The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems

Ethically Aligned Design (EAD) RFI Feedback response

The feedback in this document was submitted as part of an open Request for Information (RFI) process regarding the document created by *The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems* ("The IEEE Global Initiative") titled, *Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems.*

As stated in the submission guidelines for our RFI process, all contributions have been posted exactly as they were received. The only modification to submissions was to standardize the font and spacing in the following document for ease of readability. Committees working to update Version 2 of *Ethically Aligned Design* are currently in the process of reviewing all feedback received to help inform their updated section drafts.

The Executive Committee and all members of The IEEE Global Initiative wish to formally thank all contributors for their RFI submissions. You have contributed to the transparent, open and consensus building process that is a core part of our ethos while also helping us fulfill our mission to "ensure every technologist is educated, trained, and empowered to prioritize ethical considerations in the design and development of autonomous and intelligent systems."

Thank You.

Table of Contents for Feedback received to date (click to go right to sections)

- Tamara Hernández Alvarado, Biomedical Engineering student at Universidad Autónoma de Querétaro
- Daniel Alejandro Morales Hernández, Universidad Autónoma de Querétaro
- Sandra Daniela Carmona Martínez, Biomedical Engineering Student at Universidad Autónoma de Querétaro, México
- Pavel M. Gotovtsev, PhD, Vice-head of biotechnology and bioenergy department, National Research Centre "Kurchatov Institute"
- Aldo Aaron Gonzalez Ramirez, Autonomous University of Queretaro
- Dr. Christopher A. Tucker, Cartheur Robotics, spol. s r.o., Prague, Czech Republic
- José Eduardo Quintanar Pozos, The Autonomous University of Queretaro
- AI for Social Good, Waseda University, Tokyo, Japan - 7 March, 2017, Autonomous Weapon Systems Group
- LIU Zhanxiong（刘战雄) PhD Candidate on Philosophy of Technology in School of Humanities，Southeast University，China
- Jim Isaak, IEEE Senior Member, Computer Society President Emeritus, and past VP of the Society on Social Implications of Technology
- Alexis J. Valentin, The Secretary, WhyFuture AI Concepts
- Charles H. Jones, PhD. C.H. Jones Consulting, LLC
- Jia He, IEEE Global Initiative China Committee member
- Ansgar Koene, Senior Research Fellow at the Horizon Digital Economy Research institute, University of Nottingham, UK
- Eileen Donahoe, J.D., Ph.D. (Ethics) Executive Director, Global Digital Policy Incubator, Stanford University Center for Democracy Development and the Rule of Law
- Viola Schiaffonati, Ph.D. Associate Professor of Logic and Philosophy of Science, Artificial Intelligence and Robotics Lab, Politecnico di Milano

- Alexandre Sacco Xavier, Master of Science Researcher in Information Systems at UFRGS (Federal University of Rio Grande do Sul, Brazil)
- Frederike Kaltheuner, Policy Officer, Privacy International and Asaf Lubin, JSD Candidate, Yale Law School, Robert L. Bernstein International Human Rights Fellow, Privacy International
- Renato Opice Blum, Coordenador do Curso de Direito Digital do INSPER
- Joachim Iden, TUV Rheinland Japan
- Ilse Verdiesen MSc., Officer in the Royal Netherlands Arm, Master student TUDelft
- Pradyot Sahu, Senior Member, IEEE, Director, 3innovate
- Christina Demetriades / Deputy General Counsel, Sales & Delivery, Accenture
- Thomas Dandres, Ph.D. Research Officer / Agent de Recherche, CIRAIG, Polytechnique Montréal, dép. génie chimique
- David G. Hunt, WhyFuture AI Concepts, and Alexis J. Valentin, The Secretary, www.whyfuture.com
- Kurt Thomas, Bonn, Germany
- Ariella Berger, www.unboundedresearch.co
- The IEEE Global AI Ethics Japan Committee - Workshop Responses to EADv1 with organizers: Arisa Ema / Katsue Nagakura

I find the report very solid, encompassing, and taking us a long way towards where we need to go if we are to achieve ethically aligned AIS systems. I suggest that it be made clearer that such systems will have to use AI to oversee AI—and that this requirement calls for developing a whole new slew of AI programs. (Oren and I called them AI Guardians.) The main reason a second kind or layer of AI program is needed, is because the first layer (the AI program that guides the function of the various operating systems, whether they are cars or weapons) cannot examined by human beings without AI aid. Take, for example, the question of who is liable if an autonomous car crashes into another car—the program, owner, or the car?—which requires "reading" and "analyzing" (human terms) the operating AI program. It cannot be held up to the light, like a $20 bill, and examined. Hence, the need for AI second order programs (or Guardians).

One further notes that in the offline world, we have two or more layers—one of the operating system and one of various layers of oversight. Workers have supervisors, teachers have principals, businesses are audited, etc. AI currently is largely on the first, operating kind. The development AIS called for developing AI accountants, overseers, and in some cases even AI regulators.

Amitai Etzioni
University Professor
The George Washington University

Dear Sir or Madam,

I hereby send you the feedback of Prof. Crowcroft and myself on the EAD document.

We want to congratulate you on the excellent work and look forward to reading the next iteration.

I have pasted our comments below, and added them in Pdf format. Please let me know if you have any questions about our comments.

IEEE's Global Initiative for Ethical Considerations in the Design of Artificial Intelligence and Autonomous systems

Written comments[1] by **Corinne Cath[2] and Jon Crowcroft[3]**

Over the past months we have been closely following the IEEE's Global Initiative for Ethical Considerations in the Design of Artificial Intelligence and Autonomous systems. We believe it adds an important perspective to the debate about the ethics of AI, and has done a great job at bringing together some of the most prolific AI/AS thinkers and writers. What follows are several observations and comments on the first version of the Ethically Aligned Document (EAD), that we hope will provide a positive contribution to the ongoing work.

The Ethically Aligned Document (EAD) uses the terms AI & AS but does not specify what is meant by these terms. This leads to a situation in which the various committees are discussing different issues, all under the AI/AS umbrella. We suggest the EAD includes a comprehensive definition of what they consider AI and AS to be, and what not.

An interesting definition is that of on Russell and Norvig (1995). They mention that the history of artificial intelligence has not produced a clear definition of AI but rather is variously emphasizing four possible goals: "systems that think like humans, systems that act like humans, systems that think rationally, systems that act rationally." Stuart J. Russell and Peter Norvig, Artificial Intelligence: A Modern Approach, Englewood Cliffs, NJ: Prentice Hall, 1995: 27.

We believe that there is a need to ground the concept of AI/AS. Without defining what we are talking about when we talk AI/AS there is too much room for misunderstanding.

Similarly, there is a tension in the IEEE document between incorporating "ethics" or "ethical values" and "end-user values". Some of the committees focus on the need to incorporate high level human values like human rights, whereas others stress the need for including the values of end-users. It is important to differentiate between ethical and end-user values; these might overlap in certain cases but certainly not in all. The IEEE should clarify that it cannot incorporate end-user values when these cannot be considered ethical as per the standard the IEEE decides to set for defining ethical. We suggest that setting this standard should be done by looking at existing human rights standards, or other legal concepts like for instance human dignity.

There were also several topics we expected would be discussed but were not. We would like to see the next version of the EAD include a discussion of the following issues:

- What are the limits of autonomy? When, if ever, can we hold AI or robots responsible or liable for their actions? And what is the place of the debate about robot rights in this document?
- How should we prepare for unanticipated AI developments, like for instance collective behaviour of AI?
- What are the ethical issues related to working for AI? This already happens, for instance in the case of people working for Mechanical Turk. And in many cases, working for AI limits the ability of workers to unionize, provide feedback or address other forms of work related issues.
- How should we respond to AI developing "new" ethical principles based on the input provided by technologists?
- The introduction of AI is and will continue to bring shocks to the social, political, and economic fabric of society. What are the ethics of introducing technology like this? The document addresses ethical issues assuming technology will be introduced, but should it also discuss the limits of when a technology should not be introduced?

Comments on the Summary:

P. 9 Committee 7 (Economics/Humanitarian Issues), under its issues it says 'any AI policy might slow innovation'. Even though this statement is further nuanced in the committee's section, we suggest updating this statement to more accurately reflect the content of the committee's candidate recommendations (which does in effect not argue that policy slows down innovation).

P. 18 - Principle 2 - Recommendations: From the current language, it is unclear what the intended use of a 'system of registration' is.

P. 18 - Principle 3. - Transparency
We think it is important for the committee to also read these recent articles:

Boyd, Danah. 2017. The false hope of algorithmic transparency. https://points.datasociety.net/transparency-accountability-3c04e4804504 - .xy1sqnepl

Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. 2017. "Accountable Algorithms." University of Pennsylvania Law Review 165. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2765268.

These articles focus on the difficulties surrounding algorithmic accountability through transparency, including them would greatly add to the discussion as presented by the committee

2. Embedding Values Into Autonomous Intelligent Systems Committee

P. 25: In the last paragraph ("We also … users values.") it is suggested that interactive machine learning (AML) approaches can be used to ensure a system remains up-to-date with what its context needs from it. If this process happens within the machine independently, it is sure to raise many questions surrounding responsibility and liability. The committee should recognize, and mention these issues, or perhaps reach out to the law committee to see to what extent they have covered these issues.

P. 29: The second paragraph says: 'Computers and robots already instantiate values in their choices and actions, but these values are programmed or designed by the engineers that build the systems.' Sometimes these values are programmed into technology, but not always purposefully. And sometimes biases come from the data fed into a program. Cathy Kleiman gives an excellent example[4] of how feeding biased data into a machine learning algorithm can reaffirm the status quo. Perhaps the committee can update their language to reflect that not all values enter computers or robots purposefully.

p. 32 – 34: There is overlap between the suggestions made on transparency in the General Principles Committee and the recommendations made on this page. We think it would be interesting to do "a diff" between the two texts, and reduce the overlap.

P. 34. In relation to the GPDR's "right to be forgotten", the committee might also be interested in reading this recent article on why the 'right to explanation' does not exist in the GPDR:

Wachter, Sandra and Mittelstadt, Brent and Floridi, Luciano, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation (December 28, 2016). Available at SSRN: https://ssrn.com/abstract=2903469

3. Methodologies to Guide Ethical Research and Design

P. 45 - Transparency: some of the issues here have already been covered in committee 1 and committee 2. However, the issue of poor documentation, and the lack of an independent review organization are not mentioned by committee 1 and 2. We suggest the authors of the various sections discuss to reduce overlap and integrate new content.

Committee 4. Safety and Benefice of AGI and ASI

P. 49: Committee 4 focuses on AI researchers, but does not clarify whether AI researchers fall under the heading of 'technologists' – who are the main intended audience for this document.

P. 49: Committee 4 suggest the use of review boards; this recommendation is also made by committee 3 (p.44).

P. 50: Although most sci-fi interpretations of AI equate the increase in their capabilities with an increase in the danger they present to humanity, this is not necessarily true. For instance, in the case of automated cars the increase in their capabilities will significantly decrease the danger to individuals on the road. In the UK alone, the introduction of automated cars could save up to 3.000 lives per year, in larger countries this might be up to 30.000 lives per year. Furthermore, as AI becomes more powerful unintended behavior might also have a positive impact. We often forget that Asimov formulated a fourth – or rather a zeroth – law for robots: 'A robot may not harm humanity, or, by inaction, allow humanity to come to harm.' Fewer still remember, that it was a robot that suggested this zeroth law to Asimov. This all to say that, unintended behaviour can also have positive side-effects and that blanket fear of such behaviour is unwarranted.

Page 51 - Recommendation 2. Repeats some of the transparency recommendation made in previous committees.

Overall, the committee relies very heavily on the work of Nick Bostrom, although a formidable researcher, he presents only one part of the spectrum of the larger AI debate. The committee's work would be stronger if it would also acknowledge more of the work done by academics who have a slightly less pessimistic outlook on AI and AS.

Committee 5. Personal Data and Individual Access Control

P. 56: The committee mentions communal resource and the complexity of personal information, they might be interested in looking at this recent work by Taylor, Floridi, and van der Sloot on group privacy: http://www.springer.com/gb/book/9783319466064

P. 65: We think it might be interesting to add the groundbreaking research of Professor Sweeney to this section. She has pioneered studies on how ever greater amounts of personal data can be used for re-identification. Any discussion of this topic is incomplete without her work: http://www.forbes.com/sites/adamtanner/2013/04/25/harvard-professor-re-identifies-anonymous-volunteers-in-dna-study/ - 1a05da9c3e39

Committee 6 – Reframing Autonomous Weapons Systems

It would be great if the committee would start by outlining their definition of ‘autonomous systems’ and ‘harm’.

P. 69 – Candidate recommendation: There is an argument to be made that AWS are always unethical. This tension needs to be further addressed in candidate recommendations, as the committee does refer to resources that openly make this point but does not make any further statements about its position in this debate.

Committee 7 – Economics/Humanitarian Issues

P. 82 – Section 1: Automation and Employment

The committee holds that: ‘While there is evidence that robots and automation are taking jobs away in various sectors, a more balanced, granular, analytical, and objective treatment of this subject will more effectively help inform policy making, and has been sorely lacking to date.’ While more research is always welcome, there have been various reports that take the granular, analytical etc. approach the committee is looking for.

For instance:

Executive Office of the President National Science and Technology Council Committee on Technology. 2016. “Preparing for the Future of Artificial Intelligence.” Washington D.C. USA. https://www.whitehouse.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of _ai.pdf .

The report’s companion document, entitled the “National Artificial Intelligence Research and Development Strategic Plan”, details how to how R&D investments can be used to advance economic policies that increase economic prosperity on pp. 8-10. The plan is available at:https://www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf

European Parliament Committee on Legal Affairs. 2016. “Civil Law Rules on Robotics (2015/2103 (INL)).” Brussels Belgium: European Parliament.http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML%2BCOMPARL%2BPE-582.443%2B01%2BDOC%2BPDF%2BV0//EN.

House of Commons Science and Technology Committee. 2016. "Robotics and Artificial Intelligence." Fifth Report of Session 2016-17. London, UK.http://www.publications.parliament.uk/pa/cm201617/cmselect/cmsctech/145/145.pdf .

AI NOW recommendations report: https://artificialintelligencenow.com/media/documents/AINowSummaryReport_3_RpmwKHu.pdf

Cath, Corinne J.N. and Wachter, Sandra and Mittelstadt, Brent and Taddeo, Mariarosaria and Floridi, Luciano, Artificial Intelligence and the 'Good Society': The US, EU, and UK Approach (December 23, 2016). Available at SSRN: https://ssrn.com/abstract=2906249

P. 85: We would like to suggest that the Committee replaces the word 'underdeveloped nations' with the words "Global South". This is a more inclusive catch-all to indicate the difference in development between different parts of the world.

P. 85: The discussion on PII overlaps with the discussion in committee 5. However, the committee's focus on the Global South does not return in committee 5. Perhaps the committee chairs can discuss how to integrate the comments from committee 7 in committee 5, where they are a more natural fit?

The committee's recommendations are considerate, detailed and do a good job of providing nuanced suggestions, the introduction however is full of adjectives and reads unclear. It would be great if it could be rewritten to reflect the nuances found in the rest of the committee's text.

Committee 8 – Law

P. 89: We suggest this sentence: 'Lawyers should be part of discussions on regulation, governance, and domestic and international legislation in these areas and we welcome this opportunity given to us by The IEEE Global Initiative to ensure that the huge benefits available to humanity and our planet from AI/AS are thoughtfully stewarded for the future.' is separated into two sentences. As such:

“Lawyers should be part of discussions on regulation, governance, and domestic and international legislation in these areas. We welcome this opportunity given to us by The IEEE Global Initiative to ensure that the huge benefits available to humanity and our planet from AI/AS are thoughtfully stewarded for the future.”

P. 90: The committee holds that: ‘Although we acknowledge this cannot be done currently, AI systems should be designed so that they always are able, when asked, to show the registered process which led to their actions to their human user, identify any sources of uncertainty, and state any assumptions they relied upon.’

