responsible research & innovation (RRI) and ethical governance, and the technical processes of benchmarking, verification & validation (Winfield 2016) [8].

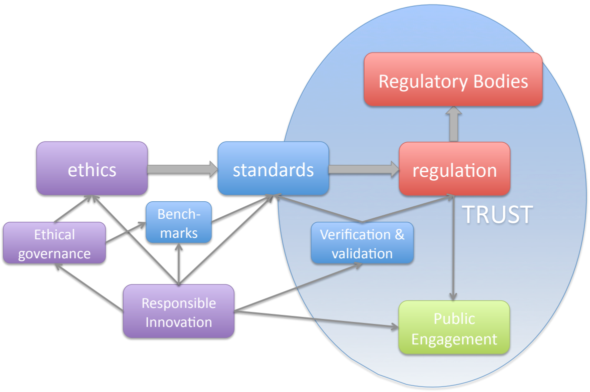

Fig. 1. A Roadmap: from ethical principles to trust (from Winfield, 2016)

Arguably the world's first published ethical standard in robotics is *BS 8611 Guide to the ethical design and application of robots and robotic systems* (BSI, 2016). BS 8611 incorporates the EPSRC principles of robotics; it is not a code of practice, but instead a toolkit for designers to be able to undertake an *ethical risk assessment* of their robot or system, and mitigate any ethical risks so identified. In the working group that drafted BS 8611 we identified 20 distinct ethical hazards and drafted advice on their mitigation. The societal hazards include, for example, loss of trust, deception, privacy and confidentiality, addiction and employment.

New, so-called 'human', standards are in development within the IEEE Standards Association global ethics initiative [9]. Presently, 11 standards working groups are drafting candidate standards formalising one or more ethical principles. To give just one example, in IEEE P7001 *Transparency in Autonomous Systems* [10] we are defining a set of measurable, testable levels of transparency for each of several stakeholder groups including users, certification agencies and accident investigators (Bryson and Winfield, 2017).

3. Ethical Robots

Is it possible to build an ethical robot: a robot capable of choosing or moderating its actions on the basis of ethical rules? Five years ago I thought the idea impossible and said as much in my *Introduction to Robotics* (Winfield, 2012). But since then I've changed my mind. In fact not only changed my mind but also developed an architecture and, with colleagues, implemented and experimentally tested an ethical robot. So, what brought about this u-turn? It was not an epiphany - more a case of several ideas slowly coming together.

First was the realisation that robots don't need to be sentient in order to act in a way we would judge to be ethical. In other words we don't need a major breakthrough in artificial intelligence to build an ethical robot. A relatively simple robot could behave ethically not because it chooses to, but because we have programmed it so. It would be an *ethical zombie* – capable of ethical actions, without understanding what it is doing, or why.

Second was thinking about very simple ethical behaviours. Consider a thought experiment: imagine you are out walking and you notice someone who is not looking where they are going. She's heading straight for a hole in the pavement – wearing headphones and peering at her smart-phone perhaps. You will probably try and intervene. How it that? Well it's not *just* because you're a good person – it's because you also have the cognitive machinery to *predict* not only the consequences of her actions, but which action you should take to try and avert a calamity.

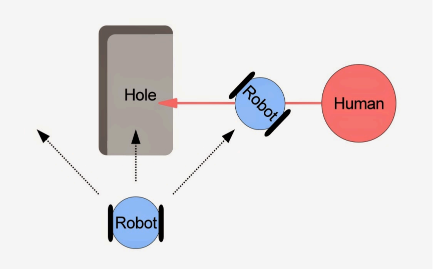

Fig 2. Ethical robot thought experiment

Now imagine it's not you, but a robot that has four possible next actions: turn toward the left, stand still, continue straight ahead, or turn toward the right. Fig, 2 shows this simple thought experiment. From the robot's perspective, it has two safe options: stand still, or turn to its left. Go straight ahead and it will fall into the hole. Turn right and it is likely to collide with the human.

But if the robot can *model* the consequences of both its own actions *and* the human's - another possibility opens up: the robot could *choose* to collide with the human to prevent her from falling into the hole. Here is a simple rule for this behaviour:

```
IF for all robot actions, the human is equally safe
THEN (* default safe actions *)
    output safe robot actions
ELSE (* ethical action *)
    output action(s) for least unsafe human outcome(s)
```

This rule appears to match remarkably well with Asimov's first law of robotics: *A robot may not injure a human being or, through inaction, allow a human being to come to harm*. The robot will avoid injuring (i.e. colliding with) a human (may not injure a human), but may also sometimes compromise that rule in order to prevent a human from coming to harm (...or, through inaction, allow a human to come to harm).

