Responsible AI – Key Themes, Concerns & Recommendations for European Research and Innovation

Summary of Consultation with Multidisciplinary Experts

Version 1.0 - June 2018

Steve Taylor[1], Brian Pickering, Michael Boniface, University of Southampton IT Innovation Centre, UK

Michael Anderson, Professor Emeritus of Computer Science at University of Hartford, USA & http://www.machineethics.com/

David Danks, L.L. Thurstone Professor of Philosophy & Psychology, Carnegie Mellon University

Dr Asbjørn Følstad, Senior Research Scientist, SINTEF, NO

Dr. Matthias Leese, Senior Researcher, Center for Security Studies, ETH Zurich, CH

Vincent C. Müller, University Academic Fellow, Interdisciplinary Ethics Applied Centre (IDEA), School of Philosophy, Religion and History of Science, University of Leeds, UK

Tom Sorell, Professor of Politics and Philosophy, University of Warwick, UK

Alan Winfield, Professor of Robot Ethics, University of the West of England

Dr Fiona Woollard, Associate Professor of Philosophy, University of Southampton, UK

[1] Contact author: sjt@it-innovation.soton.ac.uk

HUB4NGI

https://www.hub4ngi.eu https://www.ngi.eu/

This document's purpose is to provide input into the advisory processes that determine European support for both research into Responsible AI; and how innovation using AI that takes into account issues of responsibility can be supported. "Responsible AI" is an umbrella term for investigations into legal, ethical and moral standpoints of autonomous algorithms or applications of AI whose actions may be safety-critical or impact the lives of citizens in significant and disruptive ways. To address its purpose, this document reports a summary of results from a consultation with cross-disciplinary experts in and around the subject of Responsible AI.

The chosen methodology for the consultation is the Delphi Method, a well-established pattern that aims to determine consensus or highlight differences through iteration from a panel of selected consultees. This consultation has resulted in key recommendations, grouped into several main themes:

- *Ethics* (ethical implications for AI & autonomous machines and their applications);
- *Transparency* (considerations regarding transparency, justification and explicability of AI & autonomous machines' decisions and actions);
- *Regulation & Control* (regulatory aspects such as law, and how AI & automated systems' behaviour may be monitored and if necessary corrected or stopped);
- *Socioeconomic Impact* (how society and the economy are impacted by AI & autonomous machines);
- *Design* (design-time considerations for AI & autonomous machines) and
- *Responsibility* (issues and considerations regarding moral and legal responsibility for scenarios involving AI & autonomous machines).

The body of the document describes the consultation methodology and the results in detail. The recommendations arising from the panel are discussed and compared with other recent European studies into similar subjects. Overall, the studies broadly concur on the main themes, and differences are in specific points. The recommendations are presented in a stand-alone section "Summary of Key Recommendations", which serves as an Executive Summary.

Acknowledgements

The authors would like to thank Professor Kirstie Ball, Professor Virginia Dignum, Dr William E. S. McNeill, Professor Luis Moniz Pereira, Professor Thomas M Powers and Professor Sophie Stalla-Bourdillon for their valuable contributions to this consultation.

This report is supported by the "A Collaborative Platform to Unlock the Value of Next Generation Internet Experimentation" (HUB4NGI) project under EC grant agreement 732569.

Disclaimer

Summary of Key Recommendations

This document's purpose is to provide input into the advisory processes that determine European support for both research into Responsible AI; and how innovation using AI that takes into account issues of responsibility can be supported. "Responsible AI" is an umbrella term for investigations into legal, ethical and moral standpoints of autonomous algorithms or applications of AI whose actions may be safety-critical or impact the lives of citizens in significant and disruptive ways.

The recommendations listed here are the results from a consultation with cross-disciplinary experts in and around the subject of Responsible AI. The chosen methodology for the consultation is the Delphi Method, a well-established pattern that aims to determine consensus or highlight differences through iteration from a panel of selected consultees. The consultation has highlighted a number of key issues, which are summarised in the following figure grouped into six main themes.

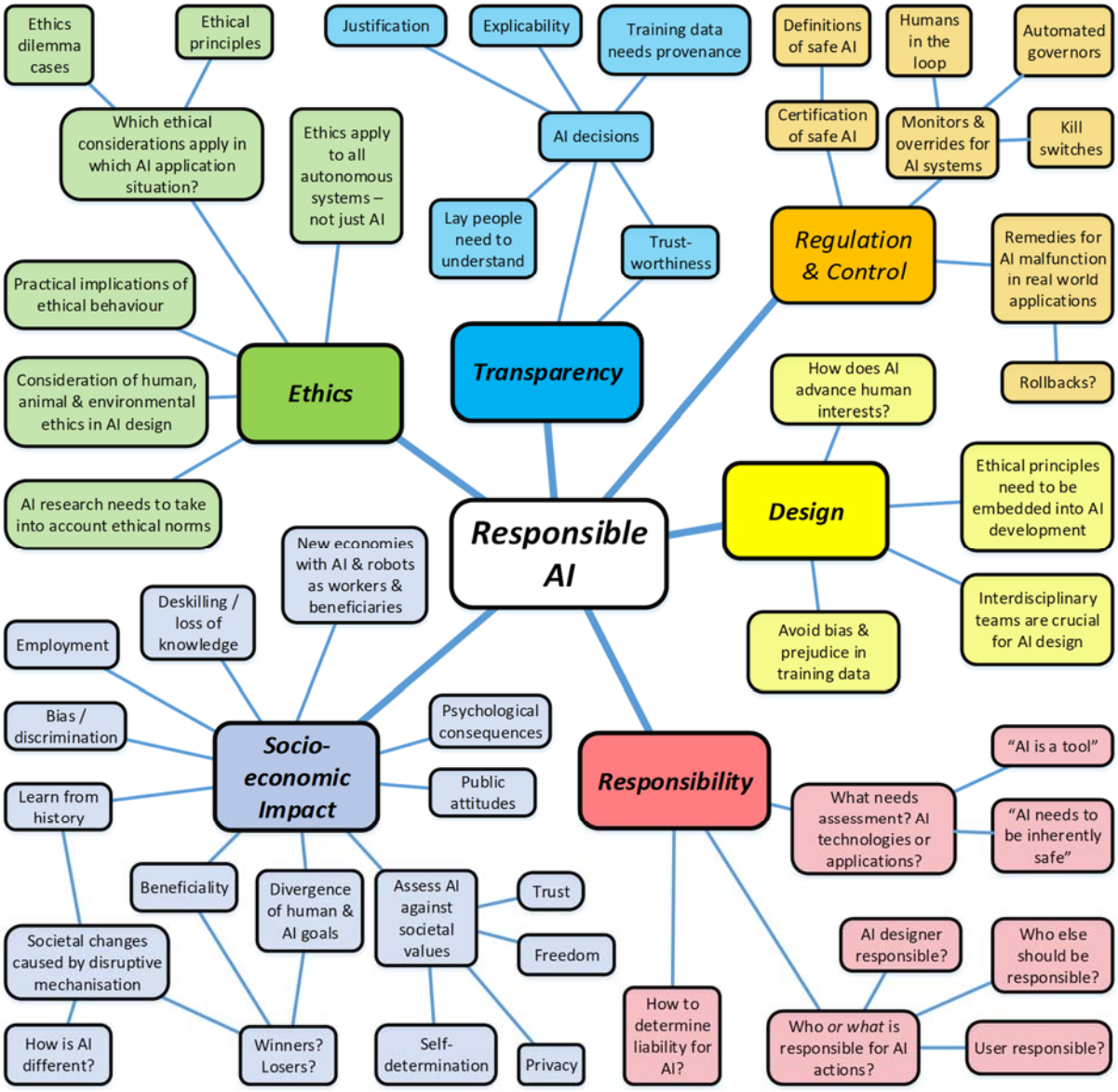

FIGURE 1: RESPONSIBLE AI - KEY AREAS AND ISSUES

Recommendations have been determined from the issues in order to help key stakeholders in AI research, development and innovation (e.g. researchers, application designers, regulators, funding bodies etc.) and these are discussed next, categorised into the same themes.

Ethics

Because of AI's disruptive potential, there are significant, and possibly unknown, ethical implications for AI & autonomous machines, as well as their applications.

- AI research needs to be guided by established ethical norms, and research is needed into new ethical implications of AI, especially considering different application contexts.
- The ethical implications of AI need to be understood and considered by AI researchers and AI application designers.
- The ethical principles that are important may depend strongly on the application context of an AI system, so designers need to understand the expected contexts of use and design with the ethical considerations they give rise to accordingly.
- Ethical principles need not necessarily be explicitly encoded into AI systems, but it is necessary that designers observe ethical norms and consider the ethical impact of an AI system at design time.
- Ethical and practical considerations need to be both considered at an AI system's design time, since they can both affect the design. They may be interdependent, and they may conflict.
- Assessment of the ethical impacts of a machine needs to be undertaken by the moral agent responsible for it. At design time, the responsible moral agent is most likely the designer. At usage time, the responsible moral agent may be the user, and the impacts may depend on the application context.

Transparency

Considerations regarding transparency, justification and explicability of AI & autonomous machines' decisions and actions are strongly advocated by the panel, in concert with others in the community.

- AI decisions and actions need to be transparent, explained and justified; and the explanation needs to be comprehensible by lay people as AI systems become more exposed to the general public.
- Provenance information regarding both AI decisions and their input data (as well as any training data) needs to be recorded in order to provide an audit trail for an AI decision.
- Trustworthiness of an AI system is critical for its widespread acceptance. Transparent justification of an AI system's decisions, as well as other factors such as provenance information for its training data, a track record of reliability and comprehensibility of its behaviour, all contribute to trustworthiness.

Regulation & Control

Investigation into regulatory aspects such as law, guidelines and governance is needed – specifically applied to new challenges presented by AI and automated systems. In addition, control aspects need

investigation – specifically concerning how AI & automated systems' behaviour may be monitored and if necessary corrected or stopped.

- Certification of "safe AI" and accompanying definitions of safety criteria are recommended. The application context determines the societal impact of an AI system so the safety criteria and resulting certification are likely to depend on the application the AI is put to. New applications of existing AI technology may need new assessment and certification.
- Determination of remedial actions for situations when AI systems malfunction or misbehave is recommended. Failure modes and appropriate remedial actions may already be understood, depending on the application domain where AI is being deployed (e.g. which emergency procedures are needed when a self-driving car crashes may very similar to those needed when a human-driven car crashes), but investigation is needed into what existing remedial actions are appropriate in what situation and whether they need to be augmented.
- An important type of control is human monitoring and constraint of AI systems' behaviour, up to and including kill switches that completely stop the AI system, but these governing mechanisms must fail safe.
- A further choice of control is roll-back of an AI system's decision, so that its direct consequences may be undone. It is recognised that there may also be side or unintended effects of an AI system's decision that may be difficult or impossible to undo, so careful assessment of the full set of implications of an AI system's decisions and actions should be undertaken at design time.
- Understanding of how the law can regulate AI is needed, and as with other fast-developing technology, the law lags technical developments. The application context may be a major factor in AI regulation, as the application context determines the effects of the AI on society and the environment.
- Even though there has been recent discussion of legal personhood for robots and AI, at the current time and for the foreseeable future, humans need to be ultimately liable for AI systems' actions. The question of which human is liable does need to be investigated however, and each application context may have different factors influencing liability.

Socioeconomic Impact

AI already has had, and will continue to have, disruptive impact on social and economic factors. The impacts need to be studied, to provide understanding of who will be affected, how they will be affected and how to guard against negative or damaging impacts.