There are clearly limits to what systems can log and explain, however there exist methods to address these issues. Stating this is entirely impossible seems like an overly broad statement.

P. 90: This sentence: ‘Although we acknowledge this cannot be done currently, AI systems should be programmed so that they proactively inform users of such uncertainty even when not asked under certain circumstances.’ The double negation in this sentence confuses the reader, perhaps rewrite?

P. 91: This sentence ‘Government increasingly automates part or all of its decision-making.’ seems like a bit of an overstatement. Undoubtedly some government decision making is automated, but certainly not all. And certainly, not beyond the Global North. A rewrite introducing some additional nuance might be useful.

P. 93: Can the committee please define what they mean by ‘turning on the AI’?

P. 92 - 93: We suggest the Committee read and refer to these two articles:

Wachter, Sandra and Mittelstadt, Brent and Floridi, Luciano, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation (December 28, 2016). Available at SSRN: https://ssrn.com/abstract=2903469

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2903469

https://www.theguardian.com/technology/2017/jan/27/ai-artificial-intelligence-watchdog-needed-to-prevent-discriminatory-automated-decisions

And consider if any of the papers of the ML and the Law conference are of interest:

http://www.mlandthelaw.org/

P. 94 – Integrity of Personal data: Although an important issue, it is unclear how it relates directly to legal issues. As such we suggest the committee speaks to committee 5 and discusses how this last set of recommendations could be integrated into their committee.

[1] The opinions expressed in this article are those of the authors and do not necessarily reflect the view of their respective universities or the Alan Turing Institute.

[2] Alan Turing Institute & University of Oxford, PhD Candidate Oxford Internet Institute

[3] Alan Turing Institute & University of Cambridge, Professor of Computer Science

[4] https://www.youtube.com/watch?v=d4L_LTkKauI

IEEE's Global Initiative for Ethical Considerations in the Design of Artificial Intelligence and Autonomous systems

Written comments by ARTICLE 19: Global Campaign for Free Expression

Introduction

ARTICLE 19: Global Campaign on Free Expression (ARTICLE 19), a global freedom of expression organisation, welcomes the initiative of the IEEE and the participants of the Global Initiative for Ethical Considerations in the Design to develop specific guidelines on ethical considerations in the creation of Artificial Intelligence and Autonomous Systems (AI/AS). We believe it is crucial to understand in which ways AI/AS facilitate and hinder the exercise of the right to freedom of expression to determinate how they should be regulated in the broad political sense, and what demands can be made on companies to develop codes of conduct for their technologists.

In this submission, we provide detailed feedback on first version of the IEEE's Ethically Aligned Design (EAD) document and highlight the main legal issues raised for freedom of opinion and expression within the document. We also offer some concrete suggestions on how the IEEE can reflect the existing standards on corporate responsibility and human rights in this respect, focusing specifically on the responsibility of industry, and individual technologists. The comments are provided in chronological fashion, following the structure of the document. We only provide comments that are relevant to our mandate (protection of the right to freedom of expression), however, the fact that we do not comment on all section should not be understood as an endorsement of respective sections.

We understand this document is a first draft that will be further developed in the upcoming months. We welcome the opportunity to provide some initial comments on this work, and look forward to working with the Global Initiative for Ethical Considerations going forward.

Overall comments:

1. *Conceptual basis for the document*: We welcome the fact that the EAD document – whether it becomes a formal code of conduct or is used as a best-practices document – starts with recognition to ensure that AI/AS do not infringe human rights (Framing the Principle of Human Rights). Indeed, it should be based on international human rights standards and standards of international humanitarian laws (for armed conflicts) throughout. The document considers some human rights standards in certain sections, but not in all. In various sections, the committees hold that incorporating end-user values is crucial. While these might sometimes overlap with legal standards, they are not necessarily the same. End-users can hold various values, based on their experience, or cultural background. However, these might not necessarily correspond with international law (e.g. some users from patriarchal societies might consider women subordinate to men while international law requires gender equality). The EAD should recognize that in these types of situations companies will need to respect the key guarantees of international human rights laws or higher ethical values than what the international law providers - to decide which of these end-users' values will (and will not) be incorporated into their AI/AS systems.

2. There are several suggestions made throughout the document that might have a negative effect on the right to freedom of expression, as for instance the AI/AS information clearinghouse, these specific issues are covered in the submission below and should be updated for the next version.

3. Implementation of the EAD document: although the document aims to educate technologists, it is unclear to what extent. Some of the committees provide very narrow detailed recommendations for technologists, whereas others present more broad principles. Sometimes these broad principles are directed at technologists, but in other cases, they are geared towards politicians and regulators. Whilst important, it would be good to further define whether such content is directly relevant to technologists (it can for instance be by providing technologists further insight into the legal and regulatory ecosystem in which they operate – and what responsibilities lie with regulators and which with them) and which is not.

4. *Definitions of key terms*: We believe that at the beginning of the document, the terms 'Artificial Intelligence and Autonomous Systems (AI/AS)' should be clearly defined and the definition should be consistently applied and cross referenced through the document (e.g. in relation to Committee 4 or Committee 7). The Initiative should ensure that all stakeholders are working with a comprehensive and shared definition of AI/AS. Any other relevant concepts for the specific committees should also be defined and explained. These shared concepts should be based on internationally recognized human rights and international law legal constructs (see above). This is necessary to ensure the recommendations made are coherent, and the discussion is accessible for outsiders. Definition of other key terms – e.g. "harm", should also be provided.

Committee 1. General Principles Committee

As noted above, we welcome the inclusion of crucial human rights documents in the language of this committee. The opening statement 1 ("AI/AS should be designed and operated in a way that respects human rights, freedoms, human dignity, and cultural diversity") is crucial and should find further resonance throughout the entirety of the document.

P. 15 The committee mentions that it is developing principles for *all types* of AI/AS – mentioning this includes both robots and software AI. If the definition is not provided in the beginning, the Committee should define AI/AS here.

P. 18 Recommendation 3: we welcome the call for the development of multi-stakeholder ecosystems to ensure norm development happens with broad stakeholder input and support.

P. 21 - Principle 4: Education and Awareness: Although important, the recommendations are aimed more at the public than at specific technologists. It would be good to further define how these recommendations are relevant to technologists.

Committee 2. Embedding Values Into Autonomous Intelligent Systems Committee

P. 22: The committee assumes that: 'a community's network of norms as a whole is likely to reflect the community's values, and AI/AS equipped with such a network would there for reflect the community's values'. We observe that this is a rather sweeping assumption. It does not consider the various power structures that go into defining overarching norms for a community. If we translate this statement to political reality, it is like stating that everyone in North Korea enjoys living under repression, or that everyone in the US is in favour of building a wall to keep out immigrants. When considering how network values arise, we need to consider the power differentials between, for instance, the people and their political establishment, or the people and their religious institutes. We cannot take 'majority' or a prevalent network of norms, to be the norm. This is acknowledged later in the document (p. 27) but perhaps this statement can be nuanced to reflect the influence of power in the opening statement as well.

It also makes a second assumption: namely that people's values are always what is ethically just. This is clearly not the case. Just because a group holds certain beliefs, does not make them ethically just. Think for instance of societies where Female Genital Mutilation (FGM) is a prevalent cultural practice, or societies where certain minorities (whether sexual, religious or otherwise) are persecuted for their belief systems. These practices are reflective of values in the particular community, but go against the standards provided in international human rights law. Hence embedding values into a system based on the prevalent norms of a certain community, does in no way guarantee and ethical (or even a legal) outcome. The committee should be wary of making such recommendations.

We suggest that they add a nuance to the language by adding to the language that refers to including end-user values into AI/AS the sentence 'to the extent that such norms or values fully comply with international human rights law'. For example, a section on p. 22 would read:

> "A community's network of norms as a whole is likely to reflect the community's values, and AIS equipped with such a network would therefore also reflect the community's values - **to the extent that such values do not violate international human rights law** - even if there are no directly identifiable computational structures that correspond to values."

P. 25 - Issue 2: Moral Overload: The background text seems to suggest that technologists can weigh legal requirements on equal footing with other requirements, like monetary constraints. Even though situations can arise in which certain countries can make unreasonable demands on technologists to build certain technology (for instance a Muslim registry), the response to a legal dilemma is fundamentally different than those to a monetary or ethical issue. The language should reflect this.

P. 34 - Paragraph 3: This text (“We also.. to be deployed”) is written as if technologies will always be deployed in a location where human rights are fully respected with the rule of law. This is not the case everywhere, and the text should go beyond focusing on ‘a minimum level of value alignment’ to include the text ‘as in concordance with international law standards’.

Committee 3. Methodologies to Guide Ethical Research and Design

Again, we reiterate the need to have a definition of AI/AS if it is not provided already at the beginning.

P. 36: There is a tension in this text, as it focuses both on the Universal Declaration of Human Rights but also on the importance of end user values. We recommend that the first paragraph is expanded to state “greater emphasis on human rights, **as provided** for in the Universal Declaration of Human Rights and **other international human rights standards**, as a primary form of human values.” This is important to reflect the developments in the international human rights framework since the Universal Declaration.

p.42: The ‘Lack of Values Aware Leadership’ recommendations should include as one of its recommendations that companies consider their obligation to respect international human rights, as laid out in the UN Guiding Principles for Business and Human Rights[1], also known as the Ruggie principles. The same holds for the lack of ownership and responsibility issue.

When discussing the responsibility of private actors, the UN Guiding Principles on Business and Human Rights should be fully reflected. These principles have been already widely referenced and endorsed by corporations and led to the adoption of

[1] http://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf

several corporate social responsibility (CSR) policies in various companies. As such they have led to a better understanding of the role of businesses in protection and promotion of human rights. The IEEE should consider developing their understanding of ethically aligned design based on the Ruggie principles, to ensure the most crucial human values and legal standards of human rights are respected by technologists."

Committee 4. Safety and Benefice of AGI and ASI

P. 49: This section would also benefit from a definition of the AI systems. It should also clarify the differentiation between AGI and ASI.

Page 52: The candidate recommendations should include a reference to the work done by the IETF on security recommendations, which asks individual engineers to think through the security implications of their work. The security considerations need to be within a reasonable limit, or a standard will not be approved for the standards track. This system provides a measure that both engenders a safety-by-design approach in engineers, and an organizational stop on bad technology.

Similarly, the Committee might be interested in the work of the Human Rights Protocol Considerations Group (HRPC) at the Internet Research Task Force. This group is developing human rights protocol considerations for engineers, asking them to document (and mitigate) the potential impact of their technology on human rights.

P. 53: Committee 4 (as well as Committee 3 on p. 44) suggest the use of review boards. In this respect, it would be useful to specify minimum requirements for these boards. For instance, it should clarify that the board should be open and transparent (like with Lucid.AI) and not opaque (such as Google Deepmind). It should also clarify that whether those should be build on industry input, or rather on academic best-practices.

Committee 5. Personal Data and Individual Access Control

P. 56: We believe that the committee should clarify how do they define concepts such as 'personal information' or 'data asymmetry' (unless this is provided overall for a whole document as per our suggestion above).

P. 56: We appreciate that the Committee wishes to inform itself of the standards on data protection and the standards such as The General Data Protection Regulation, while recognising that the non-Western standards should be reflected. We note an ongoing efforts to provide for data protection rights (usually based on the right to privacy) and we suggest that the Committee takes a note of the international standards, including the resolutions of the Human Rights Committee or the initiatives of the civil society (e.g. the forthcoming Expression and Privacy: Principles on the right to freedom of expression and privacy in the digital age, by ARTICLE 19).

P. 59: The Committee might want to consider reflecting the discussions on the issues raised in the new Data Protection Directive to the text. It might also be worth to include additional information about the data protection regulation[2] that will apply from 25 May 2018 onwards.

P. 62 – 63: Issue - How to redefine data access to honor the individual? ARTICLE 19 welcomes the recommendations made by the committee, especially their focus on consent, open standards and interoperability. We also think the who, what, why, when tool is an excellent method to ensure technologists think about these issues. We would like to see more of these practical tools for the other issues highlighted by the Committee.

We are however considered about the Committee’s use of the ‘right to be forgotten’ as “a core design capability” in “European context”. We point out that the “right to be forgotten” usually refers to a remedy which in some circumstances enables individuals to demand from search engines the de-listing of information about them which appears following a search for their name. It can also refer to demands to websites’ hosts to erase certain information. More broadly, it has been considered as a right of individuals "to determine for themselves when, how, and to what extent Information about them is communicated to others” or as a right that gives the individual increased control over information about them. It has been categorised as a privacy right even though it applies to information that is, at least to some degree, public.

[2] http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2016.119.01.0001.01.ENG&toc=OJ:L:2016:119:TOC

The "right to be forgotten" is expressly recognised neither in international human rights instruments nor in national constitutions. Its scope remains largely undefined. It came to the fore with the decision of the Court of Justice of the European Union (CJEU) in the Google Spain case of 2014, however, this issue is not limited to Europe, as since the CJEU judgement, several states outside of Europe either have adopted a dedicated "right to be forgotten" law or have been looking to adopt new laws on the subject.

ARTICLE 19 is concerned about the implications of the "right to be forgotten" for the right to freedom of expression. In a recent policy brief[3], we proposed a framework solution to the issues raised by the "right to be forgotten," grounded in international human rights law. Hence, we find it problematic that the EAD document seems to suggest that it should be granted core design capability; hence we suggest removing the reference all together.

Committee 7 – Economics/Humanitarian Issues

P. 82 – "Issue: Misinterpretation of artificial intelligence and autonomous systems in media is confusing to the public."

We find this section extremely problematic. It suggests that in order to diffuse some unspecified "confusion" of "the public", and should be remedied to ensure "objective debate" and prevent sending "a wrong message to the general public".

It is not clear what is meant here. Does the Committee mean the way feeds are organised on social media? Or is this a reference to the application of algorithms that enable visibility and findability of news produced by various actors? Does it refer to the advertising models applied by the media? It is not clear.

The creation of an "independent information clearinghouse" would be equally problematic and would seriously hamper the diversity and pluralism of the media. We note that the international freedom of expression standards guarantee the rights of individuals to access and impart information without frontiers and this right is guaranteed to everyone, regardless of professional association.

[3] https://www.article19.org/resources.php/resource/38318/en/policy-brief:-the-right-to-be-forgotten

Assigning one entity with the mission of fact checking and information would inevitably create a chilling effect upon the media and anyone who contribute to public debate. Facts are by their nature complex and intricate, to the point that it is truly impossible to avoid slight inaccuracies. Suggestion that a clearing house would ensure "objective statistics" or "fact-check" information that are absolutely true would simply be impractical. International case law has indeed recognised that journalists contributing to public debates on topics of general interest have the right to a certain degree of exaggeration or even provocation as well as satire, humour or provocation. Doing otherwise via AI/AS would enable abuse, and silencing critical voices.

Committee 8 – Law

P 89 (onwards): ARTICLE 19 welcomes that the recognition of legal implications of AI/AS and the recommendation that the solutions in this area must comply with international law. We make the following recommendations to this section:

- On p. 89, the Committee states that development, design and implementation of AI/AS should comply with international and domestic laws. We note that as far as human rights protection is concerned, many domestic laws fail to meet international human rights standards and/or are in direct violation of these standards. Hence, the compliance with the international human rights standards and their progressive implementation should be ensured.

- The section provides the key principle areas for recommendations – "governance and liability, societal impact and human in the loop." This is certainly and interesting framework, however, we suggest that "human rights-based approach" is applied to AI/AS instead and it should underline any recommendations in the Law section.

  A rights-based approach is a conceptual framework for a process of development that is based on international human rights standards and directed at promoting and protecting human rights, analysing inequalities, and redressing discriminatory practices and the unjust distribution of power.[4]

[4] Human rights-based approaches have been applied to development, education and reproductive health. See: the UN Practitioner's Portal on Human Rights Based Programming: http://hrbaportal.org.

