UK-RAS White papers

ISSN 2516-5011

UK-RAS NETWORK
ROBOTICS & AUTONOMOUS SYSTEMS

Ethical Issues for Robotics and Autonomous Systems

www.ukras.org

UK-RAS
NETWORK
ROBOTICS & AUTONOMOUS SYSTEMS

UKRAS.ORG

FOREWORD

Welcome to the UK-RAS White Paper Series on Robotics and Autonomous Systems (RAS). This is one of the core activities of UK-RAS Network, funded by the Engineering and Physical Sciences Research Council (EPSRC). By bringing together academic centres of excellence, industry, government, funding bodies and charities, the Network provides academic leadership, expands collaboration with industry while integrating and coordinating activities at EPSRC funded RAS capital facilities, Centres for Doctoral Training and partner universities.

With rapid technological advances of robotics and AI, it is timely to address the associated ethical issues. Many reports predict a huge increase in the number of robots in the future, with many of these being service robots. The technological transition from industrial robots to service robots represents an evolution into more personalized systems with an increasing degree of autonomy, however, robots and autonomous systems are gradually expected to have widespread exploitation in society.

While the impact of industrial robots has been present for a number of years, the impact of service robots in workplaces and at home is still to be seen and assessed. Progress in artificial intelligence research will have a major impact on how quickly we see intelligent and autonomous service robots. This paper reviews work considering both the regulation of future potential of robotics and AI systems, and the ethical considerations that need to be taken. References to recent initiatives to outline ethical guidelines for both the design of systems and how they should operate are also included.

The UK-RAS white papers are intended to serve as a basis for discussing the future technological roadmaps, engaging the wider community and stakeholders, as well as policy makers, in assessing the potential social, economic and ethical/legal impact of RAS. It is our plan to provide annual updates for these white papers, so your feedback is essential - whether it is to point out inadvertent omissions of specific areas of development that need to be covered or to suggest major future trends that deserve further debate and in-depth analysis.

Please direct all your feedback to info@ukras.org.

We look forward to hearing from you!

Prof Guang-Zhong Yang, CBE, FREng
Chairman, UK-RAS Network

AUTHORS





Prof. Alan Winfield
University of the West of England
alan.winfield@uwe.ac.uk

Prof. John McDermid OBE FREng
The University of York
john.mcdermid@york.ac.uk

Dr Vincent C Müller
Professor, Ethics of Technology
TU Eindhoven
University Academic Fellow
University of Leeds
Turing Fellow
Alan Turing Institute
v.c.muller@leeds.ac.uk

Ms. Zoë Porter
The University of York
zoe.porter@york.ac.uk

Prof. Tony Pipe
University of the West of England
tony.pipe@brl.ac.uk

Reviewer feedback acknowledgements:
Ana MacIntosh, Carsten Maple, Anna Angus-Smyth, Mark Gaskarth, Guang-Zhong Yang

CONTENTS

> The aim is to be pragmatic, rather than philosophical, offering designers and operators of RAS some perspectives that may help in reflecting on the design or operation of RAS.

1. INTRODUCTION

Ethics are moral principles that govern a person's behaviour or the conduct of an activity. As a practical example, a principle might be to treat everyone with respect. Philosophers have considered ethics over many centuries, and there are various well-known principles, perhaps one of the most famous being Kant's categorical imperative "act as you would want all other people to act towards all other people"[1]. Our concern in this paper is on how ethics and ethical principles should apply in the context of robotics and autonomous systems (RAS).

RAS may operate autonomously, i.e. independent of human control, but they are designed by humans, so there are several different ethical perspectives to consider:

- For designers and developers of RAS;
- For operators of RAS;
- For the RAS, where we consider RAS as "moral machines" in themselves.

Some robots are quite simple with well-defined safety mechanisms and any ethical issues are probably adequately covered by normal engineering ethics, e.g. the joint Statement of Ethical Principles[2] produced by the Engineering Council and Royal Academy of Engineering. However, the ethical issues become more complex when decisions that are normally undertaken by humans, e.g. for driving a car or piloting an aircraft, are transferred to the RAS. In this case, the ethical concerns that might attach to a human-made decision can be seen as relevant to the RAS – with ethical responsibility transferred to the developers, operators, or perhaps to the RAS itself.

Many RAS use Artificial Intelligence (AI), which we interpret here as any kind of computational system that shows "intelligent" behaviour, i.e. complex problem-solving capabilities. In general RAS, where they use AI, are solving specific tasks, e.g., route planning, and do not involve artificial general intelligence (AGI). In the rest of this paper we shall assume that any use of AI is in for a specific or "narrow" purpose and does not involve AGI. Whilst the use of AI, and more particularly Machine Leaning (ML), does not, of itself, create ethical issues, it can introduce ethical problems in that, for example, the learning might introduce biases, which would be considered unethical (or even illegal if they are seen as discriminatory). Whilst ethics of RAS and AI are not synonymous, we treat them as very strongly related as many of the challenges for RAS arise from the use of AI and ML in their development.