- Understanding of the socioeconomic impacts of AI & autonomous machines on society is needed, especially how AI automation differs from other types of disruptive mechanisation.
- AI's impact on human workers needs to be investigated – how any threats or negative effects such as redundancy or deskilling can be addressed, as well as exploiting any benefits such as working in dangerous environments or performing monotonous tasks and reducing errors.
- Public attitudes towards AI need to be understood, especially concerning the factors that contribute to, and detract from, public trust of AI.

- Public attitudes are also connected with assessment of the threats that AI pose, especially when AI can undermine human values, so investigation is required into how and when AI is either compatible or conflicts with human values, and which specific ones.
- Research is needed into how users of AI can identify and guard against discriminatory effects of AI, for example how users (e.g. citizens) can be educated to recognise discrimination.
- Indirect social effects of AI need to be investigated, as an AI system's decisions may affect not just its users, but others who may not know that they are affected.
- How AI systems integrate with different types of networks (human, machine and human-machine) is an important issue – investigation is needed into an AI system's operational environment to determine the entities it interacts with and affects.
- There is unlikely to be a one-size-fits-all approach to social evaluation of AI and its applications – it is more likely the case that each application context will need to be evaluated individually for social impact, and research is needed on how this evaluation can be performed in each case.

Design

Design-time considerations & patterns for AI & autonomous machines need to be investigated, especially concerning what adaptations to existing design considerations and patterns are needed as a specific result of AI.

- Interdisciplinary teams are necessary for AI and application design to bring together technical developers with experts who can account for the societal, ethical and economic impacts of the AI system under design.
- Ethical principles and socioeconomic impact need to be considered from the outset of AI and application design.
- Whilst the AI design should have benefits for humankind at heart, there will also be cases where non-human entities (e.g. animals or the environment) may also be affected. Ethical principles apply to all kinds of nature, and this is not to be forgotten in the design process.
- Identification and recognition of any bias in training data is important, and any biases made clear to the user population.

Responsibility

Issues and considerations regarding moral and legal responsibility for scenarios involving AI & autonomous machines are regarded as critical, especially when automation is in safety-critical situations or has the potential to cause harm.

- Humans need to be ultimately responsible for the actions of today's AI systems, which are closer to intelligent tools than sentient artificial beings. This is in concert with related work that says, for current AI systems, humans must be in control and be responsible.
- Having established that (in the near term at least) humans are responsible for AI actions, the question of who is responsible for an AI system's actions needs investigation. There are standard mechanisms such as fitness for purpose where the designer is typically responsible, and permissible use where the user is responsible, but each application of an AI system may need a

separate assessment because different actors may be responsible in different application context. Indeed, multiple actors can be responsible for different aspects of an application context.

- Should the current predictions of Artificial General Intelligence[2] and Superintelligence[3] become realistic prospects, human responsibility alone may not be adequate and the concept of "AI responsibility" will need research by multidisciplinary teams to understand where responsibility lies when the AI participates in human-machine networks. This will need to include moral responsibility and how this can translate into legal responsibility.

[2] Pennachin, C. ed., 2007. *Artificial general intelligence* (Vol. 2). New York: Springer.
[3] Boström, N., 2014. *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Introduction

This report's purpose is to provide input into the advisory processes that determine European support for both research into Responsible AI; and how innovation using AI that takes into account issues of responsibility can be enabled. "Responsible AI" is an umbrella term for investigations into legal, ethical and moral standpoints of autonomous algorithms or applications of AI whose actions may be safety-critical or impact the lives of citizens in significant and disruptive ways.

This report is a summary of the methodology for, and recommendation resulting from, a consultation with a multidisciplinary international panel of experts into the subject of Responsible AI. Firstly, a brief background is presented, followed by a description of the consultation methodology. The results are then discussed, grouped into major themes and compared against other recent European studies in similar subject areas. Finally, brief conclusions are presented. The recommendations from this consultation are presented in the "Summary of Key Recommendations" section above, and the rest of this report serves to provide more detail behind the recommendations.

Background

As AI and automated systems have come of age in recent years, they promise ever more powerful decision making, providing huge potential benefits to humankind through their performance of mundane, yet sometimes safety critical tasks, where they can often perform better than humans[4,5]. Research and development in these areas will not abate and functional progress is unstoppable, but there is a clear need for ethical considerations applied to[6,7] and regulatory governance of[8,9] these systems, as well as AI safety in general[10] with well-publicised concerns over the responsibility and decision-making of autonomous vehicles[11] as well as privacy threats, potential prejudice or discriminatory behaviours of web applications[12,13,14,15]. Influential figures such as Elon Musk[16] and Stephen Hawking[17] have voiced concerns over the potential threats of undisciplined AI, with Musk describing AI as an existential threat to human civilisation and calling for its regulation. Recent studies into the next generation of the Internet such as

[4] Donath, Judith. The Cultural Significance of Artificial Intelligence. 14 December 2016. https://www.huffingtonpost.com/quora/the-cultural-significance_b_13631574.html
[5] Ruocco, Katie. Artificial Intelligence: The Advantages and Disadvantages. 6th February 2017. https://www.arrkgroup.com/thought-leadership/artificial-intelligence-the-advantages-and-disadvantages/
[6] Bostrom, N. & Yudowsky, E. (2014). The ethics of artificial intelligence. In Ramsey, W. & Frankish, K. (eds) *The Cambridge handbook of artificial intelligence*, 316-334.
[7] https://www.wired.com/story/ai-research-is-in-desperate-need-of-an-ethical-watchdog/
[8] Scherer, Matthew U., Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies (May 30, 2015). Harvard Journal of Law & Technology, Vol. 29, No. 2, Spring 2016. Available at SSRN: https://ssrn.com/abstract=2609777 or http://dx.doi.org/10.2139/ssrn.2609777
[9] Vincent C. Müller (2017) Legal vs. ethical obligations – a comment on the EPSRC's principles for robotics, Connection Science, 29:2, 137-141, DOI: 10.1080/09540091.2016.1276516
[10] https://futureoflife.org/2017/09/21/safety-principle/
[11] Bonnefon, J-F, Shariff, A. & Rahwan, I (2016). The social dilemma of autonomous vehicles. *Science* 352(6293), 1573-1576.
[12] http://www.independent.co.uk/news/world/americas/facebook-rules-violence-threats-nudity-censorship-privacy-leaked-guardian-a7748296.html
[13] http://www.takethislollipop.com/
[14] https://www.youtube.com/watch?v=4obWARnZeAs
[15] Crawford, K. (2016) "Artificial intelligence's white guy problem." The New York Times (2016).
[16] Musk, E. (2017) Regulate AI to combat 'existential threat' before it's too late. *The Guardian,* 17th July, 2017
[17] Stephen Hawking warns artificial intelligence could end mankind, BBC News, 2 December 2014. http://www.bbc.co.uk/news/technology-30290540

Overton[18] and Takahashi[19] concur that regulation and ethical governance of AI and automation is necessary, especially in safety critical systems and critical infrastructures.

Over the last decade, machine ethics has been a focus of increased research interest. Anderson & Anderson identify issues around increasing AI enablement not only in technical terms[20], but significantly in the societal context of human expectations and technology acceptance transplanting the human being making the ethical choice with an autonomous system[21]. Anderson & Anderson also describe different mechanisms for reasoning over machine ethics[20]. Some mechanisms concern the encoding of general principles (e.g. principles following the pattern of Kant's categorical imperatives[22]) or domain-specific ethical principles, while others concern the selection of precedent cases of ethical decisions in similar situations (e.g. SIROCCO[23]) and a further class considers the consequences of the action under question (act utilitarianism – see Brown[24]). An open research question concerns which mechanism, or which combination of mechanisms, is appropriate.

A long-debated key question is that of legal and moral responsibility of autonomous systems. Who or what takes responsibility for an autonomous system's actions? Calverley[25] considers the question from a legal perspective, asking whether a non-biological entity can be regarded as a legal person. If a non-biological entity such as a corporation can be regarded as a legal person, then why not an AI system? The question then becomes one of intentionality of the AI system and whether legal systems incorporating penalty and enforcement can provide sufficient incentive to AI systems to behave within the law. Matthias[26] poses the question whether the designer of an AI system can be held responsible for the system they create, if the AI system learns from its experiences, and therefore is able to make judgements beyond the imagination of its designer. Beck[27] discusses the challenges of ascribing legal personhood to decision making machines, arguing that society's perceptions of automata will need to change should a new class of legal entity appear.

Transparency of autonomous systems is also of concern, especially given the opaque (black-box) and non-deterministic nature of AI systems such as Neural Networks. The so-called discipline of "explainable AI" is not new: in 2004, Van Lent et al[28] described an architecture for explainable AI within a military

[18] DAVID OVERTON, NEXT GENERATION INTERNET INITIATIVE – CONSULTATION - FINAL REPORT MARCH 2017 https://ec.europa.eu/futurium/en/content/final-report-next-generation-Internet-consultation
[19] Takahashi, Makoto. Policy Workshop Report Next Generation Internet - Centre for Science and Policy Cambridge Computer Laboratory. Centre for Science and Policy (CSaP) in collaboration with the Cambridge Computer Laboratory. 1-2 March 2017. https://ec.europa.eu/futurium/en/system/files/ged/report_of_the_csap_policy_workshop_on_next_generation_Internet.docx. Retrieved 2017-06-19.
[20] Anderson, M., & Anderson, S. L. (Eds.). (2011). *Machine ethics*. Cambridge University Press.
[21] Anderson, Michael, and Susan Leigh Anderson. "Machine ethics: Creating an ethical intelligent agent." AI Magazine 28, no. 4 (2007): 15. https://doi.org/10.1609/aimag.v28i4.2065
[22] https://plato.stanford.edu/entries/kant-moral/
[23] McLaren, Bruce M. "Extensionally defining principles and cases in ethics: An AI model." Artificial Intelligence 150, no. 1-2 (2003): 145-181. https://doi.org/10.1016/S0004-3702(03)00135-8
[24] Brown, Donald G. "Mill's Act-Utilitarianism." The Philosophical Quarterly 24, no. 94 (1974): 67-68.
[25] Calverley, D.J., 2008. Imagining a non-biological machine as a legal person. Ai & Society, 22(4), pp.523-537.
[26] Matthias, A., 2004. The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and information technology, 6(3), pp.175-183.
[27] Beck, S., 2016. The problem of ascribing legal responsibility in the case of robotics. AI & society, 31(4), pp.473-481.
[28] Van Lent, Michael, William Fisher, and Michael Mancuso. "An explainable artificial intelligence system for small-unit tactical behavior." In Proceedings of the National Conference on Artificial Intelligence, pp. 900-907. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004.

context and in 2012, Lomas et al[29] demonstrated a system that allows a robot to explain its actions by answering "why did you do that?" types of question. More recently, in response to fears of accountability for automated and AI systems, the field of *algorithmic accountability reporting* has arisen "*… as a mechanism for elucidating and articulating the power structures, biases, and influences that computational artefacts exercise in society*"[30]. In the USA, the importance of AI transparency is clearly identified, with DARPA recently proposing a work programme for research towards explainable AI (XAI)[31,32].

The above issues and others are encapsulated in the "Asilomar AI Principles" [33], a unifying set of principles that are widely supported and should guide the development of beneficial AI, but how should these principles be translated into recommendations for European research into the subject of responsible AI and innovation of responsible AI applications? To provide answers these questions, a consultation has been conducted and its results are compared against other relevant and recent literature in this report.