Borrowing from this concept, the rights-based approach to AI/AS should be based on

- linkage to human rights standards: human rights standards contained in, and principles derived from, international human rights instruments, should guide the policy development and implementation of AI/AS. As such, the rights-based approach shall identify the rights holders and the duty bearers, and ensure that duty bearers have an obligation to realise all human rights;
- accountability: the state should be accountable for its policy in support of AI/AS. As duty bearers, should be obliged to behave responsibly, seek to represent the greater public interest and be open to public scrutiny;
- participation: the rights-based approach demands a high degree of participation of all interested parties
- non-discrimination: principles of non-discrimination, equality and inclusiveness should underlie the practice of AI/AS. The rights-based approach should also ensure that particular focus is given to vulnerable groups, to be determined locally, such as minorities, indigenous peoples or persons with disabilities;
- empowerment: the rights-based approach to AI/AS should empower rights holders to claim and exercise their rights.

This conceptualise framework is better suited for the issues addressed in this section and recommendations contained there would still be applicable.

Tom Kuriahra, TKstds Management, an independent consultancy

Ethical Design

Pg 96. "…urgent need to broaden "traditional" ethics beyond the scope of "Western" ethics, e.g., utilitarianism, deontology, and virtue ethics; and include other traditions of ethics, e.g., Buddhism, Confucianism, etc."

Note. "…other traditions of ethics," cite religious beliefs that are unrelated to the concept of "Western" ethics.

Pg 96. "The attempt to implant human morality and human emotion into AI is a misguided attempt to designing value-based systems."

Observation. Reality is that many writers and speakers attribute human characteristics to inanimate objects, psychologically perhaps, to communication familiar concepts to those who may or may not be familiar with the technical topics of discourse, e.g., car that think, whereas in reality the logic is rule-based and not at all similar to the human thought process. The resulting misinformation creates unreasonable expectations of technologies. Designs of devices are many times are not intuitive to new users who are lacking cognitive skills to adopt and to adapt to new devices. Question. Is ethical design directed to address what has been observed?

Pg 98. "…VR systems…ultimately it could be a way to teach ourselves new ways to think and create content…."

Interpretation. Based on the second note on Pg 96, implies that VR systems can teach to think differently; to the contrary, VR systems may create different neuronal pathways for processing external sensory input, but without the stimulus may not affect the creation of "new" thought processes.

Pg 98. "…does it (mixed reality) exacerbate existing power inequities?

Affirmative. Reasoning. Gravitation to pleasure centers and preferred realities based on psychological inclinations has been witnessed with video games, "reality-type" television programs, "soap opera" serial programs, and "strange news" such as videos or snapshots of Walmart customers. Unpredictable, irrational, and impulsive behavioral and ethical issues may very well be the result, complicating the understanding of human behavior and ethical reasoning.

Pg 99. "…nudging for social good."

Question. What is the baseline or basis for determining "social good." Differs with different governance styles, familial relationships, individual set-point for ethical thinking, and social and environmental conditioning.

Pg 99. "Should we, and if so how do we, regulate computing and robotic artifacts that are able to tap into the affective system of humans…?

Have there been studies and models developed for those who watch television programs, binge on streaming content, play video games, play competitive interactive games, and obsessive gamblers? All are being influenced by the artifacts with which they are interacting; and in many instances carry over the influences in personal relationships, and interaction with those who have different views and experiences.

Pg 99. Can intimate devices such as robotic fuzzy animals with AI learning capabilities assist in therapeutic uses?

Witnessed demonstration of use in hospice and elderly patients in Japan, and imputed positive result and comfort to varying degrees. Most likely the degree of effectiveness is based on the psychological need and ability to react to the perceived intimacy.

Pg 100. "…lead humans to falsely identify with the AI."

Point. If used in circumstances where its applicability is minimal or not intended. Thus far, the purpose of AI is to promote the symbiosis with intended human operators and users.

Pg 100. "…are there fundamental individual rights that transcend these utilitarian arguments?"

Note that individual rights are affected by the governance systems under which live and the need to differentiate between perceive rights and explicitly or traditionally given rights. Opinion. Questions and issues may be reconstructed to address the geographic, societal, ethnicity, and individual rights. Rhetoric used about the "world" as fast changing, however, individual values, ethics, and morality changes more slowly and affected by the environment, living, conditions, and familial influences that may not have evolved to any extent. AI and autonomous systems tend to be based on past experiences and not what has not been known or all instances of conditions that creates circumstances that has no precedence.

Pg 102. "IEEE can help ensure these efforts are guided by appropriate ethical principles."

Opinion. The term "ensure" implies a "guarantee." Uncertain that this is the intent, however, IEEE writes seem enamored with the use of "ensure." Example of policy in the IEEE CS LMSC in the prohibition of the term in standards may provide insight into why the term is likely to increase IEEE liability if the adherence to IEEE does not or perceived not to "help ensure" the outcome.

Pg 102, Issue 2.

Critical Consideration. Silent on cost and effect on budgetary and skills required to accommodate AI/AS. Recognizing the inherent weakness may well be the "Achilles heel" of Ethical Design.

Pg 103, end Notes.

Observation. Focus is on ethics, obviously, but not so obvious is that it is from the inanimate object perspective and not the human psychology, emotional intelligence, and social and cultural influences that predate the introduction of AI/AS. Conclusion, framing of issues and supporting background may be focused too heavily on the machine aspects and not fully factoring in the complexity of human character, personalities, learning skills, and incentives to adapt outside of a "comfort zone." Security in living and safety considerations may transcend ethical design precepts and accommodating AI/AS devices. Emotional intelligence, lesser known than AI or tested-based IQ. Unknown are the factors that permits some individuals to attain high scores on tests, yet lacking in other aspects of learning and wisdom, that is, the circumstantial application of knowledge.

Pg 108, ExCom

Observation and Personal Opinion. Highly visible, academic and research oriented backgrounds, with “fingers in many pots” except social sciences, human behavior, normal and abnormal psychology, and religious studies. Has a perceived “western” leaning.

Pg 109, Global Initiative.

Observation and Personal Opinion. Some familiar names, for good or not. Academic and technology research-oriented backgrounds with Western leaning with no background in “non-western” aspects of human interaction and behavior, and religion.

Pg 115 to end, Committee Description. Similar opinion as in comments for page 108 and 109; and very “Western.” I would have been pleasantly surprised if the volunteers and recruits were better balanced in human and machine behavior, ethic being intangible and machine technology more tangible than not and the functionality is all machine technology-based (software created by programmers).

Written by unqualified and uncredentialled IEEE Affiliate member.

Thomas M Kurihara

April 4-6, 2017

February 21st:

My short title would be, **Dr. Yanqing Hong**, Researcher at the Cybersecurity Research Institute of Sichuan University; Lead, the Standardization Task Force of the Personal Data Protection Specification, National Information Security Standardization Technical Committee of China.

Unfortunately, I will not be able to join the coming Committee call as I will be on a trip to the EU Commission as a member of Chinese delegation. However, I would like to take this opportunity to offer one humble suggestion, knowing that I am very very new to this committee.

Maybe we could start, in the Chapter of Personal Data and Individual Access Control in the EAD (Page 56), not with the "data asymmetry", but with "loss of identity development and autonomy". Below are my thinkings:

1. the whole purpose of this project is to align the AI/AS to the ethical considerations. Data asymmetry is not one of the ethical considerations per se, but more of a description of imbalance position of data subject and the parties that collect, use personal data. So it might be better to start the whole chapter with a risk to ethical consideration that may be brought by the technology development.

2. I particularly like how the ECSD frames the issues regarding the Personal Data and Individual Access Control, I find that quite comprehensive. And I think all of the issues in that chapter point to one great danger: losing control of one's personal data means being deprived of the chances to fully develop one's identities and ultimately leading to loss of one's autonomy. Data is being recorded anytime, anywhere without one's full knowledge and control, further more data is analysed in a manner that escapes one's attention and expectations. This means one can not make mistake, one can not change or evolve, as everything is being documented and will come back to haunt you. Meanwhile it also means that one can not have multiple identities in different contexts. All of these lead to diminishing one's potentials and identities. So I think regaining control of one's personal data means to be able to steer one's life surrounded by IoTs and AI/AS.

So I think it would be better to articulate our ethical consideration in the very beginning of the chapter so readers will understand what we have in mind right from the start. And this way could provide a good context for readers to navigate during reading the chapter.

Comments on The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems

Manfred Bürger, 24/2/2017 – mod 7/5/2017

General comments

The initiative is highly important in view of the deep interventions envisaged and to be expected from the technical development of AI means and their implementation in the working process as well as social processes and relations, especially concerning the changes induced in human – machine, human – human and machine – machine interactions and relationships, the generalized crosslinking and networking concerning all production and social areas. The perspective of joint design of a future society is an inherent possibility in these technologies. However, the risks are also high. They are connected with abuse, but also with difficulties of technical control, with uncertainties involved in their functioning and behavior. Therefore, the emphasis in the initiative on a human perspective, on ethics, on transparency of the technical means is to be agreed.

However, the criteria as well as steps to be undertaken, the possibilities and means of implementation discussed and envisaged remain rather general or too much focused on technics. I doubt that it is possible to concretize in a way that machines can be formed with a behavior implemented in advance which guarantees human orientation and goals. Obviously, such a goal was even not reached with humans by general formulations of ethics, corresponding education and implemented rules and laws.

Even to concretize goals and ethics remains difficult, problematic and controversial, e.g. looking at critical decisions about helping people with severe illness, from palliation to assisted suicide, use of weapons, risk evaluations or even social regulations about equity of wealth. In order to concretize, ethics must be considered in a practical perspective, i.e. taking into account the conditions and interests. It requires concrete analysis of situations and conscious social debates, struggling about a collective perspective.

Thus, a perspective of implementing ethics in advance in AI appears to be limited or even questionable in principle, although general requirements as supporting, adapting, user-friendly, transparent behavior, avoiding hurting, can be considered in software and already construction as covering approaches, as e.g. done in the work of I. Boblan, Berlin (www.biorobotiklabor.de).

Further, concerning implementations in AI, it appears hardly to be possible to evaluate the way artificial learning processes go and which results they produce, especially with most effective methods of AI as neural networks. Thus, even transparency can hardly be reached (see e.g. L. Muelhauser: https://intelligence.org/2013/08/25/transparency-in-safety-critical-systems/). E.g., with pattern recognition, a spectrum is opened from realistic recognition from few pixels, e.g. R. Dahl (https://arxiv.org/abs/1702.00783) to dreaming about realistic pictures (https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html). Or: Success of AI in Go play was connected to unexpected moves, unusual for human players and not yet understood in the underlying reasoning of the machine (see e.g Muehlhauser).

In the survey of Muehlhauser, the questionable transparency just of the most effective AI methods is clearly outlined, especially concerning machine learning methods as ANN (Artificial Neural Networks) or Deep Learning techniques, but also concerning EAS (Evolutionary Algorithms) and even Logical Methods. Although the latter are considered to be more amenable to formal checking methods, thus verification, even this remains limited to “relatively modest applications” (p. 10) and fails with open systems, operational environment and user errors, finally also even more practically with limitations from resource, time and financial aspects (which should not be a limitation!).

However, I miss in the paper the most severe limitation which in my view is due to model uncertainties including unforeseen effects of nonlinear behavior. Validation with this respect must be based on empiricism. Formal verification (using this term as usually done in quality procedures) is by far not sufficient, even if possible (rather for logical methods, but even there with limitations, see Muehlhauser), since it only can yield assurance of programming consistency. Footnote 12 in Muehlhausen report may indicate this concern by referring to countless and unforeseeable interactions. The question of change with scale raised in this report in the final section and also in the IEEE paper on p. 51 is important but mainly a sub-question about model reliability, which is not to be restricted to questions of code scale.

Thus, instead of human requirements to be implemented into machines, what is needed are social implementations about procedures to handle the machines, to regulate conflicts, to elaborate common goals, to decide about them, more technically to evaluate risks, to understand essential features in complex processes. Risk evaluations and measures against risk are tasks which lead to implementations in machines and also in handling. Some guidelines in the context of nuclear safety, where considerations and implementations in advance are required to a high extent due to the risks of severe accidents, have been given by T.G. Theofanous (ROAAM – Risk Oriented Accident Analysis Methodology, see “Risk, Severe Accidents, and Thermohydraulics”, NURETH 10, 2003). They are based technically on distinguishing between areas where clear conclusions can be drawn from physical principles and knowledge and uncertain ranges. Even in accessible areas, due to the complexity, the establishment of different, conflicting views between institutions and implementation of joint elaboration procedures is required, yielding consolidation of knowledge, defense-in-depth in evaluation and application as well as convincement about fitting for purpose based on understanding. The same appears to be necessary in the AI field.

Such procedures are in principle in line with democratic rules. But, we realize how their functioning is in danger, on the one hand, by insufficient knowledge, by dominance of emotions raised from social disruption, missing perspectives and activated by influences from media, by fear about new technologies and their threats, such as loss of jobs, extended qualification and flexibility requirements, orientation problems in an increasingly complex world with loss of pre-defined identities and transparency. On the other hand, the given orientation on capitalist profit maximization and economic growth produces inherent constraints and excludes alternative paths. Problematic requirements and mere destruction of traditional ways and values produce uncertainties. These restrictions undermine free decisions and possible ways to draw use from technological development for the whole society. Due also to the increasing complexity, decisions are more and more delegated to experts who are usually linked to the given constraints, seemingly existing technical needs and the interests of dominating power. Real democratic evaluation and decision processes can therefore not develop and are even endangered.

E.g., decisions on requiring driverless cars are linked to the implicit goals of an individualized traffic system, linked itself to the interests of car industry. Other possibilities are excluded, e.g. financially supported public long distance transport combined with local public or cheap renting car systems. Or: genetic technologies may be used to help people against illness, but may become dangerous if requirements to optimize humans for other purposes dominate. Biotechnologies can help, but can also destroy if e.g. just applied to optimize output in agriculture in short-time perspective. Fuel production problems may be solved by genetically transformed bacteria, but needs and risks can be reduced by alternative technical and especially social organization means (see above for traffic). Similar problems can be envisaged for the use of AI, just due to the given constitution of society.

Further, failures as well as unexpected and not intended behavior with complex systems can in principle not be excluded completely (see above). Therefore, concepts of redundant barriers and counteractions as well as procedures covering critical ranges and conditions have e.g. been introduced in reactor safety to exclude especially risks for the environment. Dealing with such systems requires in principle permanent considerations on possible behavior based on permanent observation, handling and analyzing of experiences with feedback to procedures and design developments. This must be done best in a way allowing, even requiring different views and initiating joint clarification processes (see also work of Theofanous cited above). Catastrophic failures in such accidents as Fukushima demonstrate the insufficient implementation of such control, if not the impossibility to control the underlying technics at all.

Correspondingly, it is also not sufficient, even not in the sense of reaching best practice, to have only one person in a control room for an automatic chemical production (which was demonstrated to us as a success in a visit at a famous developer, ABB), although the controlling system may detect failures and initiate actions from the human observer. If the model behind has errors, considerations may be required deeper than expected. Thus, an educated and experienced team may be required for solution. Further, capabilities to solve in unforeseen events, but also to consider improvements and optimization as well as innovation, require permanent experience and permanent thinking about experiences.

These requirements come increasingly from the new high-technology processes and yield in principle arguments against just dropping of jobs and replacement by "intelligent" machines. The general shift of working from manual operation to control and regulation activities results from increasing capacities of "intelligent" machines and related to this the increasing needs to deal with system behavior, failure problems, safety, etc.. Adequate working processes require qualified and trained teams understanding the processes and the system behavior, able to conclude from experiences on failure behavior and possible improvements, even or just if the main production and in general working processes including control are delegated to machines and automatic control systems. This social implementation is also the best way towards safe systems. And, of course, this is also the best way to establish ethics control and decision making for AI techniques with human goals.