Ethics of AI and RAS is a very active area, with many substantial initiatives being undertaken, e.g. by the Institute of Electrical and Electronic Engineers (IEEE)[3] and the European Commission[4]. There is also extensive domain-specific activity, for example related to healthcare[5] and to autonomous vehicles (AVs). Further, the rate of production of new material on ethics of RAS and AI makes it very hard to provide a treatment of the issues which is both up-to-date and which adds value to what has already been published.

Thus this paper has a focused ambition: to set out clearly some ethical concerns and principles that are relevant in the context of developing and operating RAS; to identify some of the questions to be answered if we are to consider RAS as "ethically-aligned machines", and to provide an annotated bibliography of some of the more relevant developments in the domain. The aim is to be pragmatic, rather than philosophical in nature, offering designers and operators of RAS some perspectives that may help in reflecting on the design or operation of RAS. The paper also considers some open issues that do not seem to have received sufficient attention, in the literature, at this time. It is intended that this paper will be supplemented by a companion document that looks into some of these issues more deeply, with the aim of providing more specific guidance on ethics for RAS.

[1] This drawn from his 1785 book "Groundwork of the Metaphysics of Morals", and has a variety of translations from the original German.
[2] https://www.raeng.org.uk/publications/reports/statement-of-ethical-principles (accessed April 2019)
[3] https://ethicsinaction.ieee.org (accessed April 2019)
[4] https://ec.europa.eu/digital-single-market/en/news/draft-ethics-guidelines-trustworthy-ai (accessed April 2019)
[5] https://topol.hee.nhs.uk/ (accessed May 2019)

ETHICAL CONCERNS

The ethical concerns raised by RAS depend on their capabilities and domain of usage. The following outlines some ethical concerns that might arise in a range of domains; there is no implication that these concerns are complete or exhaustive; the order is alphabetical.

- Bias – RAS might have bias in their decision-making (based on their learning), e.g., if AVs have been trained on an ethnically biased set of images they may be more likely to fail to recognise certain ethnic groups as human (due to skin colour or clothing, for example), and make decisions that place such groups at greater risk;
- Deception – humanoid or zoomorphic robots present the risk, especially to naïve or vulnerable users, of emotional attachment or dependency (given that it is relatively easy to design a robot to behave as if it has feelings). The 4th EPSRC Principle of Robotics states "Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent"[6];
- Employment – introduction of RAS might displace certain classes of workers, e.g., taxi drivers and operators of quarrying machines; this might also involve bias[7];
- Opacity – where decisions are not transparent, i.e., open to scrutiny, there is a possibility that they are both unfair (unjust) and not open to correction; the introduction of the General Data Protection Regulations (GDPR)[8] brings with it a "right to explanation", motivated by the problem of opacity;
- Safety – RAS can both positively and negatively impact safety; the original motivation for research on AVs was to improve road safety, by reducing or removing human errors as an accident cause; however, as recent accidents with AVs in the US have shown such technology can also cause fatalities; ethical issues here include the safety (and fail-safety) of RAS, per se, and any redistribution of risk that might arise from introducing RAS;
- Oversight – the ability to oversee, or govern[9], RAS is an ethical issue as operators should be able to understand and manage the behaviour of systems for which they are responsible; this is linked to opacity, but also comes from RAS operating in open environments where it is difficult to monitor and assess their behaviour;
- Privacy – RAS may contain, and be able to provide to third parties, data which could violate an individual's right to privacy; for example, an AV is likely to know where the owner or occupant travelled, and this might, for example, allow a stalker to track them, or to show they were involved in criminal activity, or not sick at home as they claimed to be.

None of these concerns are "black and white". For example, autonomous agricultural machinery, e.g., combine harvesters, might affect (reduce) the employment of agricultural workers, whilst also dramatically contributing to their safety – and agriculture is one of the most dangerous occupations in the UK[10].

[6] Boden, M. et al. (2017) Principles of robotics: Regulating robots in the real world. Connection Science, 29 (2). pp. 124-129. ISSN 0954-0091

[7] We are aware of cases in Australia where operators of quarrying equipment were predominantly Aboriginal females, thus the introduction of autonomous equipment disproportionately affected one of the most disadvantaged groups in Australian society.

[8] https://www.gov.uk/government/publications/guide-to-the-general-data-protection-regulation (accessed April 2019)

[9] Winfield, A. F. and Jirotka, M. (2018) Ethical governance is essential to building trust in robotics and AI systems. Philosophical Transactions A: Mathematical, Physical and Engineering Sciences, 376 (2133). ISSN 1364-503X

[10] http://www.hse.gov.uk/statistics/pdf/fatalinjuries.pdf (accessed April 2019)

ETHICAL PRINCIPLES

The Engineering Council and Royal Academy of Engineering ethical principles contain some basic guidelines, e.g. that engineers should:

- Hold paramount the health and safety of others and draw attention to hazards;
- Ensure their work is lawful and justified.