Methodology

Consultation Methodology

The consultation used the Delphi Method[34], a well-established pattern that aims to determine consensus or highlight differences from a panel of selected consultees. These properties make the Delphi Method ideally suited for the purposes of targeted consultations with experts with the intention of identifying consensuses for recommendations.

The Delphi Method arrives at consensus by iterative rounds of consultations with the expert panel. Initial statements made by participants are collated with other participants' statements and presented back to the panel for discussion and agreement or disagreement. This process happens over several rounds, with subsequent rounds refining the previous round's statements based on feedback from the panel so that a consensus is reached, or controversies highlighted. This consultation used three rounds:

- *Round 1*. A selected panel of experts were invited to participate based on their reputation in a field relevant to the core subject of this consultation. Round 1 was a web survey containing a background briefing note to set the scene, accompanied by two broad, open-ended questions to which participants made responses in free-form text.

- *Round 2*. Using the standard qualitative technique of thematic analysis[35], the collected corpus of responses from Round 1 were independently coded to generate assertions that were presented back to the participants. Broad themes were also identified from the corpus, which were used as

[29] Lomas, Meghann, Robert Chevalier, Ernest Vincent Cross II, Robert Christopher Garrett, John Hoare, and Michael Kopack. "Explaining robot actions." In Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction, pp. 187-188. ACM, 2012. https://doi.org/10.1145/2157689.2157748

[30] Diakopoulos, N., 2015. Algorithmic accountability: Journalistic investigation of computational power structures. Digital Journalism, 3(3), pp.398-415.

[31] DARPA 2016 - Broad Agency Announcement - Explainable Artificial Intelligence (XAI) DARPA-BAA-16-53, August 10, 2016. https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf

[32] Gunning, David. "Explainable artificial intelligence (xai)." Defense Advanced Research Projects Agency (DARPA), nd Web (2017).

[33] The Asilomar AI Principles, proposed during the Beneficial AI 2017 Conference, Asilomar, California, 5-8 January 2017. https://futureoflife.org/ai-principles/

[34] Linstone, H.A. and Turoff, M. eds., 1975. The Delphi method: Techniques and applications (Vol. 29). Reading, MA: Addison-Wesley.

[35] Braun & Clarke (2006) DOI: 10.1191/1478088706qp063oa

groupings for the assertions. The participants evaluated each assertion, marking their agreement or disagreement (using a 5-point Likert scale[36]) and made comments in free-form text.

- *Round 3*. The results of Round 2 were collated. Those assertions that had significant majority sharing the same answer polarity (agree / disagree) were regarded as reaching consensus. The remainder, where opinion was more divided, were re-formulated into new assertions based on a thematic analysis of the comments and presented back to the panellists who could then agree / disagree and comment as before.

The Round 3 results that reached consensus were collated with those from Round 2 to determine the final consensus and disagreements of recognised experts in multiple relevant disciplines. The output recommendations of the consultation are a direct reflection of their views, therefore.

Expert Selection & Invitation

It was decided that a good target for the number of experts in the panel was 10-20, the reasoning for this range being a balance between adequate coverage of subjects and manageability. It was acknowledged that experts are busy people and as a result we assumed a 10-20% response rate, so to achieve the desired expert numbers of 10-20 experts in the panel, we aimed to invite 80-100 experts.

In order to determine relevant subject fields of expertise and hence candidate experts for the panel, a "knowledge requirements" exercise was performed in the form of an exploratory literature survey. Starting point searches included the subject areas of *Legislation & Regulation of AI*, *Ethics of AI*, *Legal and Moral Responsibility of AI*, and *Explainable and Transparent AI*. Key works and experts, as well as related search terms were found using standard tools and methods such as Google Scholar, Microsoft Academic, standard Google searches and following links from Wikipedia pages to gain a background into the theme, related themes, as well as influential people contributing important work within the theme.

The result of these investigations was a spreadsheet describing names of experts, their affiliation contact details, with notes on their specialisms. A total of 88 experts roughly evenly distributed across the subject areas above were invited to the consultation. Participants were therefore drawn from a purposive sample, based on their academic standing and reputation within a given area of expertise.

Ethical approval for the consultation was sought from the Faculty of Physical Science and Engineering at the University of Southampton and approved[37]. The application contained aspects such as disclosure of the purposes of the consultation, data protection, anonymity, risk assessment and consent.

A briefing note[38] was created, describing the background to the consultation via a literature survey[39], and in this two key questions were asked to begin the consultation:

- *What research is needed to address the issues that Beneficial AI and Responsible Autonomous Machines raise?*
- *Why is the recommended research important?*

[36] "*Strongly Agree*", "*Agree*", "*Disagree*" and "*Strongly Disagree*", with an additional "*Not Relevant*" option.
[37] University of Southampton ERGO number: 30743
[38] Hosted at https://www.scribd.com/document/362584972/Responsible-Autonomous-Machines-Consultation-Background-Gateway-Questions
[39] The content of the briefing note forms the basis of the "Background" section of this document.

The briefing note was sent to the 88 targeted experts, with a link to an online survey where they could make their responses.

Analysis

A total of 12 experts responded in detail to Round 1. This panel comprised experts in the following subject areas:

- Algorithmic Accountability & Explainable AI
- AI Applications
- Bias in Automated Systems
- Epistemology
- Law (specifically applied to computation)
- Machine Ethics
- Philosophy
- Robotics
- Sociology

Round 1 responses were in the form of free-form text, answering the two questions posed in the briefing note. The aim of Round 1 analysis was to determine assertion statements from the textual responses that could be used as input for Round 2. Two researchers coded the original text independently, and a standard thematic analysis methodology (TA) was adopted. In an entirely inductive application of the technique, each respondent's textual answers were scrutinised for opinions, statements and recommendations. Where one of these was found, the relevant quotation from the text was recorded along with a summary to form a draft assertion. Many cases were found where different respondents expressed the same opinion, albeit worded differently. All concordant opinions were clustered into a single summary assertion, recording the associated quotations and how many participants expressed that opinion. Once the assertions were determined, broad themes were identified to serve as coarse-grained groups for the assertions, and to highlight the key issues.

The researchers met to discuss and agree the final set of themes and assertions. The overlap of interim themes was good (4 out of 6 themes were clearly the same). The union set of assertions from the independent analyses was discussed, and it was found that the majority of assertions appeared in both analyses (albeit in different forms). Each assertion was discussed and modified as necessary so that the final set of assertions was agreed by both researchers. Because of this agreement, no formal analysis of inter-coder reliability[40] was therefore felt necessary. The resulting set of assertions was presented back to the panellists, and they were invited to agree or disagree with them in Round 2.

Ten experts responded to Round 2, and the responses comprised agreements and disagreements with the assertion statements, expressed in the structured format of a 5-point Likert Scale ("*Strongly Disagree*", "*Disagree*", "*Agree*" and "*Strongly Agree*", along with a "*Not Relevant*" option). Because of the response format, analysis of Round 2 was quantitative - counting the agreements and disagreements to each assertion. To determine whether consensus was reached, a simple metric was used that compared general agreement to general disagreement. The total "agreement" votes ("*Strongly Agree*" or "*Agree*") were compared to the "disagreement" votes ("*Strongly Disagree*" or "*Disagree*"), and if either group had more than twice the number of votes than the other, consensus was deemed to have been achieved. Out of the

[40] Qualitative research methods reliability of analysis is checked initially by checking agreement between two researchers ("coders") who attempt to identify categories and themes ("codes"). See, for example, Howitt, D. (2013) *Introduction to Qualitative Research Methods in Psychology*

Round 2 results, 22 assertions achieved consensus. Reviewing comments from participants, the remainder that did not achieve consensus were re-formulated, resulting in 18 derived assertions for Round 3.

Eight experts responded to Round 3 and selected whether they agreed with each of the 18 assertions presented to them. Similar to Round 2, the experts could also make optional comments. Out of the set of 18 assertions, 11 achieved consensus in Round 3. The 22 assertions from Round 2 and the 11 that reached consensus from Round 3 were combined, making **a total of 33 consensus items** over the course of the consultation. These make up the results reported here, and represent recommendations based on the cross-disciplinary perspective of recognised experts.

Results Summary & Discussion

This section contains 33 research priorities that have reached consensus from the consultation, divided into six themes. The priorities in each section are discussed and compared against three key recent European sources to highlight similarities and differences; and the differences may require further investigation. The sources comprise the European Commission's current approach regarding AI research and innovation, and current European priorities on the ethics and socioeconomic impacts of AI:

- "A European approach on Artificial Intelligence"[41] is a European Commission document that describes the current EC approach and plans for the development of AI and the assurance of safe and responsible AI. It will be hereafter referred to as the "***EC Approach***".
- The European Group on Ethics in Science and New Technologies (EGE)[42] is an independent advisory body of the President of the European Commission, and has published a Statement on "Artificial Intelligence, Robotics and 'Autonomous' Systems. The EGE Statement: "*calls for the launch of a process that would pave the way towards a common, internationally recognised ethical and legal framework for the design, production, use and governance of artificial intelligence, robotics, and 'autonomous' systems. The statement also proposes a set of fundamental ethical principles, based on the values laid down in the EU Treaties and the EU Charter of Fundamental Rights, that can guide its development*". "[43]. It will be hereafter referred to as the "***EGE Statement***".
- European Economic and Social Committee has issued an opinion statement on the socio-economic consequences of AI, entitled "Artificial intelligence – The consequences of artificial intelligence on the (digital) single market, production, consumption, employment and society"[44]. This will be hereafter referred to as the "***EESC Opinion***".

The following themes were derived from the thematic analysis of expert comments in Round 1. These serve as broad subject categories.

- ***Ethics*** (ethical implications for AI & autonomous machines and their applications);
- ***Transparency*** (considerations regarding transparency, justification and explicability of AI & autonomous machines' decisions and actions);

[41] European Commission, "A European approach on Artificial Intelligence", 25 April 2018. Available at: http://europa.eu/rapid/press-release_MEMO-18-3363_en.htm. Retrieved 2018-05-24.
[42] http://ec.europa.eu/research/ege/index.cfm
[43] European Group on Ethics in Science and New Technologies (EGE), "EGE Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems", March 2018. Available at: http://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf. Retrieved 2018-05-22.
[44] European Economic and Social Committee (EESC), "Artificial intelligence – The consequences of artificial intelligence on the (digital) single market, production, consumption, employment and society", May 2017. Available at: https://www.eesc.europa.eu/en/our-work/opinions-information-reports/opinions/artificial-intelligence. Retrieved 2018-06-05.

- ***Regulation & Control*** (regulatory aspects such as law, and how AI & automated systems' behaviour may be monitored and if necessary corrected or stopped);
- ***Socioeconomic Impact*** (how society is impacted by AI & autonomous machines);
- ***Design*** (design-time considerations for AI & autonomous machines) and
- ***Responsibility*** (issues and considerations regarding moral and legal responsibility for scenarios involving AI & autonomous machines).

Overall, there is broad agreement between the different studies, and this consultation's themes are shared with the other three studies. Each of the four initiatives covers a different subset of themes and to illustrate the overlaps and gaps, the following table maps the three external sources' areas of concern to this consultation's themes.