Thus, a major task is, in addition to elaborating rules and regulations, to establish a culture for and in the working processes, combining machines and humans, better, forming conditions under which humans can really control the machine controlled processes, can gain experiences, to be jointly converted into know-how, important just with the complex processes in order to develop a joint understanding of the essential features, i.e. about what is important, what are major influences and effects. This can only be done in teams with different, even controversial views. Employers, companies, managers have to be convinced to establish sufficient and well-educated workforce with the new techniques, in their own interest, although there exists a conflict with profit interests, at least in short term. Thus, trade-unions must also be active with this goal.

The organization of work must reflect the necessary joint interplay, i.e. a culture of cooperative work must be developed as best means for permanent control, failure and misbehavior detection, improvement and risk prevention. Finally, this must also be extended to use the technical options for gaining a human perspective in general and for defining and pursuing corresponding goals. Such goals are indicated in the initiative, but need more for realization than considerations for implementations in techniques and installing committees. Social processes and politics are required and are to be initiated.

In Germany, since 1991, the Forum Soziale Technikgestaltung (FST, forum for socially sustainable design of technology: http://sustainabilitymaker.org/partner/forum-soziale-technikgestaltung/), guided by W. Schröter, is working, linked to the German Federation of Trade Unions, with the objective of developing human and sustainable working environments in view of new technological developments and to establish social standards under discussion with industry, science and politics. This is a process including the employees of relevant companies in which changes are taking place (see also: http://www.blog-zukunft-der-arbeit.de/).

Theoretical and empirical work on the technically influenced development of working types, to which such approaches are related, go back to the 1970s, being e.g. performed in depth by a research group "Automation und Qualifikation", led by F. Haug (see e.g. Das Argument 154, 1985, p. 813). This work is now continued in InkriT (Institute of Critical Theory (http://www.inkrit.de/neuinkrit/index.php/en/). I am participating in both activities mentioned. In addition, F. Haug has developed a vision of work in society (http://www.friggahaug.inkrit.de/documents/4in1_englisch_fin.pdf) which accounts for the dramatic technical changes as well as gained productive power and gives a perspective of participation of all people in producing and designing the future of society. This can be taken as a guideline for the necessary social processes to be induced in view of the technical development, in order to get capacities for decision making and control.

Comments addressing the different sections

Ad 1) General Principles

Here, ethical principles for AI/AS lead to goals of human benefit, responsibility, transparency with education and awareness to be activated for benefit and against misuse. In addition to committees for ethical and quality survey of AI programming and implementation of automation means, it appears necessary to envisage especially ways how to influence organization of work, best already during planning and implementation of technical changes, in discussion with employers and employees, activating also trade unions for this.

General declarations of human rights are not sufficient, also not appeals to design groups. Real processes must be established and developed between the major agents on how to implement techniques, how to organize working processes, the relation between experience in process and developers, how many employees are required for certain tasks, for yielding good work, good control. For safety and transparency: technical implementations, rules, to be developed also in permanent survey and debate. Emphasis should be put on this.

Concretize in this direction: Means, steps to be undertaken to activate for such necessary processes of interaction. Experts to be included, but as partners, supporters in these processes, forced to give understandable explanations about possibilities, risks, alternatives. Education requirements (basic and permanent) to enable in general dealing with complex processes. Learning in an activity-, project-oriented way, enabling different views on a subject, rather by self-constructing than by being instructed, enabling own paths, cooperative working, dealing with conflicting goals in subject as well as in team, weighting about goals and values. This all is required as a basis for being able to understand complexity, to understand what is important, which are the relationships, the interrelations, the context-dependences, to get an overall perspective which also includes ethics.

Ad 2) Embedding Values into AI Systems

Due to the limitations of this embedding (see general comments and indications in text of initiative), there are needs to develop social processes as sketched with Ad 1). "Iterative process …proactive inclusion of users" on p.23 of the initiative may be further interpreted in this sense, especially related to point 3. on p. 22. Conflicting aspects, not only concerning ethical norms – "moral overload" (p.23) – , are essential features of complex systems, also technically, thus to deal with is important for their appropriate functioning, for safety, for understanding and transparency. May be added in text. Essential features are to be elaborated to get not lost in details. Task of experts with requirement to be understandable. Only possible in cooperative approaches, including developers and users, theory and practical experience. Candidate Recommendation on p. 24 goes into this direction although still addressing researchers and designers, only.

"Moral overload" (p. 23) depends on conditions, i.e. decisions not to be limited to the frame of given systems (e.g. individualized traffic), but allowing alternative possibilities.

Formulations as "we strongly encourage the inclusion of intended stakeholders in the entire engineering process, from design and implementation to testing and marketing" or "we also recommend, …, that designers take on an interdisciplinary approach and involve relevant experts or advisory group(s) into the design process…" on p.27 go in the direction of organizing processes of debate. However, this should be done in general, not only or especially for vulnerable people addressed here and should not only take place within committees of experts and representatives in the design process. It should be part of reorganization of the whole working and social processes in a way outlined above, with permanent control and feedback in cooperative work.

Ad 3) Methodologies to Guide Ethical Research and Design

Values-aligned design methodologies are considered as essential focus, with the major goal "that machines should serve humans". Again, emphasis should be laid on the social processes to reach this, as basis for embedding values. More important than final technical embedding in machines is embedding in the organization of such processes, kind of working processes, culture of cooperative work, of social debating.

The issue "Lack of value-based ethical culture and practices for industry" on p. 6 (Executive Summary) also indicates this.

On the other hand, the need of interdisciplinary and cooperative approaches in education and technical development should not be derived only from requirements of ethics (p. 37: "We also recommend establishing an intercultural and interdisciplinary curriculum that is informed by ethicists, scientists, philosophers, psychologists, engineers and subject matter experts from a variety of cultural backgrounds that can be used to inform and teach aspiring engineers"). Such requirements would then only be introduced as external ones to the design process, i.e. usually not taken too serious. However, interdisciplinary procedures are required from the core of complex processes, to be able to understand and handle them adequately.

They are also required since the goals of design processes are more and more extended beyond narrow technical aims, e.g. considering environmental demands such as reduction of waste gases. The requirements have increased and become more complex due to interrelations of processes and results as well as significance of effects for nature and humans. Thus, ethical considerations are required in a necessary frame of developing generalized views, different approaches, alternatives, finally on how to develop society. This should be more clearly outlined. Not just need of interdisciplinary procedures, but of general view, general goals, also not only additional view for developers, especially not ethics on top of usual work, but general tasks for society. E.g., ethical considerations of a committee in a hospital about how to treat specific cases of terminal illness miss the more general question of a culture of treatment depending on amount of employees and their qualification as a basis.

Ad 4) Safety and Beneficence of AGI/ASI

Here, a "more complex set of ethical and technical safety issues" is related to safety questions due to unanticipated behavior of AI (p. 7). On p. 49, it is recommended that AI teams cultivate a "safety mindset" and develop systems which are "safe by design". Again, it should be clarified that this means a permanent process of checking and control, not only to be established by appeals to the designers, but by their inclusion in a general process, at least as far as applications are envisaged. Introducing review boards will not be sufficient. The special problem here is the conflict that the most effective AI methods are the most uncertain and the less transparent. This can be understood due to the attempts to introduce learning behavior. Thus, methods to stop derailing behavior are to be used, similar to defense-in-depth in nuclear safety. Again, this requires permanent control in cooperative procedures. Further, on application of developed means, it has to be decided with respect to application areas: uncertain, but most effective, interesting ones may be limited in application to less critical fields – decision of society.

Emphasis should be put on understanding of models behind programming and model behavior. This is indicated on p. 51 under 2., but could be more clearly separated from just understanding and detecting programming effects (see also under General Comments). The problem of moving from a small testing environment to a large world also addresses both problem areas. What can the aim to understand reasoning processes (p. 51) mean with learning systems? In its consequences? Probably this has to go deeply into understanding how learning works depending on the kind of modeling, kind of adaption to effects, reaction of environment, etc.. I.e., again this requires deep and continuous considerations within and between development teams, to be organized.

The final recommendations of this section are only appeals to the researchers since no procedures are envisaged (p. 54), no means of developing the required culture are considered. The aim to decouple technologically implemented behavior and its results from the attitude and functioning of teams (p. 54, above Issue) appears to be unrealistic, even counterproductive. The teams must instead be included in a functioning exchange and control process.

Ad 5) Data Control, 6) Weapons, 8) Law

In my view, these are areas where control loss can yield existential results. But, they pose no additional basic questions than treated in the other areas. Basic questions on development and use of AI are treated there. AI may support specific solution options of problems, but also risks in the fields addressed here. In this sense, also other specific fields may be considered as e.g. medical applications. This does not mean that specific considerations are not necessary in these fields. Basic solutions of the specific problems can, however, not be gained by techniques but need social processes, as in general the treatment of AI (subject in the other chapters).

A major field of concern with autonomous weapons appears to be the problem of instabilities and escalations initiated or reinforced by them. However, should it be considered as a solution that "weapons must be under meaningful human control" (p. 76)? Otherwise, if killing humans, to be considered as unethical? At least, these considerations may yield some restrictions to uncontrolled and unlimited development and use of autonomous weapon systems. This may be considered as the major impact. However, is there justification by human control (p. 79) with semi-autonomous systems?

Use of autonomous weapons may be justified in specific situations where rapid reaction is required and human intervention strongly endangers humans. However, in order not to run into general autonomous procedures, with the risk of running out of control, there is a basic need to weigh such military means compared to alternative ones and especially to social processes as solution perspective. This is in general valid and may rather yield necessary limitations to military means than general considerations on ethics specifically related to autonomous systems and human control. Again, the established culture of discussion appears to be most important.

Ad 7) Economical/Humanitarian Issues

The major issues here concern employment, equal distribution and worldwide perspective (developing nations).

The wide spread in predictions of job losses indicates that this is not just a question of objective prediction possibilities. It depends on how the development is designed by the society, by politics, it depends on social aims. Aiming at cooperative design of working conditions is in principle required by the increasing technological interconnections as social character of work and the shift to regulation and control. This should be used against the tendency in automation to reduce the amount of employees, even under capitalist driving forces to decrease employment. Such objective needs for better work must be realized under pushing via debates and conflicts. Further, the needs from the increasing interconnections in social processes, in design of society due to the technological implementations (possibilities as well as risks) should be picked up by institutions trying to reorganize social affairs and constitution. I.e., in the text, such necessities should be indicated more clearly instead of only general claims on good human-oriented development and considering seemingly objective trends to be further analyzed.

I agree that changes of traditional employment structures are to be considered, not only sheer number of jobs (p. 83). The objective basis for this is a general consideration of the shift towards controlling, regulating, optimizing activities concerning automatic systems instead of hand work, including a shift to preparing and in general intellectual work, work for designing whole processes, systems. From this, requirements and conflicting points have to be derived (see above). The overall needs of designing processes extend to future design of society. This has to be done based on available forces and interests to be activated. Lines in this direction should be indicated.

I agree that proactive actions are required, not only reactive ones (p. 83). This must include education of workers (candidate recommendation, p. 83), but not only requirements to them. A new design of work, cooperative work, is needed and has to be pushed. Participation in design processes of work and social affairs should finally be opened for everybody, if democratic aims are to be maintained (see the approach of F. Haug mentioned in my general comments). A division between those able to be trained for new work and those who cannot and thus need “fallback strategies” (which?) must be objected. Re-distribution of social wealth must also be considered to enable an adequate design of society with general participation.

The requirements on regulation as given on p.84 are to be supported and extended in this direction. Innovation in a good human sense is supported by activating cooperative behavior and structures, generally required by the technical development. Rules and laws given from above are necessary but not sufficient to generate the necessary processes to establish a cooperative culture. Innovations without destructive effects depend on this.

Concerning global responsibility (p.85), cooperative development approaches are most important, which can not only mean support in developing technical capabilities to close the gap in general (which may not be realistic), but, in the sense of sharing advances, joint development. This implies transfer needs, also financial, however in joint regulation for development. P. 85: "We need to ensure the equitable distribution of the benefits of AI/AS technology worldwide" or under 3.: "Promote distribution of knowledge and wealth …, including formal financial mechanisms (such as taxation or donations to effect such equity worldwide)".

Possible means for this have to be specified. Joint development programs considering the specific conditions and possibilities of countries are required to realize a sharing of wealth. This needs specific efforts for defining and initiating, not only to be considered as technical task. Training, education as well as global standardization/harmonization should also not only be considered from AI perspectives, as done especially in Section 3, but must address the specific needs and options of a developing country. Related questions are at least raised on p. 87 (bottom): "Do the economics of developing nations allow for AI/AS implementation? How can people without technical expertise maintain these systems?". But, a perspective is not given. Adapted technologies are to be considered (e.g. probably not driverless cars!?). In Section 2, the personal information issue (privacy and safety) dominates vs. the problem of global distribution whereas I consider the latter as more important.

Economic growth, supported by technical innovations and driven by competition, without considering goals of development and society, without considering joint processes and balances in the world, must be questioned in general. Increasingly, this type of development produces destruction of our natural basis, of human and social perspectives and is in basic contrast to ethical thinking and goals. An ethical argumentation just defending this by arguments of avoiding job losses, of yielding development chances for the poor by globalization etc., i.e. by the development itself, becomes increasingly unrealistic, misses ecological and human disasters produced and can finally not be justified ethically due to missing attempts to consider alternatives, in general considering the high productivity and capacities of the leading industrial nations. Thus, such alternatives under discussion as transformation to a post-growth economy or green economy should be considered and get a major concern in the frame of the initiative.

New Committees

In contrast to the previous chapters, the possibility to implement classical ethics in AI is clearly denied by the respective committee (p. 96), due to not possible transparency with learning systems, yielding to the statement: "The attempt to implant human morality and human emotion into AI is a misguided attempt to designing value-based systems". Only valid for "classical ethics"? Which other design of value-based systems is envisaged? What about the different view as compared to previous chapters? The conclusions are missing.

Questions of identity in mixed realities are raised. These problems cannot be decoupled from conditions under which interests, possibilities and perspectives to design real life are not developed, but rather continuously new fascinations are produced to distract from such goals and disperse interests by attracting them for changing consumption. Indeed, a mixed reality world driven by AI products cannot by itself become human, rather will deteriorate.

Specific issues raised by affective computing have again to be considered in the frame of general society development and control of AI In this context.

Effective Policymaking (EpicAI) sets as cornerstone "bringing together policy with the practical considerations of industry‟ (p. 101), thus indicating a practical perspective in line with my view and requirements. However, employees, i.e. trade unions, are at least not explicitly addressed as partners in the change processes, but should be, since they are most directly affected in their life and working conditions and also carry the potential in cooperative transformation to design new ways of working and social perspectives. Further, organizations considering post-growth alternatives should be included. There is a need to initiate broader social processes, beyond the good-will of stakeholders and wisdom of technical experts.

Dear all,

I am Dr. **Stephen Rainey**, a research fellow in the Centre for Computing and Social Responsibility in De Montfort University, Leicester. I work in the EU Flagship Human Brain Project. I'm a philosopher from Ireland, and I have made the following few comments on the EAD draft, which I enjoyed reading. I have tried to keep things brief, and to the point.

All the best,

Stephen

Executive summary, p. 2

Another Aristotelian concept related to eudamonai is 'phronesis', which is the practical enacting of virtue. This might be related to AI, etc., in the sense that we should be making systems with an emphasis on matching our different ways of enacting what we think of as virtues (moral, ethical values, for example). A focus on the practical might represent a good engineering perspective as an end result is posited toward which work ought to tend, rather than a set of principles from which something must be developed. Taken this way, there are risks, for example in terms of deceit. We want AI systems to cohere with our practical ways of being, but we don't want to intentionally build systems that are intended to trick us into treating them in one way or another (Thought Ishiguro's android work suggests we can spot imposters very well). Nevertheless, a concrete, practical set of expectations -- a phronesis -- for AI systems could be a worthy ideal.