As indicated above, these may be sufficient for simple robots, and they are always relevant (although they don't cover all of the concerns above). In the rest of this paper they are taken as setting a baseline for any work on RAS, and our focus will be on the specific issues raised by RAS, including the use of AI and ML. This section identifies some ethical principles that might be of use in dealing with the design and operational perspectives noted above.

One of the most general principles that can underpin ethical design and operation of RAS is that of "distributive justice" – that goods are distributed in a way that is rational and defensible[11]. However, it can be extended to other concerns such as risk and viewed as a way of considering some of the concerns above.

For example, introducing RAS might disproportionately affect the employment of those of lower educational attainment, and who are otherwise disadvantaged. Considering "distributive justice" this might suggest that such systems shouldn't be deployed. However, a more subtle application of the principle is appropriate. For example, it might be that through retraining displaced workers have an equally good chance of employment, and the opportunity for better paid, more rewarding, and less dangerous jobs[12].

Considering privacy, there will be a balance between individual rights and those of society as a whole. Here the rational and defensible position might be to allow law-enforcement authorities to inspect data about an individual's use of an AV, when there is reason to believe that they have committed an offence. Indeed, this is little different to current rules regarding access to data not associated with RAS.

Another principle that might apply to the design and deployment of RAS is "reflective equilibrium"[13]. Reflective equilibrium is the end-point of a deliberative process in which some beliefs, thoughts and judgements about a particular topic are systematically revised in order to achieve coherence among them. The concept gained currency from the work of Rawls, and can be used in applied ethics as a method of justification: testing principles and theories against judgements about particular cases, but also testing judgements about particular cases against principles and theories, until equilibrium is achieved.

This might apply, for example, when considering safety in a design setting. The benefits – risk reduction – from autonomy due to reduction in, or elimination of, human error should be balanced against the risk attendant in the technology (including ML, and the limitations of training data, if appropriate). Thus, one principle ("overcome human error") and another principle ("acknowledge the limitations of the system") might be tested against intuitions and judgements about what would be acceptable in particular cases where this trade off would be at play. In principle, one would expect that design reviews, including ethical risk assessment[14], would offer the opportunity for such reflection between parties who have different knowledge and skills to bring to bear on the problem although project and other pressures might make this difficult[15].

In practice, it might be that reflective equilibrium is more practical in dealing with accidents or incidents. Here there is an immediate problem at hand, and there will be a complex set of factors in play, including achieving or restoring public confidence.

[11] Rawls, J. A Theory of Justice, Harvard University Press, 1971

[12] It is understood that, in the Australian quarrying case mentioned above, exactly this sort of retraining scheme was provided for affected workers.

[13] https://plato.stanford.edu/entries/reflective-equilibrium/ (accessed April 2019)

[14] BS8611:2016 Guide to the Ethical Design of Robots and Robotics Systems.

[15] Experience of at least one of the authors would suggest that design reviews rarely afford the space and time for such reflections. See also https://ieeexplore.ieee.org/document/8466102 (accessed May 2019)

A further principle, attributed to Kant[16], is "ought implies can", such that an agent is only obliged to perform an action that it is possible for him or her to perform. This has a bearing on many RAS that are not fully autonomous, but where there is a form of human-system collaboration, whether it is working collaboratively (so-called cobots) or handover, e.g., with AVs.

This principle can be seen to apply to design, for example is it reasonable to expect drivers (operators) of AVs to take back control after a period of autonomous driving? If so, how long is needed to regain situational awareness? Although this is an ethical issue it can be "tested" to an extent through simulation and experiment – for example as was done with Volvo in determining whether or not "safety drivers" could and would take over responsibility for emergency braking[17].

In the case of organisations operating RAS, the principle could also be applicable. For example, in introducing maritime autonomy operators might move ship's captains from a role of controlling vessels to monitoring them from a remote operating centre (ROC). To be economically viable, it is likely that the captains will have to monitor (oversee) multiple vessels simultaneously (if it is one-for-one then the cost of the automation and establishing the ROC is likely to outweigh and economic gains from removing other staff from the vessels). The operations should be designed so that the captains can oversee and manage the safety of all the vessels they are responsible for remotely – with changes in design, e.g., levels of automation, made if the "ought implies can" principle would be violated, due to inability to maintain situational awareness, etc.

Another principle is "participatory design"[18]. This might be viewed as a general design heuristic, not an ethical principle, but it is an important way of addressing ethical concerns. If genuine end-users are involved through participatory design, then there is the opportunity to reduce bias, to understand the impact on employment, the impact on privacy and safety, and perhaps the practicality of oversight, although this is less likely to be helpful in considering opacity. Such participatory design is an important part of Responsible Innovation[19].

The above discussion of ethical principles is not intended as a "recipe" for ethical design or operation, but to illustrate how ethical principles might affect design and operation, and how they might be used as guidelines to address the concerns illustrated above.