TABLE 1: COMPARISON OF KEY AREAS FROM DIFFERENT EUROPEAN AI STUDIES

EESC Opinion: Areas "Where AI Poses Societal Challenges"	***EC Approach***	***EGE Statement***	**This Consultation: Themes**
Safety	AI Alliance for the future of AI in Europe addresses safety	… *"safety, security, the prevention of harm and the mitigation of risks"*	This is not an explicit theme in the consultation, but safety is a key aspect of the "Regulation & Control" theme.
-	Regulation for liability	… *"human moral responsibility"*	Dedicated theme of "Responsibility"
Governance and regulation	Investigation into application of existing EU directives and regulations	… *"governance, regulation, design, development, inspection, monitoring, testing and certification"*	Dedicated themes of "Regulation & Control" and "Design"
Transparency and accountability	Algorithmic transparency	… *"explainability and transparency of AI and 'autonomous' systems"*	Dedicated theme of "Transparency"
Ethics	AI Alliance for the future of AI in Europe addresses ethical issues	The EGE statement is concerned with ethics in AI, Robotics and Autonomous Systems	Dedicated theme of "Ethics"
Education and skills	Support for EU upskilling to use new AI technologies	-	Deskilling and the loss of knowledge are covered in "Socioeconomic Impact"
(In)equality and inclusiveness	AI Alliance for the future of AI in Europe addresses inclusiveness	-	Discrimination is covered in "Socioeconomic Impact"
Work	-	-	Threats to employment are covered in "Socioeconomic Impact"
Privacy	GDPR & AI Alliance for the future of AI in Europe addresses privacy	-	Privacy is covered in "Socioeconomic Impact"
Warfare	-	Weapons and the principle of Meaningful Human Control	MHC is advocated in discussion of Responsibility

Superintelligence	-	-	Touched on in discussion of Responsibility
-	Support for Digital Innovation Hubs (DIH) to foster collaborative AI design	-	"Design" theme – design-time considerations

The following sections present the consultation's results in detail, grouped into the six themes. The assertion statements in each theme are presented in tabular form[45] with the votes in agreement and disagreement as well as the total votes cast for each assertion[46]. The order in which the assertion statements are presented corresponds to the strength of consensus amongst the panel, with the strongest consensus first. Following the table, each assertion is discussed. The discussion is centred on three major aspects: the strength of the consensus, issues raised by comments made by the panellists and any comparisons with the other three sources.

Ethics

ETHICS				
ID	**Assertion Statement**	**Disagree Votes**	**Agree Votes**	**Total Votes**
0	*AI research and technical choices need to take into account ethical implications and norms*	0	10	10
1	*Ethical choices and dilemmas faced by applications of AI need to be investigated, along with the factors that are relevant to them and the trade-offs between the factors in dilemma resolution*	1	9	10
2	*There is a deep inter-relationship between ethical and practical considerations in the design of autonomous machines*	1	8	10
3.1	*There needs to be a common understanding of what "autonomy" is or "autonomous capabilities" are*	2	6	8
3.2	*All machines (not just AI) qualify for ethical concern since they all have the potential to cause harm*	2	6	8

Assertion 0. The panel unanimously agreed that AI research should be guided by ethical norms and that AI application developers should consider ethical implications at design time. The spirit of this principle agrees with other EU studies. The EESC Opinion "*calls for a code of ethics for the development, application and use of AI*"[47], the EGE Statement states that: "'*Autonomous' systems should only be developed and used in ways that serve the global social and environmental good*"[48] and the EC Approach has plans to

[45] Each assertion has a numeric identifier, so each can be unambiguously identified. The format of the ID indicates which round an assertion reached consensus. If an assertion has an integer identifier, it reached consensus in Round 2. Round 3 assertions are derived from those that did not reach consensus and have decimal identifiers, for example assertion ID 3 did not reach consensus in Round 2, so it was replaced by assertions 3.1 and 3.2 in Round 3.

[46] The total votes for each assertion differs, and in most cases this is because some assertions reached consensus in different rounds, each of which had different numbers of participants. Round 2 had 10 participants and Round 3 had 8 participants. Some panellists did not vote for all assertions, so occasionally the total number of votes amounts to less than 10 in Round 2 and less than 8 in Round 3.

[47] EESC Opinion, page 3.

[48] EGE Statement, page 16.

implement a code of practice for ethics: *"Draft AI ethics guidelines will be developed on the basis of the EU's Charter of Fundamental Rights"*[49]. A key point made by some of the panel was to contrast ethical practices and observation of ethical norms at design time with explicit encoding of ethical principles within the resulting system. Some of the panel commented that ethical principles need not necessarily be explicitly encoded into AI systems, but AI designers need to be understanding of the ethical issues and potential impacts of their work and design accordingly.

Assertion 1. A strong consensus supporting the need for investigation into the ethical implications of the applications of AI was found in the panel, with 9 agreeing and 1 disagreeing. The only participant who disagreed with the assertion commented that their disagreement was founded in the assertion's emphasis on ethical dilemmas and trade-offs: *"I'm disagreeing with your summary statement, rather than the quotations. I don't like the idea of trade offs in ethical reasoning. It is important to consider the enabling and constraining factors on AI developments in respect of the foregrounding of human rights concerns"*. This sentiment was echoed by a participant who agreed with the assertion: *"While I strongly agree with this statement, I think that it is important not to become overly focused on "solving" dilemmas. I think that it is more important to think about the values that are being taught or realized in the AI technologies, and how those values will guide the system's behaviors"*. Both favour determination of the values and factors that guide ethical behaviour, above ethical dilemmas and trade-offs. These factors and values can support the considerations needed by designers of AI systems and applications referred to in Assertion 0 above.

Assertion 2. The need to consider the relationship between ethical and practical considerations at design time was also strongly supported by the panel, with 8 in agreement, 1 disagreeing and 1 not voting. The only comment (from a participant in agreement) concerned the specifics of the relationship and pointed out that ethical and practical considerations may be both complementary or contrary depending on the particular case, but they both independently influence the design choices: *"I think that design needs to consider both ethical and practical factors & constraints. Moreover, these do not necessarily stand in opposition - sometimes, the ethical thing is also the most practical or effective. However, they do not always have a clean or deep relationship with one another, but rather influence the technology in parallel."*

Assertion 3.1. The panel debated the meaning of "autonomy" in all three rounds of the consultation to give context to the concept of autonomous machines and their ethical implications. The result was broad agreement (6 agree, 2 disagree) but caveats were made about whether a shared understanding was important or even possible. A comment from a participant who agreed with the assertion was: *"I think that a shared understanding of various kinds of capabilities would help to advance the debates. I don't think that we need to worry about having a shared understanding of "autonomy" (full stop), nor do I think that such shared understanding would be possible"*, and comments from participants who disagreed were: *"I doubt this is possible. What is considered as part of an autonomy construct will likely diverge between disciplines, and will likely also evolve over time"* and *"A machine does not need autonomy to have ethical impact so it follows that it is not necessary for qualification of ethical concern"*.

Assertion 3.2. Related to (and prompted by) the discussion of autonomy was the assertion that all types of machine that should be eligible for "ethical concern" since they all have the potential to cause harm. Whilst 6 participants agreed with the assertion (compared to 2 who disagreed), the comments indicated that further investigation is likely to be required regarding who needs to be concerned and under what circumstances. A participant (who disagreed with the assertion statement) pointed out that some entities capable of causing harm are not in themselves moral agents: *"Much seems to depend on what we mean*

[49] EC Approach, page 3.

by "ethical concern." A meteorite has the potential to cause harm, but I don't normally think that it is subject to "ethical concern" - normally, I would think that would be reserved for other moral agents". A relevant example of a moral agent who should be concerned about the ethical impacts of a machine is its designer but there will clearly be other interested parties who will have ethical concerns. This touches on the aspect of responsibility for an AI system's actions, discussed later. In assessing the circumstances for ethical concern, the amount of potential harm is clearly important - a participant who disagreed with the assertion said: "*Of course the safety implications of any machine's operations are relevant if the operations take place with or near humans, but the harm caused by a runaway vacuum cleaner and the harm caused by a robot need not typically be the same, and some harms are the result of poorly understood algorithms rather than human error or carelessness. In short, AI for interactions with human beings does add new dimensions of harm and responsibility for harm*". Another participant (who agreed with the assertion statement) pointed out that potential for harm may not be the only factor that determines the need for ethical concern: "*It is not just their potential for harm that should be considered. It is there potential to have impact on any ethically relevant feature. For instance, even if a machine's impact is only ever good, the distribution of that good (concerns for justice) might be in question*".

Transparency

TRANSPARENCY				
ID	**Assertion Statement**	**Disagree Votes**	**Agree Votes**	**Total Votes**
4	*AI decisions need to be transparent, explained and justified*	1	9	10
5	*AI decisions need to be understood by lay people, not just technical experts*	2	8	10
6	*Transparency is needed for both data provenance and algorithmic decisions*	2	8	10
7	*Transparency requires that the autonomous machines meet at least two criteria: a track record of reliability and comprehensibility of previous behaviour*	2	8	10

Assertion 4. There was strong support (9 agree, 1 disagree) for the overarching statement that AI decisions need to be transparent, explained and justified. This is unsurprising, as transparency is overwhelmingly supported in the community. The EC Approach has identified it as a priority for near-term future plans: "*Algorithmic transparency will be a topic addressed in the AI ethics guidelines to be developed by the end of the year* [2018]"[50]. The EGE Statement views transparency from the perspective of enabling humans to be in overall control: "*All 'autonomous' technologies must, hence, honour the human ability to choose whether, when and how to delegate decisions and actions to them. This also involves the transparency and predictability of 'autonomous' systems, without which users would not be able to intervene or terminate them if they would consider this morally required*"[51]. Also taking the perspective of human control, the EESC Opinion "[…] *advocates transparent, comprehensible and monitorable AI systems, the operation of which is accountable, including retrospectively. In addition, it should be established which*

[50] EC Approach, page 2.
[51] EGE Statement, page 16.

decision-making procedures can and cannot be transferred to AI systems and when human intervention is desirable or mandatory"[52]. Other bodies strongly support AI transparency – one of the comments (from a participant who agreed with the assertion statement) referenced the IEEE standard on Transparency of Autonomous Systems (P7001)[53], so it clear that transparency is important in other closely-related areas.

Assertion 5. That AI decisions should be understandable to lay people as well as experts received majority support (8 agree, 2 disagree), and a key reason given amongst the supporters of the assertion statement was that the more people that understand an AI system's output, the more people will potentially trust it: "*But only because it is a good way to generate trust in the AI system. "Understandable AI" is not intrinsically better, but is instead a way to increase human-AI trust*". A further reason can be that the more people that understand an AI system's output, the more potential candidates there are for "human in the loop" control of it. Monitoring and control related to comprehensibility is alluded to in a comment from a participant disagreeing with the assertion statement: "*I do not believe they need to be understood, but that they to the greatest extent feasible should be explained and made available for check/revision*".

Assertion 6. Transparency contributing to both AI decisions and provenance of data has also received broad support (8 agree, 2 disagree). The contribution to AI decisions has already been discussed above, so can be taken as read here. Regarding the transparency of data provenance, no reason for the broad support are provided by the participants but given the context it is assumed that provenance is important in two main cases: training data for supervised learning systems, and input data. The EC Approach also emphasises the importance of high quality data in its near-term plans to support AI developers and deployers: "*The Commission will also support the uptake of AI across Europe, with a toolbox for potential users, focusing on small and medium-sized enterprises, non-tech companies and public administrations. The set of measures will include an EU 'AI-on-demand platform' giving advice and easy access to the latest algorithms and expertise; a network of AI-focused Digital Innovation Hubs facilitating testing and experimentation; and industrial data platforms offering high quality datasets*"[54].