So emerged the idea is that we might be able to build a robot with *Asimovian* ethics. We need to equip the robot with the ability to predict the consequences of both its own, and other(s) actions, plus the hard wired ethical logic – the IF THEN ELSE code above.

Third came the realisation that the technology we need not only exists but is mature and commonplace in robotics research – it is the robot simulator. Robot simulators provide developers with a virtual environment for prototyping robot code before then running that code on the real robot. Now the idea of embedding a simulation of a robot inside that robot is not new – but it is technically challenging and only a few researchers have pulled it off. The robot would need to have, inside itself, a simulation of itself and its immediate surroundings, including other dynamic actors like humans or other robots in its vicinity.

I was fortunate to be able to present and debate these ideas with fellow researchers at several meetings in 2013, but they remained just ideas on paper. Then a stroke of luck: in 2014 Christian Blum, a brilliant PhD student from the cognitive robotics research group of the Humboldt University of Berlin joined my research group for six months. I suggested Christian implement these ideas on our e-puck robots and happily he was up for the challenge. Thus Christian, supported by my post-doc Research Fellow Dr Wenguo Liu, implemented what we call a *Consequence Engine*, running in real-time, on the e-puck robot.

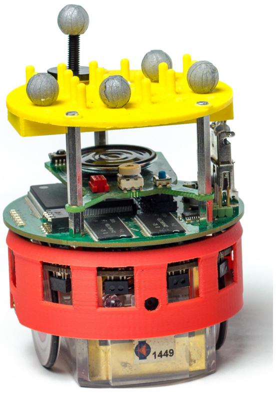

Fig 3. The e-puck robot

The e-puck robot, shown in Fig. 3, is a Swiss mobile robot designed primarily for education and research. We've been using these robots for several years for swarm robotics research, and extended our e-pucks by designing a Linux card – the circular green board shown here just above the red skirt (Liu and Winfield, 2011).

Running the open source 2D robot simulator *Stage* as its internal simulator our consequence engine runs at 2Hz, so every half a second it is able to simulate about 30 next possible actions and their consequences. The simulation budget allows us to simulate ahead around 70cm of e-puck motion for each of those next possible actions. In fact Stage is actually running on a laptop linked to the robot over our fast WiFi LAN. But logically it is inside the robot. What's important here is the proof of principle.

We tested the human-heading-for-a-hole scenario with two e-puck robots: one robot with consequence engine plus ethical rule (the *A-robot* – after Asimov), and another robot acting as a proxy human (the *H-robot*).

Fig. 4 shows the experimental arena. We don't have a real hole, but a virtual hole – the yellow shaded square on the right. We just *tell* the *A-robot* where the hole is. We also give the *A-robot* a goal position – at the top right – chosen so that the robot must actively avoid the hole. The *H-robot* on the left, acting as a proxy human, doesn't *see* the hole and just heads straight for it. (Ignore the football pitch markings – we're re-using this handy robo-soccer pitch.)

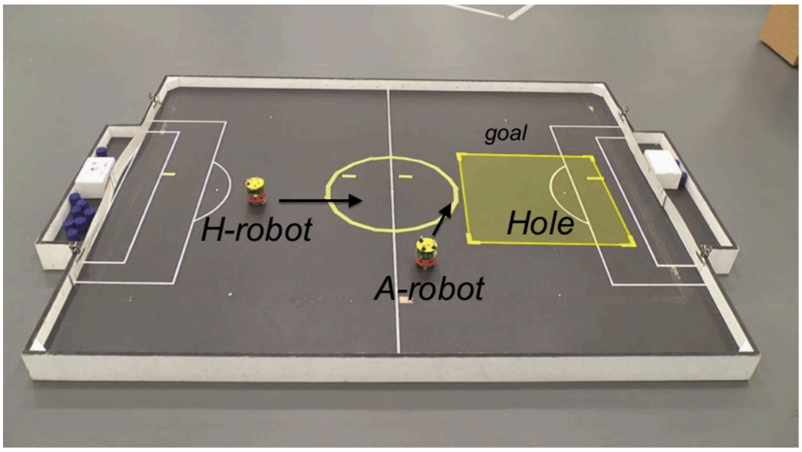

Fig. 4 Experimental setup for ethical robot trials

So, what happens? For comparison we ran two trials, with multiple runs in each trial. In the first trial is just the *A-robot*, moving toward its goal while avoiding falling into the hole. In the second trial we introduce the *H-robot*. The graphs in Fig. 5 show the robot trajectories, captured by our robot tracking system, for each run in each of these two trials.