[16] Critique of Pure Reason, trans. Norman Kemp Smith (London, 1933), p. 473.

[17] Automation Expectation Mismatch: Incorrect Prediction Despite Eyes on Threat and Hands on Wheel, Trent W. Victor, Emma Tivesten, Pär Gustavsson, Joel Johansson, Fredrik Sangberg, and Mikael Ljung Aust, Human Factors, Vol. 60, No. 8, December 2018, pp. 1095–1116

[18] C DiSalvo, I Nourbakhsh, D Holstius, A Akin, and M Louw (2008). The Neighborhood Networks project: a case study of critical engagement and creative expression through participatory design. In Proc. 11th Anniversary Conf. on Participatory Design 2008 (PDC '08). Indiana University, Indianapolis, IN, USA, 41-50.

[19] https://epsrc.ukri.org/research/framework/area/ (accessed May 2019)

LITERATURE ON AI AND RAS ETHICS

There is a very extensive literature on AI and RAS ethics, and a large number of initiatives developing guidelines and standards. The subject is multi-disciplinary, at minimum including computer scientists, experts in robotics, autonomy and AI, lawyers and philosophers. Hence, a complete literature survey would be very extensive, and likely out of date before the paper was published. Instead, we include as an Annex a selective bibliography, which in some cases provides a brief commentary, and which covers:

- Ethical Principles
- Substantive robotics and AI ethics initiatives
- Robotics and AI ethical standards and regulation
- National and International (government) Strategies
- Major Reports and recommendations
- Selected academic works

Much of the material is general, i.e., would apply to a large range of RAS application domains, but some is sector specific, e.g., pertinent to healthcare.

OPEN ISSUES

Many of the concerns raised above are "open" in the sense that there is currently no definitive solution, and it is to be expected that there will be continued work in these areas. However, there are four open concerns that seem to be less well-studied or controversial and require special attention, which we briefly discuss here.

First, as mentioned in the introduction, there is a question of whether or not RAS should be viewed as "moral machines"[20], in the sense that responsibility can be delegated to them for some ethically salient action. One of the most commonly used examples to discuss ethical dilemmas that explicitly ethical machines might face is the so-called "trolley problem" where an AV has to decide between courses of action which would lead to different numbers of fatalities. There is an extensive literature on this issue, and many variants, or refinements, of the problem. MIT has developed an on-line resource[21] where they elicit preferences about who should be "saved" if a fatal accident is inevitable. They conclude that not only should AVs be moral machines, but that they should alter their ethical stance to match the preferences in different parts of the globe[22].

However, it is not obvious whether or not it is appropriate to treat AVs or other RAS as "moral agents". Further, this is quite a complex question, as there are different levels of agency – up to full "human-like" responsibility for decision-making[23]. At minimum, we believe that one should ask whether or not it is appropriate for a RAS to be viewed as a "moral agent" if this absolves the developers or operators of the RAS from moral (and perhaps legal) responsibility for the system. There are arguments on both sides, which is why we view this as an open question[24].

Second, there is an issue of how we assure and regulate RAS. In their review of RAS[25] the Lloyd's Register Foundation identified "white spaces", or gaps, in assurance and regulation meaning aspects of RAS behaviour for which there were not adequate assessment methods. The Assuring Autonomy International Programme[26] has been set up in response to this review. It is addressing these gaps, identifying Critical Barriers to Assurance and Regulation (CBARs); these are issues, that if they cannot be resolved, might lead to unsafe systems being deployed (if the regulatory regime is permissive) or safe systems not being deployed, thus losing potential benefit (if the regulatory regime is restrictive).

At their core, the CBARs are technical, and link to some of the ethical concerns identified above. For example, the verification CBAR is concerned with verifying ML e.g., deep neural networks (DNNs). As DNNs do not make visible their learnt behaviour in a way which is human understandable, this is a source of opacity. A further CBAR is concerned with handover – the ability of a human to take control from a RAS, is the RAS is no longer capable of dealing with the situation. This CBAR is related to oversight, although oversight is rather broader in scope. There is an important ethical perspective here – is it acceptable (for a regulator) to allow deployment of a system where we know that there are not adequate assurance methods? How much is the answer to this question influenced by the benefit that might accrue from use of the system, e.g., allowing the elderly and infirm to continue living independently by use of social care robots?

Third, at a different level, many of the documents relating to ethics of AI and RAS are either posed in general terms or refer to ethical responsibility of individuals (for example, this is the focus of the Engineering Council and Royal Academy of Engineering joint statement). RAS are designed and operated by organisations, not individuals. Thus, we see the notion of corporate/organisational ethics as an open issue. In particular we believe that it is important to understand what a framework for ethical governance for organisations developing or operating RAS would look like and how it might be implemented.