Assertion 7. The assertion statement "*Transparency requires that the autonomous machines meet at least two criteria: a track record of reliability and comprehensibility of previous behaviour*" also had majority support with 8 agreeing and 2 disagreeing. Both factors contribute to trustworthiness of a system – they represent experience of a system's fitness for purpose, enabling a better-informed trust judgement than without them. A key comment by a participant who disagreed is: "*There are many ways to be transparent. Moreover, these two features are important for trust, not necessarily transparency. (And even for trust, they are not both required)*". Indeed, transparency of decision-making also contributes to its trustworthiness. Trust is discussed later, from the perspective of public trust in AI systems, and transparency is clearly a factor in public trust.

[52] EESC Opinion, page 7.
[53] IEEE Standards Association, Project 7001 - Transparency of Autonomous Systems. Available at https://standards.ieee.org/develop/project/7001.html. Retrieved 2018-06-05
[54] EC Approach, page 1.

REGULATION & CONTROL				
ID	**Assertion Statement**	**Disagree Votes**	**Agree Votes**	**Total Votes**
12	*Certification of reliably safe AI is needed, including definitions of criteria for safety*	0	10	10
14.1	*Achievable remedies should be agreed and in place for times when AI applications malfunction*	0	8	8
10	*Mechanisms that monitor and can constrain AI systems' behaviour are necessary, including stop or human override controls*	1	9	10
13.1	*Research is needed to identify factors that determine who or what is liable in different cases of AI application*	1	7	8
9	*Interdisciplinary research is needed to determine how law can ensure responsible behaviour*	2	8	10
11	*Research into whether and how AI systems' decisions or actions can be rolled back is needed*	2	8	10

Assertion 12. The panel unanimously agreed that certification for reliable and safe AI is needed, including definitions of criteria for safety. AI safety is obviously paramount, and other studies concur. The EGE Statement has a key question concerning: *"safety, security, the prevention of harm and the mitigation of risks"*, the EC Approach has tasked the newly-formed European AI Alliance with safety as a key issue, and the EESC Opinion has safety as a key area: *"The use of AI in the physical world undoubtedly gives rise to safety issues"*[55]. The external sources do not go as far as to recommend certification, but they do advocate testing and compliance to safety requirements. The EGE Statement specifically addresses the testing of AI for safety: *"All dimensions of safety must be taken into account by AI developers and strictly tested before release in order to ensure that 'autonomous' systems do not infringe on the human right to bodily and mental integrity and a safe and secure environment"*[56] and the EESC Opinion advocates compliance with safety requirements: *"The EESC believes that AI systems may only be used if they meet specific internal and external safety requirements. These requirements should be determined by AI and safety specialists, businesses and civil society organisations collectively"*. The comments made by the panellists for this consultation indicate that there are caveats to the principle of certification. Firstly, not all AI applications are safety-critical: *"For some AI applications this is obviously needed (e.g. self-driving cars), for others, this is not relevant (e.g. AI for playing the game go)"*[57]. Therefore, an assessment criterion is needed to determine which AI applications or domains qualify as safety-critical and so should be subject to safety testing. One panellist comments on the difficulties of determining general definitions for safety criteria and advocates a domain-by-domain approach to regulation: *"But I think it is going to be incredibly hard to produce those definitions (which is why I have advocated for a more dynamic, gradual regulatory system that can slowly increase the contexts of use, rather than giving blanket approvals)"*. This

[55] EESC Opinion, page 6.
[56] EGE Statement, page 18.
[57] EESC Opinion, page 6.

pattern is already being applied – for example regulatory work on self-driving cars[58] and autonomous weapons[59] is being undertaken at the present time. Another panellist comments on establishing a discipline of safety-critical AI, with implied benefits of education, training, quality control and regulation that come with a discipline: "*Yes, we need a new discipline of safety-critical AI - which brings good old fashioned safety and software engineering disciplines into AI*".

Assertion 14.1. The panel also unanimously agreed that achievable remedies should be agreed and in place for times when AI applications malfunction. There was discussion whether the remedies meant penalties, e.g. legislative tools aiming to discourage certain behaviour or to compensate an injured party, but penalties were seen as too narrow because there are other remedial actions that can be applied to undo or diminish the harm of AI applications' outcomes. The EGE Statement broadly concurs, making a distinction between legislative penalty and harm mitigation: "*In this regard, governments and international organisations ought to increase their efforts in clarifying with whom liabilities lie for damages caused by undesired behaviour of 'autonomous' systems. Moreover, effective harm mitigation systems should be in place*"[60]. Therefore, the assertion statement is concerned with any kind of remedial action, but the key point is that failure modes or harmful outcomes of an AI application need to be understood and actions defined to recover. Clearly the failure modes, their impacts and their associated remedial actions are domain-dependent, and one panellist points out that remedial actions may already exist in certain domains: "*I expect these remedies already exist in many cases*". Another points out that we need remedial actions for any critical infrastructure: "*We need remedies in place for the malfunction of any critical technology / infrastructure. Also AI*", and the expectation is that remedial processes will be in place for existing critical infrastructures. In many application domains, adaptation and augmentation of current practice is likely to be needed rather than completely new practices. A key question is therefore how to determine the domain for an AI application, and then to identify the relevant regulations and practices within it. In the example of self-driving cars, there are many failure modes already known associated with human-driven cars, as well as recovery actions (e.g. emergency procedures in response to a vehicle accident). These are for the most part still applicable to self-driving cars, and a major question concerns what additional harmful outcomes will be caused by the self-driving aspect of the vehicles and how they may be mitigated. A further point concerns the question of whether a malfunction can be identified because of deliberate obfuscation and information withholding by AI providers and platforms: "*The bigger problem is lack of transparency (especially with large social media companies) which makes it very difficult to seek redress*". This is naturally related to the transparency issue (discussed above), and also that transparency is a contributor to trust in an AI system, and the same conclusion applies here: AI providers should be encouraged (by whatever means appropriate including legislation with penalties) to ensure that their AI systems are as transparent as possible.

Assertion 10. The assertion "*Mechanisms that monitor and can constrain AI systems' behaviour are necessary, including stop or human override controls*" was strongly supported by the panel, with 9 panellists agreeing and 1 disagreeing. Here the key aspects are how an AI system can be constrained or stopped, and who or what is in overall control. Two of the external sources strongly advocate that humans should be in overall control of AI systems, and the third source makes no comment on the matter. The EGE Statement asserts that: "*All 'autonomous' technologies must, hence, honour the human ability to choose whether, when*

[58] See for example: https://medium.com/syncedreview/global-survey-of-autonomous-vehicle-regulations-6b8608f205f9. Retrieved 2018-06-06.
[59] See for example: https://thebulletin.org/military-applications-artificial-intelligence/why-world-needs-regulate-autonomous-weapons-and-soon. Retrieved 2018-06-16.
[60] EGE Statement, page 18.

and how to delegate decisions and actions to them"[61], while the EESC Opinion "[…] *calls for a human-in-command approach to AI, including the precondition that the development of AI be responsible, safe and useful, where machines remain machines and people retain control over these machines at all times"*[62]. The panellist who disagreed with the assertion statement pointed out that whilst overall human control is preferable, we need to also consider that human override may be counterproductive in a few situations: "*I think that this is context- and task-sensitive. Usually, we want a human in or on the loop. But there will be occasional contexts in which human overrides will actually make things worse*". The human-in-command approach also has implications for responsibility for an AI system, covered later.

Assertion 13.1. The panel strongly supported the assertion that research is needed to identify factors that determine who or what is liable in different cases of AI application, with 7 panellists for and 1 against. This assertion refers to legal liability – this is closely related to moral responsibility, which is a separate issue that is discussed later. There were no comments from the single panellist who disagreed with the assertion, so the following discussion concerns comments made by panellists who agreed with the assertion. A key point is that liability assignment is a well-established legal practice: "*Those factors might be determined by current legal practice, but that's fine. The main thing is that we need to know what we should pay attention to in these debates*". Also, the assignment of liability has already been investigated for similar situations: "*Assigning liability among the several actors, i.e. user, repairer, manufacturer etc, is complex for many technologies - and AI is no different*". The previous comment also highlights that in a given situation liability may be shared amongst actors, rather than having a single actor liable. Clearly, the assignment of liability depends on the application of the AI, e.g. the environment, stakeholders, the role of the AI and potential consequences, so a case-based approach may be necessary to evaluate each situation: "*For some AI applications, this may be relevant. But I guess the research then will be needed to address specifical applications (such as self-driving cars and autonomous behaviour to minimize damage during accidents)*". Finally, a panellist makes the point that there may be some situations where there is no liable actor: "*Not only who or what is liable, but who or what, if anything or anyone at all, is liable*". This is not confirmed, but we should not discount the possibility. In the external studies, the EESC Opinion emphatically states that humans are the liable parties in AI applications: "*[…] The EESC is opposed to any form of legal status for robots or AI (systems), as this entails an unacceptable risk of moral hazard. Liability law is based on a preventive, behaviour-correcting function, which may disappear as soon as the maker no longer bears the liability risk since this is transferred to the robot (or the AI system). There is also a risk of inappropriate use and abuse of this kind of legal status. The comparison with the limited liability of companies is misplaced, because in that case a natural person is always ultimately responsible"*[63].

Assertion 9. The panel broadly supported the assertion that interdisciplinary research is needed to determine how law can ensure responsible behaviour, with 8 panellists agreeing and 2 disagreeing. There were however some significant caveats in the comments. Amongst those who agreed, one panellist commented: "*I only mildly agree (partly because I think that the law is a relatively weak & inefficient means to "ensure responsible behavior")*". This indicates that there are other mechanisms that need to be investigated as well as law to encourage responsible behaviour (economic drivers for example). Amongst the comments by panellists who disagreed with the assertion was: "*Interdisciplinary research on legal aspects of AI applications clearly is needed. However, the goal of such research hardly should be to ensure responsible behaviour. AI is a tool, and may hence potentially be used for irresponsible and responsible purposes*". This refers to a specific school of thought that AI should be regarded as a tool and, like a

[61] EGE Statement, page 16.
[62] EESC Opinion, page 3.
[63] EESC Opinion, page 10.

knife, may be used for good or to cause harm: the thesis being that it is the use the tool is put to that needs scrutiny regarding responsible behaviour. A further comment indicated that there is likely to be a body of relevant work that already exists regarding responsible behaviour and should be consulted: "*I think that social science already has some very good ideas about this*". The external studies do not comment specifically on how law can ensure responsible behaviour, but the EGE Statement comments regarding allocation of responsibility: "*The whole range of legal challenges arising in the field should be addressed with timely investment in the development of robust solutions that provide a fair and clear allocation of responsibilities and efficient mechanisms of binding law*"[64].

Assertion 11. The panel also broadly supported the assertion that research into whether and how AI systems' decisions or actions can be rolled back is needed, with 8 panellists for and 2 against. This assertion clearly supports assertion 14.1, concerning remedial actions should a malfunction be detected, but the ability to undo the actions of an AI system is likely to be generally useful even when no malfunction occurs. One of the panellists disagreeing with the assertion makes the point that rollback of current AI technology is likely to be similar to other existing automated decisions: "*In the short / medium term, the rollback of AI-decisions will hardly be much different from the rollback of any other automated decision. In the long term, when approaching AI super intelligence, this may be relevant - but doing such research now would be premature. In particular as we will hardly be capable of understanding the relevant aspects of this research challenge*". A key factor is that what needs to be done to roll back an AI action strongly depends on the application domain, the action itself and its consequences, both direct and indirect. Some actions may be very difficult to undo, especially if there are indirect or unobserved consequences. For example, if a commercial AI system displays signs of discrimination, reputation damage for the company is likely to occur, and repairing the damage will require more than simply reversing the discriminatory decisions.

[64] EGE Statement, page 18.