P. 6 & p. 24ff

Methods to embed ethical reflection, and reflection upon that reflection (second-order reflexivity) can be found in the outputs from the ETICA project (http://www.etica-project.eu/). Other ideas aiming at the same theme can be seen in work by Sylvain Lavelle and Stephen Rainey on 'transforming proceduralism', into a value sensitive method

(https://www.researchgate.net/publication/293305872_Transformation_of_proceduralism_from_contextual_to_comprehensive)

This is a way of trying to find how interactions of values, contexts, and intentional attitudes shape a conceptual space in which technologies appear, hence for understanding the meaning of the technology for those around it.

p. 16 Candidate recommendations, 2 & p. 22 Section 2

If we treat AIS as interlocutors in a generalised, hypothetical, public discourse, we can generate focal points to enable ethical reflection on their nature and use. This is because as objects, AIS nevertheless enter into relations with established social realities, and the individuals who partake in them. Depending on different perspectives taken up, the AIS might stand as a challenge; a question; an aid; a new opportunity, for the individuals or social systems they appear in. As a method, this approach can yield a set of issues worth pursuing by engineers and developers that are highly relevant to the individuals and social systems the AIS appear in, by treating them as contextualised, and relative to the users (broadly construed).

p.23, on "transparent signals"

Whether or not a machine can be said to *really* be 'explaining itself' (http://news.mit.edu/2016/making-computers-explain-themselves-machine-learning-1028) the idea that a system would have a straightforward reporting mechanism that would describe the rationale for the actions it has taken/is going to take would be useful in understanding the system as such. Moreover, it would facilitate reactions to the system in giving a report to which reaction can be directed. Where multiple rationales are available to the system, it would also be of interest which ones are chosen, and why (this could aid future refining and development of the system, as well as understanding more its perception within a social setting).

p.26, recommendation on stakeholder involvement

This is very useful, and very complex. Including stakeholders in value-sensitive ways requires a lot of procedural sophistication. Some approaches are analysed and built upon in Lavelle, S. and Rainey, S., 2013. Transformation of Proceduralism from Contextual to Comprehensive. *Ethical Governance of Emerging Technologies Development*, pp.312-343.

p.47-48, candidate recommendations

These are very good, and timely, especially in a context of growing confidence in big data analytics among researchers. Approaching 'black box' techniques with a critical, but not necessarily cynical, frame of mind will be of tremendous value in the future.

Comments about Ethically Aligned Design - Version 1

Name: Reji M. Issac

Affiliation: B. P. C. College, Piravom, Ernakulam, Kerala, India, PIN - 686664

Thank you for taking interest in Ethically Aligned Design for nurturing great technologists of the future towards right direction. I highly appreciate for initiating such an attempt to formally develop an IEEE document towards a new generation of technologists who can overcome great challenges in scientific and technological world which is inclusive of Artificial Intelligence and Autonomous Systems which points towards problems like Technological Singularity. I am a person who registered for "The IEEE AI & ETHICS SUMMIT 2016: Artificial Intelligence and Ethics – Who does the thinking?" at Brussels, Belgium who could not attend the said meeting on 15 November 2016. I am also a senior member of IEEE. I would like to submit the following comments with regard to "The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems" version 1 for public discussion.

- First I introduce a microcosm symbol (representing Logos which consist of the word which is inherent in Technology showing the structure of knowledge) for the consideration of "The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems" which consists of the core values that we need to make effective in the technological world for the wellbeing of humanity, as published in the paper Geometric and Physical Modeling of Natural Intelligence . This explains why we are in need of Ethically Aligned Design in AI/AS. Being Cybernetics is based on the word, which is the source of life, and provides leadership for all other subjects, we can also include 'Ethical Considerations' as part of Cybernetics.

- With regard to The Mission of The IEEE Global Initiative, which states as "To ensure every technologist is educated, trained, and empowered to prioritize ethical considerations in the design and development of autonomous and intelligent systems", I must comment that it may be modified with the following statements as "To ensure every technologist is nurtured, educated, trained, and empowered to prioritize ethical considerations in the design and development of autonomous and intelligent systems".

- After General Principles (1) section in pages 15-21, we must add a new section as Section 2 with title "Embedding Empowerment into Autonomous Intelligent Systems", which describes in detail about what kind of empowerment an Autonomous Intelligent System should possess and which must be embedded into AIS, which forms as part of Ethical Considerations. This empowerment must be connected to the *Wisdom* point in the structure of the Word. Cybernetics can take a leadership role with respect to empowerment.

Further Resources

1. Wikipedia https://en.wikipedia.org/wiki/Empowerment

2. U.N. General Assembly, 55th Session. "United Nations Millennium Declaration." (A/55/L.2). 8 September 2000. (Online) Available: www.un.org/millennium/declaration/ares552e.pdf (accessed January 2, 2008)

3. Zimmerman, M.A. (2000). Empowerment Theory: Psychological, Organizational and Community Levels of Analysis. "Handbook of Community Psychology,"

4. Rappaport, Julian. In praise of paradox. A social policy of empowerment over prevention, in: American Journal of Community Psychology, Vol. 9 (1), 1981

5. Wilkinson, A. 1998. Empowerment: theory and practice. Personnel Review. [online]. Vol. 27, No. 1 http://dx.doi.org/10.1108/00483489810368549

- Page 37-38. Issue: Ethics is not part of degree programs. (Section 3. Methodologies to Guide Ethical Research and Design). This issue can be solved by including Cybernetics in degree programs, which provides a concentrated and converged structure of values with Logos which acts as coherent source of words, which acts as a guide for Ethical Research and Design. The problems we discuss in these pages can be solved by adopting the structure of the word as published in Geometric and Physical Modeling of Natural Intelligence solving the issues described in pages 37-38. Cross-pollination between disciplines can be achieved through this model of the word, which is the basic constituent of Cybernetics on which Cybernetics works with. Human values transcend all academic areas of focus through this structure. The presented symbol is embodying the highest ideals of human rights and transparency of the law. Through the structure presented here I am trying to bring the tacit knowledge about ethics into an explicit knowledge about ethics bringing transparency. This picture is giving a delineation of the ethical standard of the basic autonomous system of the world which is the word.This model also can be considered for interdisciplinary and intercultural education to account for the distinct issues of AI/AS.

Further Resources

1. Reji M. Issac, "Communication and Control through Words and Power", Proceedings of IEEE International Conference on Control, Robotics and Cybernetics (ICCRC 2011), March 21-23, 2011, New Delhi, India, Vol.2, pp. 414-421, ISBN: 978-1-4244-9709-6 (Print), IEEE Catalog Number: CFP1176M-PRT (Print), ISBN: 978-1-4244-9711-9 (electronic - CD), IEEE Catalog Number:CFP1176M-ART (electronic - CD),©2011 IEEE

2. Reji M. Issac, "Communication and Control through Words and Power", Advanced Materials Research (AMR – Volumes 403-408) MEMS, NANO and Smart Systems, pages 982-993, doi:10.4028/www.scientific.net/AMR.403-408.982, ©2012, Trans Tech Publications, Switzerland, ISBN: 978-3-03785-312-2. ISSN(Web) :1662-8985 ISSN (Print): 1022-6680 ISBN (CD): 1022-668

3. Reji M. Issac, "Geometric and Physical Modeling of Natural Intelligence", Bonfring International Journal of Man Machine Interface Volume 4, Issue 1, April 2016, pp.01-06, Bonfring Online ISSN: 2277-5064 | Print ISSN: 2250-1061 DOI:10.9756/BIJMMI.8137

Submitted by **Marijn van der Pas**, campaigner at FullAI

Feedback by FullAI on the document 'Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems':

- As an extra bullet on page 15 FullAI would like to suggest including this sentence: Hold an individual human being or a recognized legal entity accountable for automated AI decisions that directly influence a human life.
- On page 18 and also page 19 please also refer to the 'Ethics and Values' section of the Asilomar AI Principles: https://futureoflife.org/ai-principles/
- As a fifth principle (after page 21) we would like to suggest including the principle of what Google calls the Big Red Button by adding this sentence: In which concrete way can we control and contain artificial intelligence? (Please also see: http://www.fullai.org/european-parliamentary-committee-asks-ai-kill-switch)
- On page 24 please add a reference to the Theory of Basic Human values developed by Shalom H. Schwartz (https://en.wikipedia.org/wiki/Theory_of_Basic_Human_Values and http://valuesandframes.org/wp-content/uploads/2011/04/schwartz_circumplex.png)

Prof. Dr. Oliver Bendel
School of Business, University of Applied Sciences and Arts Northwestern Switzerland FHNW

Abstract

My most important findings are: 1. The document is too anthropocentric in most parts. KI and robotics are relevant for animals too, and there are research fields like animal-computer interaction and animal-machine interaction that should be taken into account. 2. The discipline of machine ethics is a young but very dynamic discipline, not only in the U.S. but also in Europe. It should have a greater significance and more space in the document. Besides machine ethics, there are several fields of applied ethics that are relevant here, e.g., information ethics, technology ethics and business ethics. The document should mention these technical terms and should use the specific concepts and methods of these specific ethics. 3. The document is too U.S.-oriented. I miss the articles and books of some important researchers in Europe. Since 2012, there have been several important publications in the field of machine ethics in Europe, and various conferences and workshops have taken place.

Comments

The document mentions several times (e.g., on page 2) that the technologies in question are aligned to humans and to human welfare. However, animals and animal well-being and welfare are also important and subject of disciplines and research fields like animal-computer interaction and animal-machine interaction (Mancini 2011; Bendel 2015). Furthermore, some articles try to combine machine and animal ethics (Bendel 2016e).

On page 2, the document mentions “our moral values and ethical principles”. But which moral values and which ethical principles? The values and principles of the members of the IEEE Global Initiative, of programmers and developers or users, of groups or individuals, of western or eastern societies and cultures, within a framework of duty ethics or consequentialist ethics?

On page 6, you write: “Ethics is not part of degree programs.” At Swiss universities, there are several offers in the field of ethics. To give an example: At the University of Applied Sciences and Arts Northwestern Switzerland FHNW (http://www.fhnw.ch) we teach information ethics, technology ethics and business ethics, and some of the courses are mandatory. Information ethics dedicates itself to students of business information systems. Also in Germany, ethics is often part of technical curricula.

The topic on page 10 is law and robotics. Crucial in this context are questions of liability. Another issue is whether we should equate artificial moral agents and natural moral agents (humans). In the German-speaking countries, there are several experts for robot law (Würzburg, Basel). Some of them suggest an “electronic person”, similar to a legal person. The idea is that such a person will assume del credere liability.

On page 15, you write: “Prioritize the maximum benefit to humanity and the natural environment.” The term “natural environment” is much too wide for me. You should focus more on moral patients, both on humans and animals. Of course, plants are also important, as are biospheres and biodiversity. “Designers and developers of autonomous and intelligent systems should remain aware of, and take into account when [!] relevant, the diversity of existing cultural norms among the groups of users of these AI/AS.” (p. 18) But what if these cultural norms contradict human rights? Not all cultural norms are good, and some of them should not be respected.

On page 19, you write: “Thus, lack of transparency both increases the risk and magnitude of harm (users not understanding the systems they are using) and also increases the difficulty of ensuring accountability.” I would like to add the following: Systems can generate transparency about their functions and sources (e.g., Google Now or Google Assistant declares very often: “Wikipedia says ...”). Furthermore, managers and developers can achieve transparency about their systems.

Education and awareness are the issues on page 21. Point 1: “Providing ethics education and security awareness that sensitizes society to the potential risks of misuse of AI/AS.” Perhaps you could be more concrete: Education on information ethics, technology ethics and machine ethics. Point 2: “Delivering this education in new ways, beginning with those having the greatest impact that also minimize generalized (e.g., non-productive) fear about AI/AS (e.g., via accessible science communication on social media such as Facebook or YouTube).” However, the methods and the behavior of Facebook, Google and Co. contradict personal rights and human rights in many cases. In Europe, some of their activities and services are illegal or under observation. In my opinion, we need legal and neutral platforms.

The chapter "Embedding Norms and Values in Autonomous Intelligent Systems" (p. 29 ff.) focusses on research in the U.S. However, there is also research in Europe by Luís Moniz Pereira (Pereira 2016) and Oliver Bendel (Bendel 2016a-d, Bendel 2015). There will be a "Handbuch Maschinenethik" ("Handbook Machine Ethics") in German and English at the end of 2018 (Bendel 2018). There are several current books on information ethics and machine ethics (Bendel 2016c; Bendel 2016d; Trappl 2015; Rötzer 2016). In general, I would clearly distinguish between "machine ethics" (the discipline or the research field) and "machine morality" (the morality of the machine itself). Machine morality is not necessarily associated with machine learning. Most of the existing prototypes and simulations have nothing to do with machine learning and deep learning. The machines just follow simple rules.

On page 31, you write: "If a community's systems [!] of norms (and their underlying values) has been identified, and if this process has successfully guided the implementation of norms in AIS, then the third step in value embedding must take place: rigorous testing and evaluation of the resulting human-machine interactions regarding these norms." In my opinion, not only the community counts, but also the individual. The rights of our minorities and single persons must be safeguarded too.

On page 38, you write: "We also recommend establishing an intercultural and interdisciplinary curriculum that is informed by ethicists, scientists, philosophers, psychologists, engineers and subject matter experts …" I would like to add the following: Ethicists are philosophers. If not, they are theologians – but because theologians are not scientists that should not matter in this context.

In relation to "Methodologies to Guide Ethical Research and Design": In the different cultures, there are different individuals with different opinions. Most of my students do not understand the rights to their own image any longer; I do, and these rights are important for me. Furthermore, scientific ethics has emerged from western philosophy. This is very important in this context. In philosophical ethics, we use, e.g., logical, dialectic and discursive methods. And finally, I would clearly distinguish between the disciplines and approaches on the one hand and the subjects and topics on the other hand. Of course, religion can be considered, but not as a method.

References

Anderson, Michael; Anderson, Susan Leigh (eds.). Machine Ethics. Cambridge University Press, Cambridge 2011.

Bendel, Oliver (ed.). Handbuch Maschinenethik. Springer, Wiesbaden 2018. (Will be published in 2018.)

Bendel, Oliver. Überlegungen zu moralischen und unmoralischen Maschinen: Neben die Moralphilosophie als Menschenethik ist im 21. Jahrhundert die Maschinenethik getreten. In: Rötzer, Florian (ed.): Programmierte Ethik: Brauchen Roboter Regeln oder Moral? Heise Medien, Hannover 2016a.

Bendel, Oliver. Annotated Decision Trees for Simple Moral Machines. In: The 2016 AAAI Spring Symposium Series. AAAI Press, Palo Alto 2016b. pp. 195 – 201.

Bendel, Oliver. Die Moral in der Maschine: Beiträge zu Roboter- und Maschinenethik. Heise Medien, Hannover 2016c.

Bendel, Oliver. 300 Keywords Informationsethik: Grundwissen aus Computer-, Netz- und Neue-Medien-Ethik sowie Maschinenethik. Springer Gabler, Wiesbaden 2016d.

Bendel, Oliver. Considerations about the relationship between animal and machine ethics. In: AI & SOCIETY, 31 (1), 2016e. pp. 103 – 108.

Bendel, Oliver. Überlegungen zur Disziplin der Tier-Maschine-Interaktion. In: gbs-schweiz.org, 14. Februar 2015. Via http://gbs-schweiz.org/blog/ueberlegungen-zur-disziplin-der-tier-maschine-interaktion/.

Bendel, Oliver. Maschinenethik. Contribution for the Gabler Wirtschaftslexikon. Springer Gabler, Wiesbaden 2012. Via http://wirtschaftslexikon.gabler.de/Definition/maschinenethik.html.

Lin, Patrick; Abney, Keith; Bekey, George A. (eds.). Robot Ethics: The Ethical and Social Implications of Robotics. The Mit Press, Cambridge, MA 2012.

Mancini, Clara. Animal-Computer Interaction (ACI): A Manifesto. Interactions, 18 (4) 2011. pp. 69 – 73.