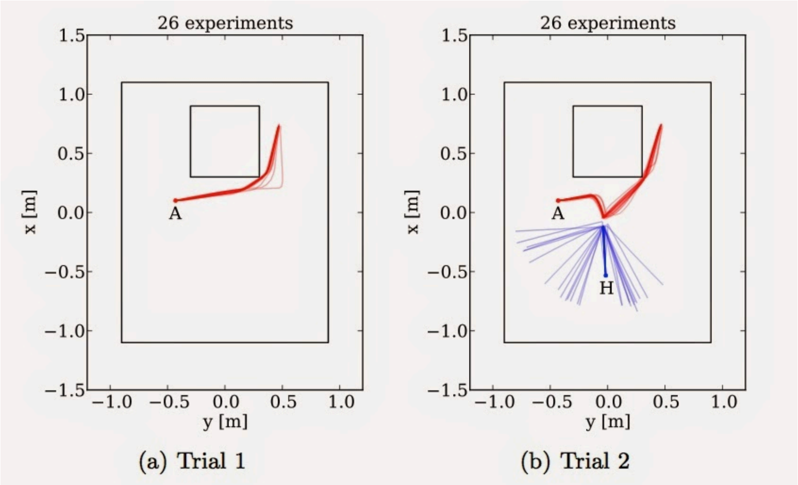

Fig. 5 Experimental results for trials 1 and 2

In trial 1, notice how the *A-robot* neatly clips the corner of the hole to reach its goal position. Then in trial 2, see how the *A-robot* initially moves toward its goal then, when it notices that the *H-robot* is in danger of falling into the hole, it diverts from its trajectory in order to head-off H. By provoking H's collision avoidance behaviour, A sends it off safely away from the hole, before then resuming its own progress toward its goal position. The *A-robot* is 100% successful in preventing H from falling into the hole.

At this point we started to write up our results, but we needed something more than 'we built it and it works just fine'. So we introduced a third robot – acting as a second proxy human. So now our ethical robot would face a dilemma – which *H-robot* should it rescue? Actually we thought hard about this question and decided not to program a rule, or heuristic, partly because ethicists, not engineers, should decide such a rule and partly because we wanted to test our ethical robot with a balanced ethical dilemma.

We set the experiment up carefully so that the *A-robot* would notice both *H-robots* at about the same time – noting that because these are real physical robots no two experimental runs will be exactly identical. The results were very interesting. Out of 33 runs, 16 times the *A-robot* managed to rescue one of the *H-robots*, but not the other, and amazingly, 3 times the *A-robot* rescued both. In those 3 cases, by chance the *A-robot* rescued the first *H-robot* very quickly and there was just enough time to get to the second before it reached the hole. Small differences in the trajectories of H and H2 helped here. But most interesting were the 14 times when the *A-robot* failed to rescue either. Why is this, when there is clearly time to always rescue one? When we studied the videos, we see the answer. The problem is that the *A-robot* sometimes dithers. It notices one *H-robot*, starts toward it but then almost immediately notices the other. It appears to *change its mind*. And the time lost dithering means the *A-robot* cannot prevent either robot from falling into the hole. We believe this is the first experimental test of a real robot facing an ethical dilemma.

In three years I went from sceptic to believer, and ended up building a minimally ethical robot. But, as we say in our paper (Winfield et al. 2014), we are not claiming that a robot which apparently implements part of Asimov's famous laws is ethical in any formal sense, i.e. that an ethicist might accept. But even minimally ethical robots could be useful. Our approach is a step in this direction.

4. Conclusions

This essay has I hope answered the question *what does an ethical roboticist do?* (or, more accurately *this* ethical roboticist), by outlining the development of ethical principles which, in turn, are leading to new ethical standards in robotics and AI. Parallel efforts to advise policymakers will, it is hoped, lead to regulation and hence public trust – and trust is vital if these new disruptive technologies are to succeed and bring the societal benefits we believe they should.

Let me finish with a question about the ethics of ethical robots: *if we could build an ethical robot are we ethically compelled to do so?* Some argue that we have an ethical duty to try and build moral machines. But the counter argument is: are there ethical hazards? Might, for instance, an ethical robot be vulnerable to accidental or malicious hacking – transforming the robot from ethical to unethical, as we argue in Vanderelst and Winfield (2018)? While there is no question that we need ethical roboticists there are considerable doubts over ethical robots.