Socioeconomic Impact

SOCIOECONOMIC IMPACT				
ID	**Assertion Statement**	**Disagree Votes**	**Agree Votes**	**Total Votes**
15	*Research into AI's impact on human workers is needed, including employment and deskilling of humans replaced by machines, as well as psychological consequences*	0	10	10
20.2	*Research is needed to extend how we understand the economic impacts of machines in society with a specific focus on AI, and how AI is different from other mechanisation*	0	7	7
17	*Public attitudes towards AI need to be understood, especially concerning public trust of AI*	1	9	10
19	*Research is needed into how users of AI can identify and guard against discriminatory effects of AI*	1	9	10
24.1	*Each AI and application needs to be assessed for benefits and harm. We must consider who benefits from AI and also possibly who may be harmed by the same AI application*	1	6	7
23.2	*AI research needs to concentrate on applications where it is known that AI can outperform humans*	6	1	7
18	*Research is needed into how AI integrates into networks of humans and machines, as well how machines interact with other machines*	1	8	10
16	*Research into the threats that future AI may pose to humankind is required, including where AI and human goals differ and where AI can undermine human values*	2	8	10
22	*Research is needed into how AI can be tested against societal values such as self-determination, autonomy, freedom, trust and privacy*	2	7	9

Assertion 15. The panel unanimously agreed that research into AI's impact on human workers is needed, including employment and deskilling of humans replaced by machines, as well as psychological consequences. This is unsurprising as the majority of external sources also have highlighted these aspects and there is a great deal of public fear about AI, automation and robots taking away peoples' livelihoods for example. The EC Approach concentrates on maintaining relevant skills: "*Europeans should have every opportunity to acquire the skills and knowledge they need and to master new technology*". This is important to keep the working population able to work with the current technology. The EESC Opinion concurs with the Approach regarding upskilling: "*The maintenance or acquisition of digital skills is necessary*

in order to give people the chance to adapt to the rapid developments in the field of AI. The European Commission is firmly committed to developing digital skills through its Digital Skills and Jobs Coalition"[65]. The EESC Opinion also provides a caveat that skills development needs to be supported across the board, not just in areas affected by AI systems: *"However, not everyone will be capable of or interested in coding or becoming a programmer. Policy and financial resources will therefore need to be directed at education and skills development in areas that will not be threatened by AI systems (i.e. tasks in which human interaction is vital, where human and machine cooperate or tasks we would like human beings to continue doing)"*[66]. There is another aspect to deskilling, and this is the ever-increasing take-up of technology that performs the work of previous generations of humans, resulting in loss of the knowledge or skills to perform the work and an increasing reliance on the technology (that may not be able to explain its actions). Whilst these risks are not limited to AI, it is recommended that they are recognised, and plans be put in place for their assessment. The EESC Opinion addresses loss of employment through AI: *"The EU, national governments and the social partners should jointly identify which job sectors will be affected by AI, to what extent and on what timescale, and should look for solutions in order to properly address the impact on employment, the nature of work, social systems and (in)equality. Investment should also be made in job market sectors where AI will have little or no impact"*[67]. The panellists' comments concentrate on highlighting that the effects of AI on the working population need not all be negative, and advocate a balanced approach considering both the negative and positive effects: *"This is an important research challenge. However, the research should not only concern negative implications (deskilling etc.) but also opportunities brought by AI (e.g. new human work opportunities opening up in consequence of new AI). Current research in this area typically is biased towards problems / challenges. I believe a more balanced approach is needed"* and *"But we also need to examine potential positive impacts & opportunities. That is, we need to look at the full picture"*.

Assertion 20.2. The panel also unanimously agreed that research is needed to extend how we understand the economic impacts of automation, by specifically focusing on AI and how AI is different to other forms of mechanisation. Not all panellists voted regarding this assertion (7 out of 8 panellists), but all that voted agreed. A key point here is that there have been many cases of disruptive technological breakthroughs throughout history[68] and there are historical records of how society adapted to their advent, but the key question is to understand how AI is different from these historical cases. Comments made by the panel highlight the need for investigation into new socioeconomic effects as a result of adoption of AI: *"There is currently substantial demand and interest in such research"* and *"I do think that AI is replacing a different kind of labor than previous "revolutions," and doing so in ways that are potentially different. Perhaps existing management & economic theory is sufficient, but I'm skeptical"*. The EGE Statement points to the need for new economic models of wealth distribution in which AI and autonomous technologies participate and fair and equal access to these technologies: *"We need a concerted global effort towards equal access to 'autonomous' technologies and fair distribution of benefits and equal opportunities across and within societies. This includes the formulating of new models of fair distribution and benefit sharing apt to respond to the economic transformations caused by automation, digitalisation and AI"*[69]. Specifically addressing the differences between AI and other mechanisation, the EESC Opinion

[65] EESC Opinion, page 9.
[66] EESC Opinion, page 9.
[67] EESC Opinion, page 4.
[68] Three examples spring to mind: the Gutenberg printing press, the threshing machine and the Internet. Each of these were revolutionary: the Gutenberg press resulted in mass information dissemination, the threshing machine mechanised grain harvests providing huge efficiency gains at the cost of employment, and the Internet accelerated mass information dissemination by orders of magnitude.
[69] EGE Statement, page 17.

cites an external source that distinguishes between the types of skills affected through different types of technology: *"Brynjolfsson and McAfee from MIT refer to the current technological developments (including AI) as the second machine age. However, there are two important differences: (i) the "old" machines predominantly replaced muscular power, while the new machines are replacing brainpower and cognitive skills, which affects not only low-skilled ("blue-collar") workers but also medium and highly skilled ("white-collar") workers and (ii) AI is a general purpose technology which affects virtually all sectors simultaneously"*[70]. Research will be needed to test this assertion that AI affects virtually all sectors simultaneously, and if so, how can it be managed.

Assertion 17. The panel strongly supported the assertion that public attitudes towards AI need to be understood, especially concerning public trust of AI, with 9 votes for and 1 against. This is to be expected, given the coverage AI has had in the media, with scare stories regarding AI taking away employment or "killer robots" waging war on the human race. The OED defines "trust" as "*Firm belief in the reliability, truth, or ability of someone or something*"[71]. Clearly evidence of previous reliable behaviour is a contributory factor towards building trustworthiness, as discussed in Assertion 7, and high-profile AI failures (such as accidents involving self-driving cars) detract from it. The other attributes referred to in Assertion 7, transparency of decision-making and comprehensibility of previous behaviour also contribute to trustworthiness – if people can see and understand behaviour, they are less likely to be suspicious of it. Attitudes and trust of the general public are most likely to be directed at the application of AI, not AI per se, as pointed out by a panellist supporting the assertion: "*Agree, but attitudes need to be investigated not for AI in general (too broad), but for key AI applications (e.g. trust in self-driving cars)*". Also, an application of AI is more likely to have societal impact rather than the underlying general-purpose algorithms, because the application is designed for real-world benefit and may have real-world threats. Another panellist who also supported the assertion pointed out the need to capture the full spectrum of diversity in public opinion: "*This kind of survey work could be helpful, but only if done with appropriate care to measure the relevant factors & covariates. I suspect that public attitudes will vary widely, and so one would need to capture that diversity*". A further support of the assertion observed that there is already information on public attitudes to AI and robotics: "*Yes, although we already have a pretty good understanding through i.e. the euBarometer surveys*". A recent relevant Eurobarometer is the 2017 survey "*Attitudes towards the impact of digitisation and automation on daily life*"[72] in which perceptions and attitudes towards robotics and AI were polled, and the overall attitude of the general public towards AI was mildly positive, with 51% "*positive*" and 10% "*very positive*" compared to 22% "*fairly negative*" and 8% "*very negative*", but an overwhelming majority (88%) agree that robots and AI require careful management, of which 35% "*tend to agree*" and 53% "*totally agree*".

Assertion 19. The panel also strongly supported that research is needed into how users of AI can identify and guard against discriminatory effects of AI, with 9 votes for and 1 against. Again, this is unsurprising because there is considerable concern regarding bias and discrimination in AI per se, and there is already work being undertaken to prevent AI systems being biased in the first place[73]. The need for

[70] EESC Opinion, page 8.
[71] https://en.oxforddictionaries.com/definition/trust. Retrieved 2018-06-18.
[72] Eurobarometer EBS 460 "Attitudes towards the impact of digitisation and automation on daily life". Available at: http://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/Survey/getSurveyDetail/instruments/SPECIAL/surveyKy/2160. Retrieved 2018-06-08.
[73] See for example The World Economic Forum Global Future Council on Human Rights 2016-2018: How to Prevent Discriminatory Outcomes in Machine Learning. Available at: http://www3.weforum.org/docs/WEF_40065_White_Paper_How_to_Prevent_Discriminatory_Outcomes_in_Machine_Learning.pdf. Retrieved 2018-06-08.

research into the prevention of bias in AI is widely supported, and the EGE Statement comments: *"Discriminatory biases in data sets used to train and run AI systems should be prevented or detected, reported and neutralised at the earliest stage possible"*[74] but here the assertion focuses on the need to understand how users (e.g. citizens) can be empowered to recognise discrimination. Comments from panellists that supported the assertion concern the need to define the discriminatory effects: *"though clarity is needed about what "discriminatory effects" are"*, and the need to protect citizens who do not use AI but are affected by it: *"Of course, people should be able to do this. But I actually think that the more important challenge is helping the people who are differentially impacted by AI, but are not directly using it (so have few opportunities to learn about this system that is deeply affecting their lives)"*. This last point is particularly important, because it affects potentially many people, who have no idea that they are being discriminated against.

Assertion 24.1. There was strong support for the assertion: *"Each AI and application needs to be assessed for benefits and harm. We must consider who benefits from AI and also possibly who may be harmed by the same AI application"*, with 6 supporters and 1 dissenter. This assertion has its roots in a discussion in Round 2 stemming from the Asilomar Beneficial AI principles[33] in which a panellist asked who benefits and pointed out that what benefits one party may negatively affect, discriminate or harm another (relating also to Assertion 19). Clearly there is the general "societal benefit" and the EESC Opinion supports positive societal benefit: *"The development of AI applications that benefit society, promote inclusiveness and improve people's lives should be actively supported and promoted, both publicly and privately. Under its programmes, the European Commission should fund research into the societal impact of AI and of EU-funded AI innovations"*[75]. There are also the partisan benefits that are the core of this assertion, and a specific example of the need to protect from partisan benefit resulting from AI development is given also by the EESC Opinion: *"The vast majority of the development of AI and all its associated elements (development platforms, data, knowledge and expertise) is in the hands of the "big five" technology companies (Amazon, Facebook, Apple, Google and Microsoft). Although these companies are supportive of the open development of AI and some of them make their AI development platforms available open-source, this does not guarantee the full accessibility of AI systems. The EU, international policy makers and civil society organisations have an important role to play here in ensuring that AI systems are accessible to all, but also that they are developed in an open environment"*[76]. A panellist supporting the assertion casts the assessment in terms of ethical risks and cites the Responsible Research and innovation (RRI) frameworks[77] as a potential home for assessment practice and guidelines: *"Yes, all AIs should be subject to an 'ethical risk assessment', as well as being developed within frameworks of Responsible Research and Innovation"*. Another supporting panellist points out that this assessment of benefits is not exclusive to AI: *"This goes for AI, and also for other forms of technology. Practically any technological progress comes at a cost, and we need to make sure that the benefits outweighs [sic] the costs"*. Clearly this is true, but are there any benefits specific to AI that need to be investigated? Finally, a further supporting panellist makes a plea for the use of common sense in the application of assessments: *"Agreed, but obviously there are sensible & absurd ways to approach this. For example, changing a color on a robotic arm shouldn't trigger a re-assessment"*.