Pereira, Luís Moniz; Saptawijaya, Ari. Programming Machine Ethics. Springer International Publishing Switzerland, Cham 2016.

Rötzer, Florian (ed.). Programmierte Ethik: Brauchen Roboter Regeln oder Moral? Heise Medien, Hannover 2016.

Trappl, Robert (ed.). A Construction Manual for Robots' Ethical Systems: Requirements, Methods, Implementations. Springer, Berlin and New York 2015.

Wallach, Wendell; Allen, Colin. Moral Machines: Teaching Robots Right from Wrong. Oxford University Press, New York 2009.

Dear IEEE,

I scanned the IEEE Ethically Aligned Design document. I fully support the idea behind this, but I think the approach as outlined is a bit misguided.

Morality being defined as cultural norms is relatively accurate; though not completely. Ethics is universal and transcends culture and time. I wrote a book about the distinction between morals and ethics, see www.EthicsDefined.org. A more detailed description of the differences between morality and ethics can be found here: http://www.ethicsdefined.org/what-is-ethics/morals-vs-ethics/ and here: http://www.ethicsdefined.org/what-is-ethics/moral-and-ethical-sphere-of-influence/

Trying to program AI to mirror specific human cultural norms is a fool's errand, and would likely result in unethical behavior and actions. Cultural norms are inherently ethically relative, making for a poor universal foundation.

Modeling AI to derive ethical truths ought to be relatively easy and pretty straight forward. Because AI would likely lack human vices, it should have the ability to be more accurate at deriving ethical truth.

The foundational axiom of ethics is "**I feel, therefore I know ethics**."

Being able to feel sensory and emotional input, and understand that others can feel the same input in similar ways is the basic requirement for ethical knowledge. Knowledge is know-how. If the AI could "feel" physical and emotion input (pleasure, pain, etc.), and understand others could feel the same (empathy), then it would have the basic knowledge require to then be able to derive ethical truth.

Understanding is know-why. Asking [thee] basic question "How would I feel if that was done to me?" and framing this question in terms of harm/care and fairness/reciprocity provides the foundation to derive the ethical truth in any situation. It can be used to question/challenge authority, validate whether it is ethically valid to follow group norms, what is fair treatment, what is harmful or helpful, whether just retribution for an unethical act is being followed, etc. To be ethically minded it to have an independent mind capable of deriving ethical truth.

The difficulting with AI is not running the above algorithm [the Golden Rule focused on matters of harm/care and fairness/reciprocity], rather, it's getting the AI to understand physical and emotional pain, to understand fairness, etc. Some of this could be hard wired (physical sensors programmed to understand human pain thresholds), facial recognization with microexpression algorithms to detect emotional states (fear, surprise, pleasure, hate, anger, resentment, etc.) which could act as *empathy sensors*. Aspects of fairness could be hard wired (tit-for-tat algorithms) and some learned (fair play through interactions with others).

The algorithm for deriving the correct ethical path is pretty simple, the difficulting is decoding the elusive obvious manner in which we humans do it . . . which I did :-)

The benefit of a simple set of rules, that allow for the derivation of ethical truth for any set of circumstances, and continual learning ought to be obvious. Now all you guys need to figure out how to program it effectively.

Regards,

Lawrence Sheraton
Website: EthicsDefined.org
Sculpture: LawrenceSheraton.com

(1) Name and Affiliation of individual submitting this comment
Name: Shinichi NOMOTO
Affiliation: KDDI Research, Inc.
E-mail: nomoto@ieee.org

S. Nomoto, Dr.
Executive Principal Researcher
KDDI Research, Inc.
nomoto@kddi-research.jp

(2) Page number
Page 16. (Principle 1 – Human Benefit. Issue: How can we ensure that AI/AS do not infringe human rights?)

(3) Comment [Sh-Nomoto_EADv1-01]
First, on Page 16, AI/AS is regarded as a counterpart to be under the control of Human which is assumed perfect. As the history has shown, Human is by no means flawless and AI/AS would be relatively flawless but yet may have some bugs. Human should rather coevolve with AI/AS. We (Human and AI/AS) should learn from each other through various interactions, both positive and negative, for better and prosperous future, while avoiding any major, critical and catastrophic accidents and conflicts. In this context, some level of interference (negative interaction) between Human and AI/AS should be allowed in the earlier phase, or even encouraged during the development or proof-experiment phase. (Note that, in EAD v1, there is no use of the words "interfere" and "interference" at all.)

Second, AI/AS will be embedded in a system which as a whole will do some missions or tasks. The risk management can usually be done through total system design (e.g. redundancy, functional safety (fail-safe) design). It is not clear what AI/AS precisely means in EAD v1.

Therefore, I'd recommend the first "Issue" for "General Principles" on Page 16 as follows.

- Issue: How can we best balance long-term benefits versus short-term risks of AI/AS deployment so that Human will coevolve together with AI/AS.

For the above new Issue, I'd recommend the following "Candidate Recommendation."

- Candidate Recommendation: Thorough identification and consideration are needed on what are the fundamental differences between AI/AS and other modern technologies as well as their implications so that we can go forward to effectively develop new frameworks and/or methodologies for the implementation and deployment of AI/AS.

- Candidate Recommendation: An interdisciplinary and stake holder-inclusive approach of total system design, incl. non-technical aspect, should be pursued which will allow us pragmatically implement and deploy a new system augmented by AI/AS for Human benefit.

(2) Page number
Pages 22-23.

(3) Comment [Sh-Nomoto_EADv1-02]
The description in the paragraph on Page 23 is very important. The process of "embedding values into AIS" would be and must be an iterative process. It is anticipated that, during the course of implementation, not only positive but also negative interaction (interference which may provoke trust) would happen. Human, and the society, should be patient, tolerant, and gentle enough to help new AIS be born and grow healthily. Human is a parent of AIS and should totally be responsible to the whole life of AIS. A parent will also grow by fostering and loving children.

I'd recommend adding the following paragraph at the end of Page 23.

Since the value of AI/AS varies and may be brand new, embedding new values with new AI/AS would result in difficulty that we couldn't rely on our experience and the history. The embedding process must be continuously iterative by having a "Kaizen (constant improvement in Japanese)" management scheme and, in order to get (long-term) benefit, all stakeholders should be patient, tolerant, and gentle to help new AI/AS be born and grow healthily (or let it properly extinct if necessary). Human is a parent of AI/AS and should be well prepared and totally responsible to the whole life of AI/AS.

Here, the word "AI/AS" is used instead of "AIS" for the consistency throughout the EAD v1 document.

(1) Name and Affiliation of individual submitting this comment

Name: Shinichi NOMOTO

Affiliation: KDDI Research, Inc.

E-mail: nomoto@ieee.org

(2) Page number

Pages 22-23.

(1) Name and Affiliation of individual submitting this comment

Name: Shinichi NOMOTO

Affiliation: KDDI Research, Inc.

E-mail: nomoto@ieee.org

(2) Page number

Pages 32-33. (Issue: Achieving a correct level of trust between humans and AIS

(3) Comment [Sh-Nomoto_EADv1-03]

As stated in the last sentence before "Candidate Recommendation" on Page 32, trust building should take into account Human natures. It would take time and whether the new system is acceptable by the general public is often unpredictable. This is because Human is not rational. In case of dynamic pricing, for example, even the user is fully provided with logics behind the system (accountability and transparency are sufficient), he/she may not be happy to accept the system if it is revealed one day that the higher price is always produced than his/her neighbors of similar lifestyle. Therefore, accountability, transparency and verifiability are necessary but not sufficient for acceptance by the general public. Irrationality of Human must be taken into account before deploying the new system widely in society. Acceptability by the general public should be proved by the properly designed social experiments which will take time.

I'd recommend adding the following candidate recommendation.

Candidate Recommendation: Accountability, transparency and verifiability may not be sufficient for trust building, because Human is irrational by nature. Acceptability by the general public should be proved by the societal experiments properly designed by stakeholders before wide implementation. Technologist need to be willing to take time to prove the acceptance.

Another aspect that is worth considering is "over-trust" because the issue raised here is "achieving a 'correct' level of trust." "Over-trust" or over-dependency is already happening that people believe in non-existence of something on the earth if no query answer is hit by Google search, for example.

(2) Page number

Pages 49-55.

(3) Comment [Sh-Nomoto_EADv1-04]

Chapter 4 (pp. 49-55) have a number of major problems some of which are identified below:

(a) The word "beneficence" only appears in Chapter title and Committee's name. Why?

(b) It is not clear what AGI/ASI exactly means and its difference from AI/AS that is commonly used in EAD v1. Superintelligence, in general, is defined as "intelligence after the so-called singularity." If this understanding is correct, the issues identified in Chapter 4 do not apply specifically to AGI/ASI entitled. Human has already utilized a number of complex systems with critical mission (e.g. jet plane, spaceship, nuclear power plant, regenerative medicine). In the first paragraph of Chapter 4 should describe AGI/ASI definition and identify the difference between the modern technologies already implemented with "safe/beneficence by design" and AGI/ASI in future.

(c) The authors/editors of Chapter 4 seem to have only AI research teams (or AI community) in their mind. Stakeholders, or at least "technologists" as defined on Page 4, must be taken into account and Candidate Recommendation should cover them.

(d) A number of books and papers are referred. Most of them are related only to Machine Learning which is just one of AI techniques and not representing methodologies of system design with AGS/ASI elements. Accordingly, the discussions there are too biased for the Machine Learning community. At least, "Recommendation" may not include any statement that recommends seeing some specific books/papers (e.g. "See xx"), because authors/editors are ethically prohibited trying to increase citation index or sales of someone’s research work by exploiting "Recommendation" in EAD document.

(e) There are some statements referring to "the operator" on Page 51. I believe that there will be no operator in autonomous system by definition. The supervisor or administrator may exist but their task and responsibility is vague in AGI/ASI context. Similarly, whether we can rely on "review board" that is frequently mentioned on Page 53 is questionable because review board seems to reside off-line and be reactive with latency.

(f) (Editorial comment) "Technical (Section 1)" comes first followed by "General Principles (Section 2)." It is not recommended if there is no fundamental reason.

(g) (Editorial comment) Most of draft "Candidate Recommendations" describe "what to do" (e.g. "contribute," "work to," "ensure") rather than showing "goals" as stated on Page 50.

I'd recommend revising the title and "Issues" of Chapter 4 as follows.

- Title of Chapter 4: Coevolution of Human and AGI/ASI
- Issue: Future AI/AS (incl. AGI/ASI) would become too complex for technologists to properly design, verify, and modify by using their capability with best practices over the history. Therefore, technologists use the developed AI/AS in developing new AI/AS which increases the complexity. This iterative cycle shall result in explosion of complexity implying that Human, even using collective and accumulated wisdom on earth, cannot supervise the integrity of whole system. Is Human prepared to succeed the role of supervisor of the society to AGI/ASI? If not, how and when?
- Issue: Autonomous and superintelligent systems, especially when applied to general purpose, would make a decision which implies interference with Human. The ethical discussion on culpability, responsibility, liability, etc. of AGI/ASI is yet to be clarified. How should our society become more robust and tolerant to the interference and friction caused by AGI/ASI so that Human can learn from those experiences for coevolution with AI/AS?
- Issue: As the advancement of technology is accelerating exponentially, the timescale of AI/AS has begun to divert those of Human. What kind of mechanism (not only technological but also societal one) do we need to fulfill the gap and achieve the best balance of long-term benefits versus short-term risks.

Thus, the contents of Chapter 4 would need major revision accordingly.

(2) Page number

Page 84.

(3) Comment [Sh-Nomoto_EADv1-05]

On Page 84, regarding the "Issue: AI policy may slow innovation," the "Candidate Recommendation" is on how to accelerate legislation and policy regarding AI. It can be read that Human shall adjust AI/AS in order not to hinder but to "keep up with the rapid advancement of technology." Shall we really recommend stakeholders in this direction? I wouldn't deny the proposed approach in "Candidate Recommendation" on Page 84, but the fundamental difficulty here is that we will not be able to do so, because the gap between the timescale of Human and that of AI/AS is constantly increasing exponentially. Here, "timescale" includes not only clock speed of brains/AI but also generation cycle of Human/robot, growth rate within a generation of Human/robot, adaptation speed to changes of real/cyber world, and so forth.

I think that Human and AI/AS should coevolve by adjusting optimum balance between technology innovation and societal innovation. However, the priority should be given always to Human rather than AI/AS. If Human, even with collective and accumulated intelligent, cannot keep up with the advancement of technology, we must intentionally slow the innovation (or societal implementation) by putting higher priority to the acceptance by general public (consensus based on common sense).

Therefore, I'd suggest revising "Issue" and "Candidate Recommendation" on Page 84 as follows:

- Issue: The speed of Human and societal mechanism (e.g. legislation, policy making) will not be able to keep up with the rapid advancement of AI/AS technology sooner or later (or partially already has not).
- Candidate Recommendation: Stakeholders should collaborate in minimize the gap between the timescale between Human evolution and AI/AS evolution by using collective and accumulated intelligence. Interdisciplinary discussion and collaboration would help society be prepared for unforeseen risks with some proactive schemes. However, if Human, even with collective and accumulated intelligent, cannot keep up with the advancement of technology, intentional slowing of the innovation (i.e. societal implementation and deployment) should be decided from the viewpoint of "Human First."

Artificial Intelligence

Rogelio Piña Vega Autonomous University of Querétaro

I am a student of the 5th semester of Biomedical Engineering of the Autonomous University of Querétaro (UAQ) in Querétaro, Mexico. As part of our current load of subjects, we are taught the subject of Bioethics which aims to promote the inclusion of social and humanistic sciences within our training as engineers and in this way include them in our research, prototypes, and projects that we will develop later.

The UAQ, as part of its postgraduate courses, offers the Master of Science in Artificial Intelligence. Interest in writing about Artificial Intelligence (AI) began with a talk with the coordinator about the IEEE's initiative to create a Bioethics Committee for AI. After this talk comes up in me this interest about how we can help the formation of this committee and what my contributions could be for this.

The subject that most drew my attention from the document Ethically Aligned Design: A Vision for Prioritizing Wellbeing with Artificial Intelligence and Autonomous Systems was the fourth theme, Safety and Beneficence of Artificial General Intelligence (AGI) and Artificial Superintelligence (ASI)1.

As mentioned in the document on page 51, to retrofit security to an AI system in the future will be more difficult, because if we start to create machines with IA without any ethical or bioethical guidelines it could have quite large repercussions in the future if the machine is capable of being upgraded to a more intelligent system, it will be more difficult to correct the errors that it could present.

Tomasik2 tells us in his text that with the free advances of these machines we will reach a point where we will not be able to coexist in the same world and will unleash a war between machine with AI and humans. If we as a society do not care about the development of AI, we are letting a small portion of our society grant autonomy to machines with Artificial Intelligence and ultimately those machines will be able to replace us after a while because they will be better than us.

I consider that due to the exponential growth of the technologies and systems of Artificial Intelligence we must specify certain regulations for these emerging technologies of our days creating codes of bioethics that limit the actions of these machines. Once these regulations are stablished, we could reach larger scales with the bioethical codes to also regulate the process of design and manufacture of intelligent machine systems.

Nowadays, we do not know what the world will look like in a few years due to Artificial Intelligence without bioethical guidance and growth without any control in its research, but we could know what the future of our planet with these AI systems would be if we include a bioethical debate on the progress of AI.