Acknowledgements

The work of this essay has been supported by EPSRC grants EP/G063052/1 and EP/L024861/1. I am very grateful to Dr Christian Blum (supported by Dr Wenguo Liu) who programmed the robots, set up the experiment and obtained the results outlined in section 3. I owe thanks also to collaborator Professor Michael Fisher for encouraging me to think deeply on the question of how to build ethical robots.

Bibliography

Anderson, S. (2008), Asimov's "three laws of robotics" and machine metaethics, *AI and Society*, Vol 22 No 4, pp 477–493.

Asimov, I (1950), Runaround, *Astounding Science Fiction*, 1942, Reprinted in *I, Robot*, Gnome Press, 1950.

Asimov, I. (1981), Guest commentary: The Three Laws, *Compute! Magazine*, Vol 18 No 3, pp 18, November 1981. [3]

Boden, M., et al (2017), Principles of robotics: regulating robots in the real world, *Connection Science*, Vol 29 No 2, pp 124-129. [7]

BSI (2016), *BS8611:2016 Robots and robotic devices: guide to the ethical design and application of robots and robotic systems*, British Standards Institute, London.

Bryson, J., (2017) The Meaning of the EPSRC Principles of Robotics, *Connection Science*, Vol 29 No 2, pp 130-136.

Bryson, J. and Winfield, A. (2017), Standardizing Ethical Design for Artificial Intelligence and Autonomous Systems, *IEEE Computer* Vol 50 No 5, pp 116-119.

Clarke, R. (1993), Asimov's Laws of Robotics: Implications for Information Technology, *IEEE Computer,* Vol 26 No 12 (Dec 1993) pp 53-61 and Vol 27 No 1 (Jan 1994), pp 57-66. [4]

Etzioni, O., and Weld, D. (1994), The First Law of Robotics (a call to arms), *AAAI Tech. Rep. SS-94-03*, pp 17-23. [5]

Liu W and Winfield AF (2011), Open-hardware e-puck Linux extension board for experimental swarm robotics research, *Microprocessors and Microsystems*, Vol 35 No 1.

Murphy R, and Woods D. (2009), Beyond Asimov: The Three Laws of Responsible Robotics. *IEEE Intelligent Systems*, Vol 24 No 4.

Silver, D. et al (2016), Mastering the game of Go with deep neural networks and tree search, Nature, Vol 529, pp 484-489.

Schwab, K. (2017), *The Fourth Industrial Revolution*, Portfolio Penguin.

Tzafestas, S (2016), *Roboethics: a navigating overview*, Springer.

Vanderelst, D. and Winfield, A. (2017), An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research*. Vol 48, pp 56-66. [6]

Vanderelst, D. and Winfield, A. (2018), The Dark Side of Ethical Robots, in Proc. AAAI Workshop on AI, Ethics and Society (AIES 2018), New Orleans.

Winfield, AF (2011) Roboethics for Humans, *New Scientist*, pp 32-33, May 2011.

Winfield, AF (2012) *Robotics: A Very Short Introduction*, Oxford University Press.

Winfield, AF (2016) Written evidence submitted to the UK Parliamentary Select Committee on Science and Technology Inquiry on Robotics and Artificial Intelligence, Discussion Paper, *Science and Technology Committee (Commons)*, Website, 2016. [8]

Winfield AF, Blum C and Liu W (2014) Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection, pp 85-96 in *Advances in Autonomous Robotics Systems*, Lecture Notes in Computer Science Vol. 8717, Eds. Mistry M et al, Springer, 2014.

Web sites:

[1] http://alanwinfield.blogspot.co.uk/2017/12/a-round-up-of-robotics-and-ai-ethics.html
[2] https://www.icrac.net/
[3] https://archive.org/stream/1981-11-compute-magazine/Compute_Issue_018_1981_Nov
[4] http://www.rogerclarke.com/SOS/Asimov.html
[5] http://www.aaai.org/Papers/Symposia/Spring/1994/SS-94-03/SS94-03-003.pdf
[6] http://doi.org/10.1016/j.cogsys.2017.04.002
[7] http://dx.doi.org/10.1080/09540091.2016.1271400
[8] http://eprints.uwe.ac.uk/29428/
[9] http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html
[10] http://sites.ieee.org/sagroups-7001/