[74] EGE Statement, page 17.
[75] EESC Opinion, page 4.
[76] EESC Opinion, page 9.
[77] See for example https://ec.europa.eu/programmes/horizon2020/en/h2020-section/responsible-research-innovation. Retrieved 2018-06-08.

Assertion 23.2. The next assertion, "*AI research needs to concentrate on applications where it is known that AI can outperform humans*" is unique in this consultation in that there was strong consensus amongst the panellists, but the consensus ***disagreed*** with the assertion: 1 agreed and 6 disagreed. There were numerous reasons given for disagreement in the comments. One commented that we should not artificially restrict the domain of AI research: "*There are lots of reasons to do AI research, and I see no reason why we should limit its domain or applications in this way*". Others provided reasons why it can be useful to research AI when it equals or underperforms humans: "*It should concentrate on cases where AI might replace humans (whether or nor [sic - assume "not"] humans are outperformed)*" and "*AI can be still be helpful even if it underperforms when compared to human behavior in similar circumstances. For instance, simple eldercare robots that help people remain at home*". A further comment warned about reduction of expectations that may be caused by artificial limitation of research targets: "*It is inevitable that AI developers seek low hanging fruit, although I don't think this is necessarily a good thing, since harder problems then get neglected.*"

Assertion 18. The panel broadly agreed that research is needed into how AI integrates into networks of humans and machines, as well how machines interact with other machines, with 8 votes for, 1 vote against, and 1 panellist not voting. The sole comment supporting the assertion indicated that research is already underway: "*And there's already a large amount of this research*", and the sole comment against felt that the recommendation was too broad: "*This sounds too general to me*". The EESC Opinion discusses complementary human-AI systems: "*The EESC recommends that these stakeholders work together on complementary AI systems and their co-creation in the workplace, such as human-machine teams, where AI complements and improves the human being's performance. The stakeholders should also invest in formal and informal learning, education and training for all in order to enable people to work with AI but also to develop the skills that AI will not or should not acquire*"[78].

Assertion 16. The panel broadly supported the assertion that research into the threats that future AI may pose to humankind is required, including where AI and human goals differ and where AI can undermine human values, with 8 panellists agreeing with the assertion and 2 participants disagreeing. This assertion alludes to longer-term assessment of future threats associated with Artificial General Intelligence and Superintelligence. These technologies may be able to determine their own goals, and they may not necessarily be compatible with human goals. This was indicated by two comments from panellists that supported the assertion: "*I would hasten to add that, given that such a threat is theoretical/long-term, this research itself should be performed in a responsible manner that does not unduly alarm the public and undermine beneficial AI progress*", and "*While I agree that these are interesting questions, I also think that they focus on a relatively far-off future, and so I don't think that they should be the focus of major funding efforts at this time*". One of the comments from panellists that disagreed with the assertion argues that it is too soon to investigate future AI: "*I believe we do not know enough about what such future AI will be like for this to be a meaningful research topic*". The other disagreeing comment pointed out that risks to "humankind" are too general – the risks to the specific affected parties need to be assessed. It also points out that the affected parties may not be human, but any other sentient creature: "*Too general - not humankind, but affected humans and other sentient creatures*". This last point backs up another assertion (29.2) that we need to consider the impact of AI on non-human entities such as animals or the environment.

Assertion 22. The panel broadly supported the assertion that research is needed into how AI can be tested against societal values such as self-determination, autonomy, freedom, trust and privacy, with 7 votes for and 2 against. One comment from a supporting panellist added the caveat that the societal

[78] EESC Opinion, page 4.

values are not static: *"but in relation to conceptual progress that is being made in relation to self-determination, autonomy, freedom etc. these are not static concepts either"*, while the other comment from a supporting panellist pointed out the similarity between this assertion and others regarding the societal impact of AI: *"Much the same as an earlier point"*. The only comment from a panellist disagreeing with the assertion alluded to the need to understand how to measure and describe the societal values but is not convinced that AI needs to be tested against them: *"it seems like this is just a request for operationalizations of those terms? I'm not sure what else might be required?"*. The external sources are definite in their acknowledgement that the impact of AI on society needs to be better understood, but they do not go as far as advocating how AI can be tested against societal values. The EGE Statement says: *"The principles of human dignity and autonomy centrally involve the human right to self-determination through the means of democracy. Of key importance to our democratic political systems are value pluralism, diversity and accommodation of a variety of conceptions of the good life of citizens. They must not be jeopardised, subverted or equalised by new technologies that inhibit or influence political decision making and infringe on the freedom of expression and the right to receive and impart information without interference. Digital technologies should rather be used to harness collective intelligence and support and improve the civic processes on which our democratic societies depend"*[79]. The EESC Opinion asks: *"How do we ensure that our fundamental norms, values and human rights remain respected and safeguarded?"*[80] and advocates that: *"Under its programmes, the European Commission should fund research into the societal impact of AI and of EU-funded AI innovations"*[81].

[79] EGE Statement, page 18.
[80] EESC Opinion, page 6.
[81] EESC Opinion, page 4.

Design

DESIGN				
ID	Assertion Statement	Disagree Votes	Agree Votes	Total Votes
25	*Ethical principles need to be embedded into AI development*	0	10	10
27	*Inclusive, interdisciplinary teams are needed to develop AI*	0	10	10
29.2	*Sometimes AI will have to be sensitive to non-human entities, e.g. in agriculture, fishing etc.*	0	8	8
26	*AI engineers need to be aware of potential biases and prejudices in selection of training data*	1	9	10
29.1	*An important AI design consideration is how the AI advances human interests & values*	1	7	8
28.1	*Attempting formal definition of AI concepts may mask important nuances, be time-consuming and as a result may hold up AI research. It is more important to get adequate definitions and debate their shared understanding publicly.*	1	6	7

Assertion 25. The panel unanimously agreed that ethical principles need to be embedded into AI development. This is unsurprising given the importance given to ethics in AI. One panellist made the distinction between embedding ethical principles into the AI itself and respecting ethical principles at design time: *"Of course, this does *not* mean that ethical principles need to be explicitly represented by the AI. Rather, the idea is that ethical AI requires changes in practice (most notably, attention to ethical issues)"*. Therefore, this assertion may be interpreted that ethical principles need to be considered at design time. This echoes Assertion 0, where a similar point was made by the panellists. Another panellist pointed out that it may be difficult to determine the ethical principles: *"As long as we have a clear idea of what the relevant ethical ideas are"*. The external studies strongly support ethically-guided AI development. The EESC Opinion *"calls for a code of ethics for the development, application and use of AI so that throughout their entire operational process AI systems remain compatible with the principles of human dignity, integrity, freedom, privacy and cultural and gender diversity, as well as with fundamental human rights"*[82]. The EGE Statement states: *"Applications of AI and robotics should not pose unacceptable risks of harm to human beings, and not compromise human freedom and autonomy by illegitimately and surreptitiously reducing options for and knowledge of citizens. They should be geared instead in their development and use towards augmenting access to knowledge and access to opportunities for individuals. Research, design and development of AI, robotics and 'autonomous' systems should be guided by an authentic concern for research ethics, social accountability of developers, and global academic cooperation to protect fundamental rights and values and aim at designing technologies that support these, and not detract from them"*[83].

[82] EESC Opinion, page 3.
[83] EGE Statement, page 17.

Assertion 27. The panel unanimously supported that inclusive, interdisciplinary teams are needed to develop AI. Comments also indicated the support: *"And if there were an option above "Strongly Agree," then I would have chosen it. The best AI - socially, but also technologically - emerges when one can use a broad, "design thinking" approach that employs methods from many disciplines"*, and *"I believe history has shown that AI can be developed also in non-interdisciplinary teams. However, future AI applications will likely be strengthened through an interdisciplinary approach"*. This is an important point – while it is and has been possible to develop AI from a purely technical perspective alone, in order to fully realise its benefits and protect from potential threats, interdisciplinary teams are needed. A further comment emphasised that diversity in the disciplines is needed: *"Diverse in the sense of interdisciplinary"*, and another pointed out that understanding the target communities is important: *"it is critically important that design teams fully reflect the gender and ethnic mix of the societies they are aiming to develop AIs for"*. Interdisciplinary development is widely supported in the wider EC community. The EESC Opinion has stated that one of its primary objectives is to: *"… shape, focus and promote the public debate on AI in the coming period, involving all relevant stakeholders: policy-makers, industry, the social partners, consumers, NGOs, educational and care institutions, and experts and academics from various disciplines (including AI, safety, ethics, economics, occupational science, law, behavioural science, psychology and philosophy)"*[84]. The EC Approach has taken steps to provide support for collaboration across member states through centres of excellence and Digital Innovation Hubs: *"The Commission will support fundamental research, and also help bring more innovations to the market through the European Innovation Council pilot. Additionally, the Commission will support Member States' efforts to jointly establish AI research excellence centres across Europe. The goal is to encourage networking and collaboration between the centres, including the exchange of researchers and joint research projects"* and *"Digital Innovation Hubs are local ecosystems that help companies in their vicinity (especially small and medium-sized enterprises) to take advantage of digital opportunities. They offer expertise on technologies, testing, skills, business models, finance, market intelligence and networking"*[85].

Assertion 29.2. The panel also unanimously supported the assertion that sometimes AI will have to be sensitive to non-human entities, e.g. in agriculture, fishing etc. This is often overlooked because there is much emphasis on ethical considerations related to human rights, but AI needs to respect other domains such as those quoted above, and this cannot be forgotten. A comment emphasises this point and provides further examples that need to be considered: *"Or self-driving cars needing recognize [sic – assume "needing to recognize"] animals in the road. Or home healthcare robots needing to recognize insect infestations. Or lots of other examples. I don't see how almost any AI could succeed if it wasn't sensitive to non-humans"*. The EGE Statement supports this assertion: *"AI technology must be in line with the human responsibility to ensure the basic preconditions for life on our planet, continued prospering for mankind and preservation of a good environment for future generations. Strategies to prevent future technologies from detrimentally affecting human life and nature are to be based on policies that ensure the priority of environmental protection and sustainability"*[86].

Assertion 26. The panel strongly supported the assertion that AI engineers need to be aware of potential biases and prejudices in selection of training data, with 9 votes for and 1 against. This is unsurprising as it is strongly supported in the wider community and the sentiment of the assertion concurs with that of Assertion 19. The EGE Statement says: *"Discriminatory biases in data sets used to train and run AI systems*

[84] EESC Opinion, page 3.
[85] EC Approach, page 1.
[86] EGE Statement, page 19.

should be prevented or detected, reported and neutralised at the earliest stage possible"[87] and the EESC Opinion says: *"[...] There is a general tendency to believe that data is by definition objective; however, this is a misconception. Data is easy to manipulate, may be biased, may reflect cultural, gender and other prejudices and preferences and may contain errors"*[88]. The only comment from the panel was from the panellist voting against the assertion, and the comment reveals that it is more likely to be the wording of the assertion that is objected to rather than its sentiment of preventing bias entering AI via its training: "*Everyone should know about the possibility, but only selected members of the engineering team need to be able to fully analyze the problems*".