[1] The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems. Ethically Aligned Design: A Vision For Prioritizing Wellbeing With Artificial Intelligence And Autonomous Systems, Version 1. IEEE, 2016. http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html [2] Tomasik B. (2016). *Artificial Intelligence and Its Implications for Future Suffering.* (pp. 9) Foundational Research Institute. Available at: https://foundational-research.org/files/artificial-intelligence-and-its-implications-for-future-suffering.pdf

While it would be a total mistake to stop the advance of technology and science, we cannot let this advance grow in a free way but we must regulate it to avoid future conflicts with these developing technologies, a very clear example of this is Artificial Intelligence, that is why I mention bioethics. Therefore, this science is not responsible for prohibiting behaviors that could cause conflict within our society, but rather ensures that all these behaviors have a well-defined objective and thus be able to achieve the welfare of both individual and society. So if we get bioethics into AI guidelines we can establish these codes of ethics in the machines and thus seek the welfare of mankind not to generate a future war between humans and intelligent machines. We should not give so much autonomy to the machines because we must be clear that all these technological developments and machines are tools that help us improve our quality of life, lessen our suffering or preserve life as Jonsen3 would tell us in his text.

As a Biomedical Engineering student and because of the nature of the degree, I could enter in the field of Artificial Intelligence, so I think it is not a subject foreign to me and should not be foreign to any of us because of the serious implications that could have an uncontrolled development of this technology in the future not very far to our days.

Bioethics is a science that allows us to reflect on our actions to act correctly and maintain a social balance in the environment in which we develop. That is why I consider that we are in time to relate the topic of bioethics to AI research in order to maintain this balance on our planet and to be able to coexist humans, biologic beings, and machines.

Generating a consciousness within society, that would be the main strategy that we could have in the short term to have the best benefit of AI in the field of industry and avoid future problems. With the dissemination of information, this social awareness could be generated and different professionals could join a debate to address issues such as AI techniques that are currently being developed, the social impact they could have on testing in the real world, social security, the legal responsibility of companies developing such technologies and machines, etc.

As we know, AI is developed continuously and with new projects, then with the promotion of bioethics in their areas of research we would be ensuring that all these new projects would be developed under a well-established protocol with bioethical guidelines to model the Systems of Artificial Intelligence. The fact of generating awareness within our society in the field of Artificial Intelligence would imply a great advance to promote bioethics in this technology.

[3] Jonsen, A. (2006). *A History of Religion and Bioethics. In D. Guinn, Handbook of Bioethics and Religion* (1st ed., pp. 26). Nueva York: Oxford University Press. Available at: https://books.google.com.mx/books?id=WClvsJ03vWgC&lpg=PP1&hl=es&pg=PA26 - v=onepage&q&f=false

Dear Sir or Madam,

I am hereby submitting my comments on regarding the version 1 of Ethically Aligned Design.

Name: Hyo-eun Kim (hyoekim26@hanbat.ac.kr; qualia9@gmail.com)

Affiliation: Hanbat National University, College of Liberal Arts

Area: Philosophy of Science, AI Ethics.

https://hanbat.academia.edu/HyoeunKim

https://newclass.hanbat.ac.kr/ctnt/liberal/prof.php?mno=02.01

--Comments on EAD

1) On page 32, I recommend adding the following two kinds of issues with regard to transparency:

Two technical and philosophical issues with transparency can be raised: the one is a dilemma with regard to transparency; another is a fundamental difficulty.

One way to resolve the problem of transparency is to prove the moral decision-making of artificial intelligence ("Rationalizing neural Predictions" (2016) Lei, T., Barzilay, R., & Jaakkola, T. (2016). *arXiv preprint arXiv:1606.04155*.). Even if morality-proofing artificial intelligence plays a role in making AI transparent, however, a dilemma may arise due to the inherent nature of artificial neural network: The higher level of the efficiency we seek in AI algorithm, the less the transparency of the transparency we find. Specifically artificial neural network used in machine-learning AI (compared to the symbolic processing model used in logic-based AI) would not be transparent enough to reveal the process of how the autonomous system made a decision. Yet, using only logic-based AI would not be realistic for developers. Then, the questions is how we ensure both smarter AI and transparency.

Second, explaining ‘how’ the decision-making of AI ‘was’ reached by its algorithm might be fundamentally impossible. The morality-proofing artificial intelligence might show only the *post facto* justification of the decision that the AI already made; however, it could not reflect the context of the discovery or the process of decision-making of autonomous systems.

(For the distinction between the context of discovery and justification, please see Kuhn, Thomas S. (1970), The Structure of Scientific Revolutions, Chicago: University of Chicago Press.; Popper, Karl R. (1959 [1934]), The Logic of Scientific Discovery, London: Hutchinson.)

2) On page 15, I recommend adding to the “Prioritize the maximum benefit to humanity and the natural environment”:

However, the preference of human benefit and natural environment might be ambiguous in real situations. Which one would AI choose between humanity and the natural environment when the latter goes bad? Most of all, which one would AI choose between two kinds of moral values within a society? And, would it be there only one value of humanity? Prioritizing the benefit to humanty is natural and important but it requires an answer to the questions how to decide which value is preferable in a specific situation.

Sincerely,

Hyo-eun Kim

Francisco López Caracheo.

Universidad Autónoma de Querétaro.

In this work, I am going to expose my point of view about Safety and Beneficence of Artificial General Intelligence (AGI) and Artificial Superintelligence (ASI) of the Ethically Aligned Design document.

In the page fifty as a part of the Background, we can see that capable AI systems are likely by default to adopt "convergent instrumental subgoals". In the abstract of this topic we see that "As AI systems become more capable, unanticipated or unintended behavior becomes increasingly dangerous".

In my opinion, the values that are programmed to each AI system determine the subgoals that can develop these. Some values, or standards that systems should follow, are contradictory to each other. Because of this, it is very important to determine which are the values that manage to regulate AI systems in a better way. A robust value structure should prevent these subgoals from being developed.

In this chapter I would add what values should be implemented depending on the conditions where the AI system is positioned, conditions that will vary due to society, economy, discrimination and even the religion of the environment where it is. The framework for the regulation of an AI system might be, as I said, different in each environment. The creation of many systems of regulation, each specific to each society, is a rather complicated and long work. It is worth analyzing what conditions have an impact on an AI regulation system. Once this question is clarified, we could define if a regularity system would be suitable using universal values (the same values in the regularity system for all societies) or a different regulation system would have to be created for IA systems in each society.

Later, I would add this: According to these both ideas (the ideas of the second paragraph) we could notice that as a system become more capable, it's become dangerous and unpredictable. A consequence of its capacity could be a growth in its autonomy. This new capacity and autonomy could be used to reorganize the old one's systems.

In order to propose an increasingly robust and successful model, I think it would be good for the education of new readers to include, in the next book, more opinions from different authors (even if them opinions are merely philosophical and not philosophical-technical) as background. There is an example: What would happen, for example, if an AI system is programmed not to harm people and, in any particular case, a human being is found between extreme suffering or induced death? Some authors like Fletcher [1] reject the extreme suffering and propose the death in these particular cases, using casuistry to determinate the right choice. Others, like McCormick [1], are faithful followers of the Natural Law, who seeks at any cost the preservation of life.

In cases like these are where the diversity of opinions could help us to define the values that will regulate an AI system. More diversity of opinions will allow us to establish a set of correct values more precisely.

Autobiography: I am Francisco López Caracheo, actually, I am studying biomedical engineering, in the 5th semester, in the UAQ. I am very interested in how AI works and how it can be used to make the world a better place I would like engineering, medicine and philosophy. Find a way to implement values to AI system is a problem that joints engineering and philosophy. I would like to use AI systems to solve medicine problems. I think that starting to relate me with various of these issues is high importance. That's why I am doing this work.

[1] Jonsen, A. (2006). A History of Religion and Bioethics. En D. Guinn, Handbook of Bioethics and Religion (1st ed., pp. 23-25). Nueva York: Oxford University Press.

Name: Zoe Porter

Affiliation: University of York

On page 16, I recommend**:**

Adding the word "made" after "designed" and before "operated" in point 1 of Background.

Adding a point under Background that AI/AS technologists should not only support the rights and wellbeing of users, but also the rights and wellbeing of workers in the supply chain.

The EICC Code of Conduct could be included as a list of documents in the first paragraph (and included under Further Resources).

Supporting document:

EICC Code of Conduct: http://www.eiccoalition.org/standards/code-of-conduct/

On page 27, I recommend:

In the first paragraph, fourth sentence, adding the words "and assumptions" after "values".

On page 30, a typo:

The surname initial for Patrick Lin, co-author of Robot Ethics: The Ethical and Social Implications of Robotics, is L not P.

On page 39, I recommend:

Under the issue of differentiating culturally distinctive values embedded in AI design, adding another paragraph to Background: A responsible approach also involves interrogating algorithmic and automated decision-makers for assumptions, biases and misunderstandings that may disadvantage members of particular demographic and cultural groups. Algorithms in the financial services sector, for example, may discriminate against people from a 'saving' culture rather than a 'spending' culture, by judging them as 'high risk' because they have no record of paying back loans, and automated medical diagnosis systems might not take into account cultural differences with respect to self-disclosure, affecting the effectiveness of data analysis.

Supporting document:

https://www.ft.com/content/c90e68a4-661d-11e6-8310-ecf0bddad227

Adding, under Candidate Recommendations, the possibility of AI watchdogs to regulate algorithmic decision-making.

Supporting document:

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2903469

(pg. 43 of this paper for watchdog idea. The paper as a whole is also highly relevant as a Further Resource for **page 90**)

On page 96, I recommend:

Adding the word "and assumptions" after "values" in the sentence "what/which set(s) of values guide the design..."

On page 100, I suggest:

With respect to point 4, a debate on cultural sensitivity would be required in cases where affective systems are inserted into societies where certain discriminatory attitudes are prevalent, for example, societies where discrimination against women is commonplace.

With best regards,

Zoe Porter
PhD student
Department of Philosophy
University of York
York YO10 5DD
United Kingdom

"Adequate use of personal information used through mechanisms that use artificial intelligence"

Pérez Covarrubias Juan Carlos
Biomedical engineering
Autonomous University of Queretaro

March 5, 2017

Introduction

Artificial intelligence systems and autonomous systems have a great need for data manipulation in order to generate a rapid statistical response to facilitate the work of the human being, in any type of activity that invests it economically, socially or in the fields of Health Public At the moment the usefulness of this data carries a vast compilation of personal information that must be handled with delicacy. This document proposes to generate a particular opinion on the basis of some recommendations that the article Ethically Aligned Design[1] for the manipulation of personal data of the people, how to regulate them and to fulfill the mandates of the industries, individuals, institutions and other instances that do Use of artificial intelligence (AI) and autonomous systems (SA).

In public services, such as health, it is important to keep a statistical and non-statistical control of the people who may or may not meet the expenses necessary to pay for certain services, so the collection of information is important, especially for monitoring and Improve public or private services. The text mentions two important instances for the need to provide personal information, one in the eIDAS and IDNY are two mechanisms for gathering information that help citizens to apply to different public services. In the text, the following recommendation is generated: "When available, people should identify trusted identity verification resources to validate, test and disseminate their identity" in order to help institutions provide a better public service. Basically it is summarized in the following questions:

1. Who needs access and for what duration? Is it a person, a system, a regulatory body, a legal requirement "or" an input to an algorithm?

2. What is the purpose for access? Is it read, used and discarded or collected, used and stored?
3. Why is the data required? Is it to comply with compliance, lower risk, because it is monetized, or in order to provide a better service / experience?
4. When will it be collected, for how long will it be maintained, when will it be discarded, updated and re-authenticated? How does duration affect the quality and life of the data?

Before providing any type of private information it is recommended to question the four previous points that were generated under a recommendation in the text because in many occasions it is possible to make a misinterpretation of the data or also can make inferences of the shared information that the individual He did not want them shared. It is recommended that AI and SA not only analyze the information, but also seek to help people understand the granular level consent in real time. In this way minimize risks in the long term by means of data collection.

On the other hand also it is possible to emphasize an important question that is “how we can control the use of our information?” Many times because of lack of information people do not know how they can control the correct use of their data, so the text mentions some recommendations as: "Algorithmic guardian platforms should be developed so that individuals can heal and share their personal information. The guardian could serve as an educator and negotiator on behalf of its user, suggesting how the requested data could be combined with other data already provided, informing the user if the data is being used in a way that was not authorized, or doing Recommendations to the user based on a personal profile ". This will improve the correct and appropriate use of people's information.

Conclusion

In conclusion, reference is made to the recommendations provided by the text, which help us to generate a general position, where we can propose an international committee responsible for the regulation of data management, as previously seen only in some developed countries. Have this type of identification regulation. This not only to be able to apply to public services (by each country or state) also for international cooperation such as trade. The same international committee should assume responsibilities with countries to advice and train institutions that make use of data management, through representatives capable of providing technical and advisory assistance. It is also necessary to inform society about the protection it has based on the correct use of its information through legal entities such as the Organization of American States (OAS) or the National Institute of Transparency, Access to Information and Protection of Personal Data (INAI) in the case of our country. However, such organizations should encourage international cooperation to improve in the long term the misuse of personal information, fraud and any other form of embezzlement.

It is also recommended to provide assistance to persons who are affected by the misuse of their personal information or disclosure of it in relation to institutions responsible for regulating them nationally or locally, this could be achieved through the international committee so that it plays a role of mediator between institutions and people harmed. Because it would not be possible to intervene directly with the individual out of respect for national rights but to follow up on the case.

[1] Engineers, I. o. (2016). Personal Data and Individual Access Control. En *Ethically Aligned Design: A Vision For Prioritizing Wellbeing With Artificial Intelligence And Autonomous Systems* (págs. 56-67).

Bioethics in artificial intelligence, personal data and access to individual control

Raymundo Vargas Parra
5th Semester. Biomedical engineering
Autonomous University of Queretaro
February 2017

Artificial intelligence (AI) is a discipline that gradually takes its place in our daily life; Given that it has grown at very small steps, society has not noticed this growth, neither its threats nor its great virtues, so it has not been given the importance it deserves, this is an area that begins to generate debate, Over time the limits of rightness, ethics and scientific advances have been questioned and therefore has had to balance the positions to know which decisions are correct, or at least delimit to what extent is acceptable measure or not. This essay demonstrates positions of religious sectors, political opinion and social views, in order to know what are the best decisions and attitudes that have to be taken into account as the AI is developing, as well as my point of view As a student of biomedical engineering, based on proposals and principles described in Ethically Aligned Design created by IEEE committees.

Artificial intelligence has been growing based mostly on experiences, impressions and intuitions of people competent in the area, who seek to improve the quality of life or make dreams come true, this and more have made it possible, thus contributing to the advance Medicine, engineering, and notable economic growth, but its subtle growth makes it impossible to appreciate its breakthrough as an emerging technology.

The most palpable approach that society is to science fiction films, where we end up saying "that does not happen", or "it's been many years for that to happen", the truth is that these realities are not far; they are closer than Can you imagine. Knowing that this type of technological advance will help us to progress as a society gives the animus to the science of continuing to innovate, improve and discover more about this, but just as growth is latent, it has to be defined as it has to grow .

Artificial intelligence proposes that it is fundamental for people to define, access and manage their personal data as conservators of their unique identity, which makes me a very sensitive issue because, while respecting our autonomy, we must take into account that we are the result of our past actions and decisions, I believe that manipulating personal data in this way would destroy who we are by someone who would like to be what would change the perspective of the human being.

Jonsen (Guinn, 2016), shows us religious positions where the natural law is questioned, which shows us the divine order by which a person is created, and on the other hand the era of modernization and technological advances. It rescues from reading an important principle, the will of man.

The main theme that religion asks to be careful in AI is precisely the will of the person who is defined as "Individual freedom requires reflection and conscious choice" according to the dictionary.

The AI is an area that helps us automate, and control the processes, and progress in various areas, also its misuse could lead us to this problem, access to control of information of some individual which would break it That God has taken care of.

Talking with some friends about their opinion on this subject, their position is that science grows and we must be up to the point and willing to grow technology, because if you can change some information neural, then you can change Neuronal problems that cause diseases or disorders which is a gigantic contribution to the area of medicine, but use it for human experimentation methods can bring with it a bigger problem where instead of curing the patient they leave it worse, or someone healthy leave it With cerebral traumas. But in turn a part of friends with what I could talk about this topic, commented his approval to the creation of neural database in order to give people experience of the things that are recorded in the brain memory.