Fourth, related to corporate/organisational ethics is the issue of "who decides" about the design and/or operation of a RAS. Ethically, the "obvious" answer is the person or people who might benefit from the RAS, and those who might be put at (greater) risk. Where these groups are the same it is easier to see how to manage the balance between benefits and risks. In many cases, however, there is not such a simple alignment of benefits and risks and/or such a large population is affected by a decision, e.g. to approve

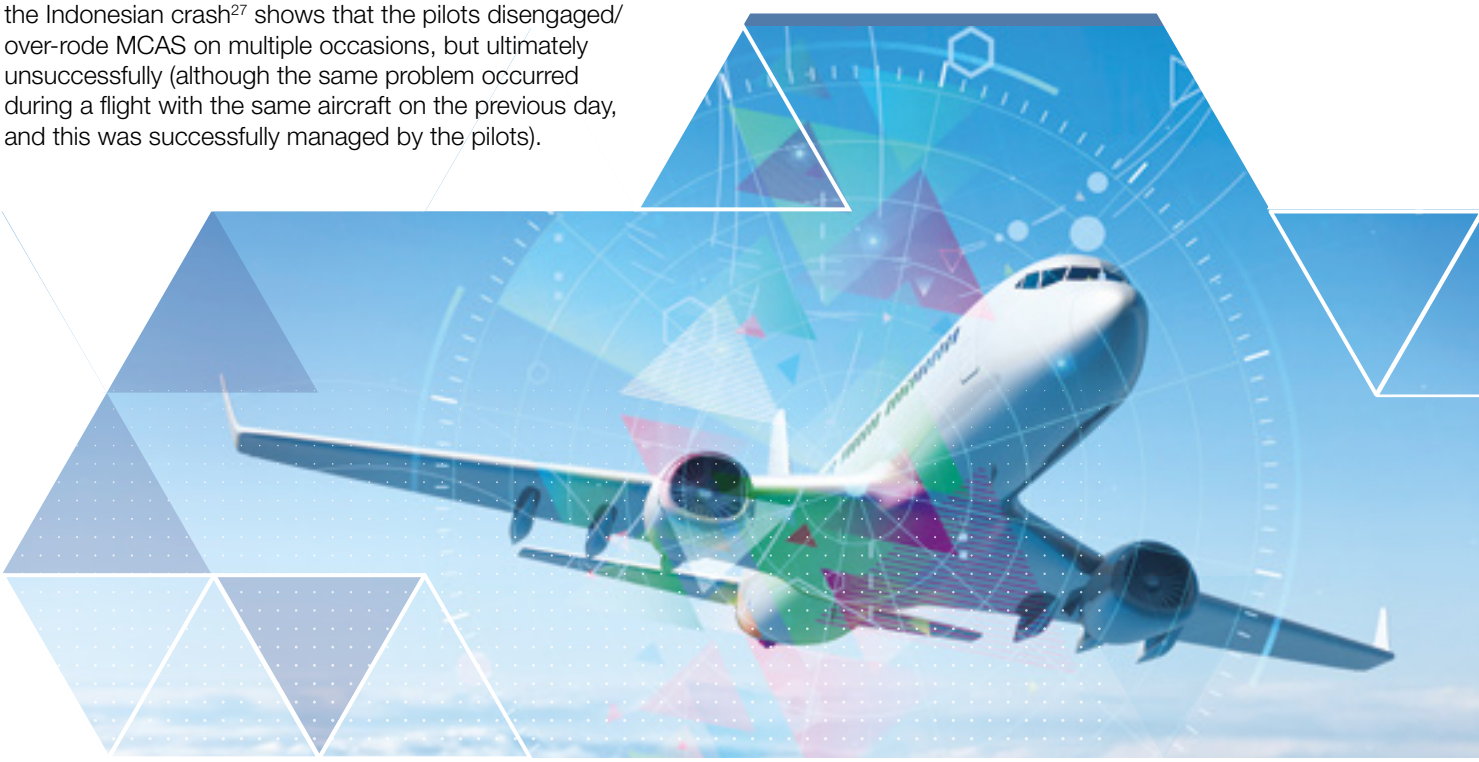
[20] https://ieeexplore.ieee.org/xpl/tocresult.jsp?isnumber=8662725&punumber=5 (accessed May 2019)
[21] http://moralmachine.mit.edu (accessed April 2019)
[22] https://www.media.mit.edu/publications/the-moral-machine-experiment/ (accessed April 2019)
[23] Winfield, AF, Michael, K, Pitt J. and Evers, V. (2019) Machine ethics: The design and governance of ethical AI and autonomous systems. Proceedings of the IEEE, 107 (3). pp. 509-517. ISSN 0018-9219
[24] https://www.lrfoundation.org.uk/en/publications/foresight-review-of-robotics-and-autonomous-systems-ras/ (accessed April 2019)
[25] https://www.york.ac.uk/assuring-autonomy/ (accessed April 2019)
[26] https://reports.aviation-safety.net/2018/20181029-0_B38M_PK-LQP_PRELIMINARY.pdf (accessed April 2019)

operation of a system such as the Docklands Light Railway (DLR), that it makes sense for the State to regulate on behalf of those affected – the public. In many cases, the state can and should carry on in this role. However, what happens where the RAS learns? For example, if an AV learns from other AVs – should the owner of a particular AV be able to "reject" this type of learning, or set a "risk appetite" as is done with financial products? What happens when AVs are provided as a service, rather than being individually owned?