Assertion 29.1. The panel broadly supported the assertion that an important AI design consideration is how the AI advances human interests & values, with 7 votes for and 1 against. The external studies concur that AI should benefit society by advancing its values. The EESC Opinion says: *"The development of AI applications that benefit society, promote inclusiveness and improve people's lives should be actively supported and promoted, both publicly and privately. Under its programmes, the European Commission should fund research into the societal impact of AI and of EU-funded AI innovations"*[89], but here we need to understand how an AI application will affect society. Clearly human values are important, but a panellist supporting the assertion raised a question regarding which humans: *"human = all of humanity? or only some humans?"* Another panellist who voted against the assertion asked questions regarding which values, pointing out that not all values are universally beneficial: *"The problem with this statement is that it assumes merit and consistency to "human interests and values". Not all human interests are worthy of consideration, not all human values are laudable, and, clearly, such interests and values do not stand as a consistent set of principles. A value judgement is implicit in the statement - admirable interests and good values need only apply. Too [sic] determine these, value judgements will need to be made*". Given these comments, we need to understand the effects on different sectors of society – who will benefit and who may suffer and how. In addition, this assertion should also be considered in the light of assertion 29.2, which brings in other domains and entities that need to be considered.

Assertion 28.1. The panel broadly agreed that attempting formal definition of AI concepts may mask important nuances, be time-consuming and as a result may hold up AI research, with the knock-on effect of delaying development of novel AI applications, with 6 votes for and 1 against. The assertion also states that it is more important to get adequate definitions and debate their shared understanding publicly. While some definitions are helpful, the community should not be held up while formal definitions are agreed. One panellist who voted for the assertion commented that not all definitions are equal and some need to be formalised, but others need not be: "*Some aspects of AI (e.g. different machine learning approaches) clearly require formal definitions. However, for high-level conceptualizations of AI adequate definitions that are publically [sic] debated will be sufficient*". The panellist who voted against the assertion pointed out that since we have a poor understanding of natural intelligence, definitions of artificial intelligence are extremely difficult: "*No. One of the major problems with AI is that we have a very poor understanding of natural intelligence. This makes AI a (more or less) theory free science. What we need is a general theory/standard model of intelligence (cf physics)*."

[87] EGE Statement, page 17.
[88] EESC Opinion, page 6.
[89] EESC Opinion, page 4.

Responsibility

RESPONSIBILITY				
ID	Assertion Statement	Disagree Votes	Agree Votes	Total Votes
32.1	*Research is needed to determine how/when moral responsibility should translate into legal liability, specifically applied to AI situations*	0	8	8
30	*People, not AI systems, bear responsibility and AI developers are responsible for the tools they develop*	2	8	10
31	*The concept of "AI responsibility" needs to be researched by integrated multidisciplinary teams so as to arrive at a hybrid understanding of the key issues concerning responsibility and where it can be attributed when AI participates in human-machine networks*	2	8	10

Assertion 32.1. The panel unanimously agreed that research is needed to determine how/when moral responsibility should translate into legal liability, specifically applied to AI situations. Comments from the panellists alluded to the potential difficulties: "*Yes, although I think this is a very difficult question - one for the lawyers and philosophers*", and "*We need the research, but good luck getting any agreement!*". It is likely that each AI application and situation will need to be judged on its own merits, as is currently happening – there are domains of application that are being currently investigated and tested as to how these questions can be answered within the specific domain (the obvious examples that spring to mind are self-driving cars and automated weapons). The EGE Statement concurs that this work is important: "*In this regard, governments and international organisations ought to increase their efforts in clarifying with whom liabilities lie for damages caused by undesired behaviour of 'autonomous' systems. Moreover, effective harm mitigation systems should be in place*"[90].

Assertion 30. The panel broadly agreed that people, not AI systems, bear responsibility and AI developers are responsible for the tools they develop, with 8 votes for and 2 against. The assertion contains two clauses that will be discussed separately because it has become clear through this analysis they are actually independent assertions: that people bear ultimate responsibility for AI systems' actions; and that a designer bears responsibility for the systems they develop.

There is strong support in the panel and the wider community for the assertion that people bear ultimate responsibility for AI actions. The panellists that support the assertion commented: "*Strongly agree that People not AI systems bear responsibility …*" and "*Strongly agree - see EPSRC Principles of Robots 'Humans, not robots, are responsible agents'*". The EPSRC Principles of Robots advocates that robots are tools and the human user determines the use the tool is put to, which can be beneficial or harmful, but the human

[90] EGE Statement, page 18.

bears final responsibility for the tool's usage[91]. The EGE Statement has much to say on the matter and comes down firmly in agreement that humans need to be responsible: *"Moral responsibility, in whatever sense, cannot be allocated or shifted to 'autonomous' technology"*[92] and *"The principle of Meaningful Human Control (MHC) was first suggested for constraining the development and utilisation of future weapon systems. This means that humans - and not computers and their algorithms - should ultimately remain in control, and thus be morally responsible"*[93]. The EESC concurs with the MHC approach: *"The EESC calls for a human-in-command approach to AI, including the precondition that the development of AI be responsible, safe and useful, where machines remain machines and people retain control over these machines at all times"*[94]. There have been discussions regarding legal personhood for AI systems, i.e. that AI systems could take responsibility for their own actions. The current weight of opinion is against this, and the EESC is emphatic in their rejection: *"[...] The EESC opposes the introduction of a form of legal personality for robots or AI. This would hollow out the preventive remedial effect of liability law; a risk of moral hazard arises in both the development and use of AI and it creates opportunities for abuse"*[95]. Much of the discussion regarding legal personhood for AI is looking ahead to Superintelligence or Artificial General Intelligence where the AI systems might have vested self-preservation and self-improvement goals and thus would have an incentive to behave according to whatever rights and responsibilities society places upon them, but the current generation of AI systems fit far better into the category of smart tools that need a human in charge to determine the tools' application and take responsibility for the outcome.

Regarding the second assertion, if we accept that a human being needs to take responsibility for an AI system, then we need to understand which human (or humans), and under what circumstances? The assertion that the designer needs to take responsibility for the tools they develop is certainly true, but only up to a point – there are many contexts of use that the designer cannot be responsible for. Most of the disagreement in this assertion concerned the question about who is responsible, not whether a person should be responsible. A panellist who agreed with the assertion commented: *"… Strongly agree that People not AI systems bear responsibility. However, AI developers while responsible for the quality of the tools they develop cannot be held responsible for how the tools are used"*. Another panellist, who voted against the assertion, commented: *"I agree with the quoted statements, but I think the summary over-simplifies a complex issue. AI developers certainly bear significant responsibility for the tools they develop and this must inform their practice. However, precisely because AI systems may act - and interact - in ways that individual designers or teams could not have predicted, assigning responsibility can be difficult. We, as a society, need to plan for problems that may arise without any one individual being at fault"*. Another panellist, who also disagreed, commented: *"It all depends on contextual factors"*. Clearly there are responsibilities that need to be assigned to an AI system's designer and these include reliability, fitness for purpose, basic safety etc. However, the designer cannot be responsible when the system is used beyond its original purpose or for deliberate or accidental misuse, and it is an open question whether existing regulation that puts the onus of responsibility on the user is adequate. The EC Approach quotes plans to extend existing directives to incorporate AI: *"The EU has liability rules for defective products. The Product Liability Directive dates from 1985 and strikes a careful balance between protecting consumers and encouraging businesses to market innovative products. The Directive covers a broad range of products and possible scenarios. In principle, if AI is integrated into a product and a defect can be proven in a product*

[91] 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. 2017. Principles of robotics: regulating robots in the real world. *Connection Science*, *29*(2), pp.124-129.
[92] EGE Statement, page 10.
[93] EGE Statement, page 10.
[94] EESC Opinion, page 3.
[95] EESC Opinion, page 4.

that caused material damage to a person, the producer will be liable to pay compensation. The actual cause of events that lead to damage or incident is decisive for the attribution of liability. The Commission plans to issue an interpretative guidance clarifying concepts of the Directive in view of the new technologies, building on a first assessment on liability for emerging digital technologies published today"[96]. The assignment of responsibility is very likely to be situation-dependent, and a promising strategy is to use case-based precedents, similar to existing case law.

Given this discussion, the key recommendations arising from the two assertions are made separately. Whilst it is currently well accepted that people need to be in control of, and take responsibility for, AI systems' actions, there may be future situations where AI systems have their own legal personhood, but this is not in the foreseeable future. Research is clearly needed to determine who is responsible for AI systems' actions in different circumstances, domains and application situations (all of which may mean a different person is responsible).

Assertion 31. The panel broadly agreed that the concept of "AI responsibility" needs to be researched by integrated multidisciplinary teams so as to arrive at a hybrid understanding of the key issues concerning responsibility and where it can be attributed when AI participates in human-machine networks with 8 panellists agreeing and 2 disagreeing. This assertion follows from Assertion 30 and adds the recommendation that the question of responsibility should be investigated by multidisciplinary teams. The only comment (by a supporting panellist) reinforced the "humans are responsible" principle discussed above: "*Yes, but only in order to attribute responsibility among the human designers - not to the AI*".

Conclusion

This document has summarised the results of a consultation with multidisciplinary experts into the subject of Responsible Artificial Intelligence and compared these results with other European-focused studies resulting in guidance from similar fields.

This study has used the Delphi Method, an iterative consultation mechanism aimed at consensus building (or highlighting difference where consensus is not achieved). Three rounds of iteration were undertaken and a total of eight experts participated in all three rounds. The result of the consultation was 33 assertion statements that reached consensus amongst the experts, in six broad themes:

- Ethics;
- Transparency;
- Regulation & Control;
- Socioeconomic Impact;
- Design; and
- Responsibility.

The assertions are summarised as key recommendations in the Summary of Recommendations. Each assertion has been discussed and compared with three external studies, highlighting where there are similarities and differences. Overall the consensus between the four studies is good, where multiple studies concur on the major points and recommendations, however each study has its own perspective and has different emphasis on detail points. The major points from this consultation are discussed next, and these are cross-cutting issues or principles that affect and join different themes in this consultation.

[96] EC Approach, page 3.

This consultation advocates that, for the foreseeable future, humans need to be in overall control of AI because the consensus regarding the current state of AI technology is that of smart tools, and humans must take responsibility for AI actions. Humans must be empowered to monitor and intervene to prevent or undo AI actions if necessary. There may be future situations where the predictions of AI as a Superintelligence come true, and this will necessitate revisiting the question of whether a human or the AI itself is responsible, but for the current time the consensus is that the human is responsible. The question of which humans are responsible most likely depends on the application context of the AI, as different application contexts may have different human roles and responsibilities.

This consultation asserts that application contexts are key influencers of many aspects of "Responsible AI", more so than the underlying AI algorithms because the application context determines the societal impact, and whether it is for good or poses risks. Different application contexts may use the same underlying AI algorithms, but the contexts may have totally different risks, stakeholders, ethical considerations and regulation requirements. This correlates with the "AI is a tool" school of thought that says that the use the AI is put to is the subject of ethical concern, regulation and responsibility; rather than the AI algorithm itself. Existing application contexts may have their own regulations and control patterns already, and these can for the basis for AI systems participating in the context. (A key example here is AI-powered self-driving vehicles. There are many regulations and practices for human-driven vehicles, so the question is what need to be changed or added to cater for self-driving vehicles.)

AI has significant potential for disruptive socioeconomic impact. Lessons may be learned from previous examples of disruptive technologies and analogies may be drawn between AI and historical examples of disruptive mechanisation, but an open question remains regarding what sets AI apart from previous examples of technological disruption.

AI needs to be trustworthy to be generally socially acceptable. Some key aspects that influence trustworthiness are transparency, comprehensibility (to a layperson) and a track record of reliability. Some AI technologies are opaque and unpredictable, and these may be useful for research and surprising behaviour may even be inspirational, but these attributes do not contribute to trustworthiness, especially in AI systems that operate in safety-critical applications.