The recent accidents involving the Boeing 737 Max 8 aircraft illustrate the last three points. Whilst the aircraft's Manoeuvring Characteristics Augmentation System (MCAS) was unlikely to have been conceived as autonomous by Boeing's engineers, it was in effect autonomous, taking decisions about stall without involving the pilots. It exhibited the handover problem (CBAR) and the preliminary report on the Indonesian crash[27] shows that the pilots disengaged/ over-rode MCAS on multiple occasions, but ultimately unsuccessfully (although the same problem occurred during a flight with the same aircraft on the previous day, and this was successfully managed by the pilots).

As well as technical issues, there are ethical queries about the design (and operation) of the aircraft which will be clarified once the final report on the Indonesian and Ethiopian crashes are published .

There has also been commentary in the media, e.g., by the Financial Times , about the way in which the aircraft was certified and whether or not the Federal Aviation Authority (FAA) relied too much on Boeing for certification. This is a governance issue, and also an ethical one. Further, this highlights the last issue – who decides about deployment, and what role should the developer have? There is now to be an audit into the way the aircraft was initially certified[30] and it might be that this will provide a wider insight into ethical governance of RAS.



[27] It would be unwise – unethical perhaps – to speculate further on the underlying causes until these reports are published.

[28] https://www.ft.com/content/715ccc92-4a7a-11e9-8b7f-d49067e0f50d (accessed April 2019)

[29] https://www.transportation.gov/sites/dot.gov/files/docs/briefing-room/334391/memorandum-secretary-audit-certification-boeing-737-max8-2012-2017.pdf (accessed April 2019)

CONCLUSIONS

There are unusual challenges in ethics for RAS. Perhaps the issue can best be summarised as needing to consider "technically informed ethics". The technology of RAS raises issues that have an ethical dimension, and perhaps uniquely so due to the possibility of moving human decision-making which is implicitly ethically informed to computer systems. Further, if seeking solutions to these problems – ethically aligned design, to use the IEEE's terminology – then the solutions must be technically meaningful, capable of realisation, capable of assurance, and suitable as a basis for regulation.

Thus, ethics for RAS is a rich, complex multi-disciplinary concern, and perhaps more complex than many other ethical issues facing society today. It is also fast-moving. This paper has endeavoured to give an accessible introduction to some of the key issues, noting that many of them are quite subtle, and it is not possible to do them full justice in such a short document. However, we have sought to counterbalance this by giving an extensive list of initiatives, standards, etc. that focus on ethics of RAS and AI, see Annex A.

Finally, it is perhaps worthwhile making an observation about terminology. The AI community uses the term "AI safety" in a way that is quite different to how a safety engineer would consider safety of a RAS. A paper from some leading AI researchers identifies "concrete problems in AI safety"[31]. Whilst this is a different conception of safety, the concerns identified are relevant to our ethical considerations as they indicate ways in which AI (reinforcement learning in particular, in this paper) may produce undesired results. A key observation is that this shows that the RAS, AI and safety engineering communities need to work together on a range of issues, including ethics and ethically informed design.

[30] https://arxiv.org/abs/1606.06565 (accessed April 2019)

ANNEX A:

Selective Bibliography

Note that this bibliography includes material on AI which does not directly refer to RAS. However, the authors take the view that anything related to AI ethics potentially applies to RAS since such systems can be regarded as "embodied AI". For ease of use, the majority of the entries in the pdf version of this report contain web links. These links were all verified prior to publishing the paper; for brevity we omit the "accessed on" information for these links.

A.1 Ethical principles:

Asimov's three laws of Robotics (1950)

For completeness – noting that Asimov was the first to establish the principle that robots should be governed by principles.

Murphy and Wood's three laws of Responsible Robotics (2009)

These were proposed in Robin Murphy and David Wood's paper Beyond Asimov: The Three Laws of Responsible Robotics.

EPSRC Principles of Robotics (2010)

These principles were drafted in 2010 and published online in 2011, but not formally published until 2017 as part of a two-part special issue of Connection Science on the principles, edited by Tony Prescott & Michael Szollosy. An accessible introduction to the EPSRC principles was published in New Scientist in 2011.

Future of Life Institute

Asilomar principles for beneficial AI (2017)

The ACM US Public Policy Council

Principles for Algorithmic Transparency and Accountability (Jan 2017)

See the ACM announcement of these principles.

The principles form part of the ACM's updated code of ethics.

Japanese Society for Artificial Intelligence (JSAI)

Ethical Guidelines (Feb 2017)

Draft principles of The Future Society's Science, Law and Society Initiative (Oct 2017)

An article by Nicolas Economou explains the 6 principles with a full commentary on each one.

Intel's recommendation for Public Policy Principles on AI (October 2017)

These principles were announced in a blog post by Naveen Rao (Intel VP AI).

Montréal Declaration for Responsible AI draft principles (Nov 2017)

The Montréal Declaration for Responsible AI proposes 7 values and draft principles above. (full with preamble, questions and definitions).

UNI Global Union Top 10 Principles for Ethical AI (Dec 2017)

Drafted by UNI Global Union's Future World of Work these 10 principles for Ethical AI "provide unions, shop stewards and workers with a set of concrete demands to the transparency, and application of AI".

Lords Select Committee 5 core principles to keep AI ethical (April 2018)

These principles appear in the UK House of Lords Select Committee on Artificial Intelligence report AI in the UK: ready, willing and able? published in April 2019. The WEF published a summary and commentary.

AI UX: 7 Principles of Designing Good AI Products, April 2018

These principles, focussed on the design of the User Interface (UI) and User Experience (UX), are from Budapest based company UX Studio.

The Toronto Declaration on equality and non-discrimination in machine learning systems (May 2018)

The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems does not succinctly articulate ethical principles but instead presents arguments to address concerns "...about the capability of [machine learning] systems to facilitate intentional or inadvertent discrimination against certain individuals or groups of people".

Google AI Principles (June 2018)

These principles were launched with a blog post and commentary by Google CEO Sundar Pichai.

IBM's 5 ethical AI principles (September 2018)

For a full account read IBM's Everyday Ethics for Artificial Intelligence.

Microsoft Responsible bots: 10 guidelines for developers of conversational AI, Nov 2018

Microsoft's guidelines for the ethical design of 'bots' (chatbots or conversational AIs).

CEPEJ European Ethical Charter on the use of artificial intelligence (AI) in judicial systems and their environment, see 5 principles, Feb 2019

The Council of Europe ethical charter principles are outlined here.

Women Leading in AI (WLinAI) 10 recommendations launch Feb 2019

Presented by the Women Leading in AI group at a meeting in parliament in February 2019, a report in Forbes by Noel Sharkey outlines both the group, their recommendations, and the meeting.

The NHS's 10 Principles for AI + Data, Feb 2019

These principles are set out with full commentary and elaboration on Artificial Lawyer.

IEEE General Principles of Ethical Autonomous and Intelligent Systems (A/IS) (March 2019)

These amended and extended general principles form part of Ethical Aligned Design 1st edition, published in March 2019. For an overview see pdf here.

The ethical issues arising from the police use of live facial recognition technology, March 2019.

The UK government's independent Biometrics and Forensics Ethics Group (BFEG) published an interim report outlining nine ethical principles forming a framework to guide policy on police facial recognition systems.

Floridi and Clement Jones, The five principles key to any ethical framework for AI

Luciano Floridi and Lord Tim Clement Jones set out in the New Statesman, these 5 general ethical principles for AI, with additional commentary.

The European Commission's High Level Expert Group on AI Ethics Guidelines for Trustworthy AI (April 2019)

Published on 8 April 2019, the EU HLEG AI ethics guidelines for trustworthy AI are detailed in full.

Draft core principles of Australia's Ethics Framework for AI (April 2019)

These draft principles are detailed in Artificial Intelligence Australia's Ethics Framework A Discussion Paper. This comprehensive paper includes detailed summaries of many of the frameworks and initiatives listed above, together with some very useful case studies.

The ethical principles referenced above are listed in full here: http://alanwinfield.blogspot.com/2019/04/an-updated-round-up-of-ethical.html

A.2 Substantive AI ethics initiatives

The EURON Roboethics Atelier (2005)

The Future of Life Institute (2015/16)

The Foundation for Responsible Robotics (Dec 2015)

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (April 2016)

The Partnership on AI (Sept 2016)

AI 4 All (March 2017)

Montreal AI ethics institute (July 2017)

The AI Now Institute, New York University (Nov 2017)

The Institute for Ethical AI & Machine Learning (UK, 2018)

The Institute for Ethical Artificial Intelligence in Education (UK, Oct 2018)

Institute for Ethics in Artificial Intelligence (Germany, Jan 2019)

Introduction to the Centre for Data Ethics and Innovation (gov.uk) https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/787205/CDEI_Introduction-booklet.pdf

Saidot: Enabling responsible AI ecosystems https://www.saidot.ai/

A.3 Robotics and AI ethics standards and regulation

British Standard BS 8611 (2016) Guide to the Ethical Design of Robots and Robotic Systems https://shop.bsigroup.com/ProductDetail?pid=000000000030320089

IEEE 'human' standards currently in draft:

P7000 – Model Process for Addressing Ethical Concerns During System Design

P7001 – Transparency of Autonomous Systems

P7002 – Data Privacy Process

P7003 – Algorithmic Bias Considerations

P7004 – Standard for Child and Student Data Governance

P7005 – Standard for Transparent Employer Data Governance

P7006 – Standard for Personal Data Artificial Intelligence (AI) Agent

P7007 – Ontological Standard for Ethically Driven Robotics and Automation Systems

P7008 – Standard for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems

P7009 – Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems

P7010 – Wellbeing Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems

P7011 – Standard for the Process of Identifying and Rating the Trustworthiness of News Sources

P7012 – Standard for Machine Readable Personal Privacy Terms

P7013 – Inclusion and Application Standards for Automated Facial Analysis Technology

See these two articles on ethical standards in robotics and AI:

Bryson and Winfield (2017) in IEEE Computer https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7924235

Winfield (2019) In Nature Electronics https://www.nature.com/articles/s41928-019-0213-6

(preprint here https://www.researchgate.net/publication/331138667_Ethical_standards_in_robotics_and_AI)

The Open Community for Ethics in Autonomous and Intelligent Systems (OCEANIS) https://ethicsstandards.org/

A.4 National/international (governmental) Strategies

For an excellent roundup of national strategic initiatives in AI to date, see

https://medium.com/politics-ai/an-overview-of- national-ai-strategies-2a70ec6edfd

Note that a US initiative has been launched since this article.

A Proposed Model Artificial Intelligence Governance Framework, Singapore. https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/A-Proposed-Model-AI-Governance-Framework-January-2019.pdf This has been winning awards.

See also EC HLEG report below.

A.5 Major reports

UK Commons Select committee inquiry on robotics and AI, 2016:

https://www.parliament.uk/business/committees/committees-a-z/commons-select/science-and-technology-committee/inquiries/parliament-2015/robotics-and-artificial-intelligence-inquiry-15-16/

Lords Select committee inquiry on AI AI in the UK: ready, willing and able?, 2017, final report: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf

Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems European Group on Ethics in Science and New Technologies, March 2018 https://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf

Monitoring the evolution and benefits of Responsible Research and Innovation in Europe, Policy Brief for the EC by Jack Stilgoe, Oct 2018 https://ec.europa.eu/programmes/horizon2020/en/news/monitoring-evolution-and-benefits-responsible-research-and-innovation-morri

Nuffield/CFI report Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research, 2019: https://www.nuffieldfoundation.org/sites/default/files/files/Ethical-and-Societal-Implications-of-Data-and-AI-report-Nuffield-Foundat.pdf

The Topol Review (NHS HEE): Preparing the healthcare workforce to deliver the digital future, Feb 2019 https://topol.hee.nhs.uk/ (covers the application of robotics and AI in healthcare).

IEEE Ethical Aligned Design 1st Edition – note that the 1st 2 versions (2016 and 2017) were drafts: https://ethicsinaction.ieee.org/ March 2019, Important and comprehensive – this initiative has spun out 14 new standards working groups d date.

AI Sustainability report, April 2019 http://www.aisustainability.org/wp-content/uploads/2019/04/SUSTAINABLE-AI.pdf

Note the AI sustainability centre is also another initiative.

EC High Level Expert Group on AI report and recommendations: Ethics guidelines for trustworthy AI, 8 April 2019. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai Note: hugely important but already controversial.

A.6 Selected academic works:

[1] Boddington, P., Millican, P. & Wooldridge, M. (eds) Special Issue: Ethics and Artificial Intelligence, Minds & Machines 27: 569, Springer 2017. https://doi.org/10.1007/s11023-017-9449-y

[2] Boddington, P, Towards a Code of Ethics for Artificial Intelligence, Springer 2017.

[3] Cath, C (ed), Governing artificial intelligence: ethical, legal and technical opportunities and challenges. Philosophical Transactions of the Royal Society A 376, 2018. http://doi.org/10.1098/rsta.2018.0080

[4] Dignum, V (ed) Ethics in Artificial Intelligence, Ethics Inf. Technol. 20:1, Springer 2018. https://doi.org/10.1007/s10676-018-9450-z

[5] Lin P, Jenkins R and Abney K (eds) Robot Ethics 2.0, Oxford University Press 2017.

[6] Müller, Vincent C. (forthcoming), 'Ethics of artificial intelligence and robotics', in Edward N. Zalta (ed.), Stanford Encyclopedia of Philosophy (Palo Alto: CSLI, Stanford University). https://plato.stanford.edu

[7] Prescott T & Szollosy M (eds) (2017) Ethical principles of robotics, Connection Science, 29:2 and 29:3, DOI: 10.1080/09540091.2017.1312800

[8] Winfield AF, Michael K, Pitt J and Evers, V (eds) (2019) Machine ethics: The design and governance of ethical AI and autonomous systems. Proceedings of the IEEE, 107 (3). pp. 509-517. ISSN 0018-9219

Ethics for RAS is a rich, complex multi-disciplinary concern, and perhaps more complex than many other ethical issues facing society today

UK-RAS NETWORK
ROBOTICS & AUTONOMOUS SYSTEMS

www.ukras.org