Future informed consent must be based on a limited and specific exchange of data against the long-term sacrifice of the active means of information, since a certain part of the brain is manipulated and controlled, it is important that the person who performs it completely Aware of the processes that are performed.

Doing an investigation of the subject I realize that respect for the autonomy of the person. It is necessary to defend, no one has the right to violate the opportunity that a person has to be autonomous, although God respects this principle, we should not do it, because that prevent

Bibliography

Guinn, D. (2016). *Handbook of Bioethics and Religion.* Nueva York.

Kim E. Barrett, S. M. (2016). *Ganongs Review of Medical Physiology.* Mc. GrawHill.

Artificial Superintelligence

Tamara Hernández Alvarado1

With the growth of Artificial Intelligence, certain concerns are created, as mentioned in the issue "As AI systems become more capable, as measured by the ability to optimize more complex objective functions with greater autonomy across a wider variety of domains, unanticipated or unintended behavior becomes increasingly dangerous2" mentioned in the page 50, ethical principles have been sought that incorporate the highest ideals of human rights, that maximize the benefits for humanity and the environment and that manage to mitigate the risks or negative impacts that the evolution of the Artificial Intelligence brings with it.

The technology has an exponential growth, so we can expect that at some point the machines will become more intelligent than the humans, thus having an intelligent explosion, which Tomasik3 describes as the clock that runs between biological and digital minds. According to Chalmers4 this event will be followed by an explosion of increasing levels of intelligence, as each generation of machines will create smarter machines in turn. This explosion of intelligence is now better known as singularity. I totally agree with this hypothesis, where it is clear that the machines can be self-sufficient to fulfill the tasks assigned to them without the need for human help.

The key idea is that a machine that is smarter than humans will be better than humans in machine design. That way you'll be able to design a machine smarter than the most intelligent machine humans can design5. With this reasoning, we can intuit that this machine will be able to design a machine smarter than itself. So we would expect a sequence of smarter machines each time. We must be ready for when this development arrives, since it will be necessary that we also change and adopt a new society that will be dependent of technology.

Technological changes are faster growing and can lead to unpredictable consequences. If there is artificial intelligence, then there will be Superintelligence. It is inevitable that gradually anger artificial intelligence leaving behind biological beings, while technological progress is increasing. It is enough to look back and note that the history of life on Earth consists of one species that overcomes another, over and over again, where two species competing for the same resources cannot coexist, suggesting that In the long term, humans or machines will ultimately occupy the role of the most intelligent beings on the planet, and that as soon as one species has the advantage over another, then it will dominate in the long run.

If we create a world with Artificial Intelligence and Superintelligence, what will be our place on the planet? Chalmers6 exposes 4 options: extinction, isolation, inferiority or integration. It is important to foreground these possible scenarios, because it depends on us which of them will be the one that defines the course of our species.

[1] Biomedical Engineering student at Universidad Autónoma de Querétaro.

[2] The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems. Ethically Aligned Design: A Vision For Prioritizing Wellbeing With Artificial Intelligence And Autonomous Systems, Version 1. IEEE, 2016. http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html.

[3] Tomasik, Brian (2016). Artificial intelligence and its implications for future suffering. *Foundational Research Institute*. https://foundational-research.org/files/artificial-intelligence-and-its-implications-for-future-suffering.pdf

[4] Charlmers, David (2010). The Singurality: A philosophical Analysis. *Journal of Consciousness Studies, 17*(9-10), 7-65. http://consc.net/papers/singularity.pdf

[5] Charlmers, David (2010). The Singurality: A philosophical Analysis. *Journal of Consciousness Studies, 17*(9-10), 7-65. http://consc.net/papers/singularity.pdf

[6] Charlmers, David (2010). The Singurality: A philosophical Analysis. *Journal of Consciousness Studies, 17*(9-10), 7-65. http://consc.net/papers/singularity.pdf

- Extinction: Being self-sufficient technology, they will not need us, so we will stop being a priority. During the fight between species, smart machines could win.

- Isolation: continue to exist, but without interaction with Artificial Intelligence, or with minimal interaction, being intelligence totally independent of humanity.

- Inferiority: Digital minds can outnumber us without any problems, they will not need sleep, they will only focus on their tasks, being able to think and make decisions faster than any human. Which leads us to live in common with more intelligent beings which will not allow themselves to be dominated by any specie. Living beings will adapt to a world governed by technology.

- Integration: we can use the technological development to our benefit, generating improvements to our biological system that allow us to develop activities of superintelligent machines.

- Integration seems to be the most reasonable option, but the improvement of the human species is not as easy as designing a robot, it is not only about improving the brain or replacing it with a super advanced one, it is about giving it the characteristics necessary to continue offering humanity.

- Ordinary humans are aware. If we lose the capacity of conscience we would lose our subjective character, which is what gives our lives value and morality to our actions, this would cause in a certain way, the cease to exist. It is important to point out that the machines do not have consciousness, they do not act for good or for evil, they simply fulfill the task assigned to them. The risk lies in the extent to which a machine can be used to perform such tasks. The arrival of superintelligence presents us with an ethically wide panorama, where the development of a superintelligence with a supermoral that with each innovation improves for our well-being, or a system capable of rejecting and re-establishing values previously established, generating their own ethical values and acting in a convenient way for the accomplishment of tasks.

- We must act from this moment to have the control. It is necessary to create a bioethical center, as did Canada in 1976 before the need to have an ethical laboratory that fosters interdisciplinary discussion on the moral and social problems of developments in biomedicine7. In this way, ethical reflection and self-regulation of the development of biological sciences can begin, while artificial intelligence grows, thus generating better benefits for society. We must maintain a slow development to have control of every technological improvement that is going to have, in order to prove, discuss, act and prevent any behavior that may cause risk to society.

- Artificial intelligence will bring great benefits to humanity, such as the improvement of our species, so that we do not have to compete for being superiors on the planet, but where we can coexist in such a way that we have the same capacity as any of the Machines, taking into account that, although we dispose of certain biological parts, the humanitarian part must be maintained. But these benefits must be progressive, so as not to lose the importance of technology regulation, in order to improve biological–technological interaction.

7 Stanton-Jean, Doucet, Leroux & Cousineau (2014). Canada. En Henk A.M.J. ten Have, Bert Gordijn (ed.), Handbook of Global Bioethics (pp. 959-992). Springer Reference.

Daniel Alejandro Morales Hernández
UAQ

Ethically aligned design (págs. 36-48)

Artificial intelligence is a subject in exploration yet, a technology in full research which gives it the name of emerging technology, taking into account this AI is not yet, somehow saying, regularized which means that there are still no organisms Regulators of the advancement of this technology concerning what is ethically good or bad, in this document I want to highlight the importance of the constitution of a regulatory body.

This is because one of the main problems is the lack of ethical and social sciences in the scientific field, requiring a multidisciplinary work with other sciences, thus achieving a balance between all the sciences, and in this way to be able to continue on the right path And beneficial for the human being.

I agree with the statement of "Ethics is not part of degree programs"[1] because not all universities are ethical as a subject, fortunately in the UAQ, the curriculum integrates bioethics as a curriculum, this helps me to begin to know About how futures have to conceive certain points in order to be ethically responsible.

Also the point of "Lack of an independent review organization"[2] this makes me a key point, since there is no ethically regulating body on the subject of artificial intelligence, and in the absence of this body, all research could be left with their own subjective .

Ethics is necessary for all kinds of technology, since this has to be beneficial for all human beings, in addition to its development does not affect the environment and looking for some harmony with all computers.
To harmony I mean to avoid designing new emerging technologies with the idea that the human being is superior, because with this idea and with the development of AGI can get up against the human being.

The human being in his need to dominate everything, is destroying itself, since it can not control or control itself, that is why the need to have agencies that regulate activities and projects, although it remains a body governed by people , Are experts in the subject, besides they do not decide for themselves if not that each action is reviewed by various entities and can give a more objective answer

[1] IEEE. (2016). En *Ethically aligned design* (pág. 37).

[2] IEEE. (2016). En *Ethically aligned design* (pág. 46).

"Incorporation of Values in Autonomous Intelligent Systems" The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems

By Sandra Daniela Carmona Martínez[1]

Even though the current situation is that there are more and more scientific and technological advances increasing and improving the AIS and the actions they perform, together with the fact that there is no standard ethical design that helps to include human norms or moral values in autonomous intelligent systems (AIS), should be well defined that ethics[2], which is a branch of philosophy that studies human behavior taking as the unit a moral act, justifying and making a judgment on how beneficent (good) or inconvenient (bad) is this. Therefore it is necessary to define this concept specifically for AIS as figures that are increasingly taking greater value in society and greater interaction in various areas with humans.

We see that ethics as we know it, is carried out within a community or civilization, is "respect for life"[3], but modern civilizations have no intention of showing love or even interest, and this is why too many human problems such occur as misunderstandings and conflicts that lead to wars, famines, poverty, overpopulation, pollution, etc. But now we have the alternative of solving these problems in a technological way, prioritizing in the first place to protect life, to reduce the pain and suffering of the individual. So we can define an individual as "subject of a life" and this simple fact has an inherent back by a set of rights value and because of its reasoning, has beliefs, desires, perception, memory and sentience[4] (capacity of feeling pain and pleasure). These rights are privileges, jurisprudence that is assigned to each person regardless of nationality, age, race, gender, social class, etc. The most essential rights are right to life, to integrity and security, equality, freedom, work, nationality, etc. Then we can say that rights are the power of the community.

Regarding the identification of norms and values of a specific community, it can conceive culture as a new human genome[5], where ever will self improving and progressing slowly but surely, impacting those who are immersed in that way and causing a collective event that has repercussions in each individual because they are actors who assume an active role in that social environment. An alternative schedule AIS with the same mechanism as something simple that evolves[6] together with the community and culture in which AIS are present. By making a comparison, the brain operates with a computer-like mechanism, receives the information, stores it and processes it with various algorithms and modules, so from the way the information was obtained, the path that was taken to process it and finally the result obtained, will be the key to the decision that is next to be taken, trying to be the one that produces the greatest benefit, optimization of resources, simplicity in the operation, etc. However, there is a probability that these algorithms whose result implies a decision to carry out a determined action will enter into conflict or state of moral overload due to the clash of values, where it is preferable to follow the priority rules and to break the secondary ones, especially in the case in which a benefit is obtained, to define this priority and to introduce it to the order of operation in AIS.

All technological progress has been rejected at the time of its appearance[7] and once people realize that this has resulted in a benefit that improves quality of life, saves time, resources and human effort, now it becomes a necessity created[8] wherein now members of the community involved with technology, or in this case the AIS, experience training consumption, putting their production and obteinment as a priority, even when it is not satisfying a basic need, especially if your operation includes knowledge or procedure easy to understand and even more especially if there is the presence of the primacy of the image[9], which favors the impoverishment of the ability to understand, where it brings apparently learning, but this is not significant, resulting in a waste of time disguised as productivity.

In conclusion we can point to a new concept in ethics that encompasses new technologies automated as AIS, something like roboethics[10] to standardize and analyze movements on the machines and their interaction with humans, to be established and recognized with its respective principles and the issuing of judgments such as ethics currently, where the benefit that AIS bring to our societies is always greater than the disadvantages or prejudices existing to represent a profit and a gain due to its utility, avoiding in as far as possible, moral overload[11], favoring in its design and programming those values or community rights that are of vital importance such as human rights or those norms that have the greatest value for society where there is presence of AIS, taking into account that this, as well as the progress of technology, can evolve socially, where these changes must be contemplated in the AIS system, along with that improvement of human activities, but with a mechanism that is not only complacent in terms of harmony In the human-machine interaction, but also improves the capabilities and learning of its users without dehumanizing the human by stripping him of his opinions, concepts and skills, but rather by increasing his critical capacity, abstraction, activating his memory, discernment and revitalizing their way of learning, complementing it as a rational living being.

Footnotes page and References
(1) Biomedical Engineering Student at Universidad Autónoma de Querétaro, Mexico
(2) Von Hildebrand, Dietrich (1983). Ethics. Pp. 12. Spanish edition, Madrid. I meet Ediciones. Https://books.google.com/books?id=m9f7ThvF6j4C&printsec=frontcover&dq=%C3%A9tica&hl=en&sa=X&ved=0ahUKEwjo9dyU7L_SAhUP8WMKHUioC9sQuwUIGzAA#v=onepage&q=%C3%A9tica&f=false
(3) Urdaneta-Carruyo, Eliexer. Albert Schweitzer: Man as a symbol. Venezuelan Society of History of Medicine and Latin American Institute of Bioethics and Human Rights (ILABID). Gac Méd Méx Vol. 143 No. 2, 2007. Los Andes University Hospital, Merida, Venezuela. Http://www.medigraphic.com/pdfs/gaceta/gm-2007/gm072l.pdf
(4) Leyton, Fabiola. Environmental ethics: a review of biocentric ethics. Number 16 - April 2009. From the Master. Journal of Bioethics and Law. University of Barcelona. Http://www.ub.edu/fildt/revista/RByD16_master.htm
(5,6) Tomasik, Brian. Artificial Intelligence and Its Implications for Future ff ering His. Foundational Research Institute. June 2016. pp. 4; 16-17. Https://foundational-research.org/files/artificial-intelligence-and-its-implications-for-future-suffering.pdf

(7) Sartori, Giovanni (1997). Homovidens: The society Remote control. Pp. 29-33. Ed. Taurus. Http://centromemoria.gov.co/wp-content/uploads/2013/11/Homo_Videns_La_sociedad_teledirigida.pdf

(8) Zuleta Salas, Guillermo León. The emergence of bioethics and the reason for it. Lasallian Research Journal. Vol. 11 No.1 - 2014. Pp. 29-30. Http://www.scielo.org.co/pdf/rlsi/v11n1/v11n1a03.pdf

(9) Savater, Fernando (1997). The value of educating. Pp. 50-62. Barcelona, Spain. Ed. Ariel. Http://www.ivanillich.org.mx/Conversar-educar.pdf

(10) Roboethics http://www.openroboethics.org/results-how-much-interaction-with-a-robot-is-socially-acceptable/

(11) The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems.
Ethically Aligned Design: A Vision For Prioritizing Wellbeing With Artificial Intelligence And Autonomous
Systems, Version 1 IEEE 2016 pp. 24-37

http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html

Comments from
Pavel M. Gotovtsev, PhD,

Vice-head of biotechnology and bioenergy department, National Research Centre "Kurchatov Institute"

Dear authors of IEEE *Ethically Aligned Design* Document, thank you for preparation of this extremely required document. This document looks solid and very useful. I'm not a specialist in AI and Ethics questions but I hope my several comments and questions can be useful for you. Next, after question numbers documents page numbers presented.

Q1 p.24. Does in possible to develop different ethic norms for different communities? Are there universal ethic norms that partially implemented in every community? These norms can be base for every AGI/ASI.

Q2 p.25. Is the possible to provide algorithmizing of GLP? Today looks this is very difficult question that require additional researches and I think it is necessary to additionally highlight in Candidate recommendation section in this page, or in the p.30. That can lead to additional attention of governments and can improve support of researches in this field.

Q3 p.29-30. It is clear that Machine Learning (ML) aspects of AGI/ASI is one of the most challenging questions today. And for me it is not very clear defined main issues of ML applications from the side of goals of its operation. For example - AGI/ASI "survival", by the word "survival" I mean AGI/ASI subsystems, programs, subprograms or algorithms that responsible for stable and full operation of AGI/ASI. In case of ML application for "survival" tasks: how we can control this learning with correspondence to not only ethics but also safety for humans. Does exist possibility that survival goals will be more significant than ethical after periods of ML during AGI/ASI operation? Especially it looks critical for military applications. I think we can clarify several of such issues but mention that in future exist possibility of new issues due to fast progress in AI field.

Q4 p 31. The main issue – norms implemented in AIS must be compatible with norms in the relevant community – lead to the next questions: Can ethical norms, that we implemented in AIS, influence on AIS efficiency? Can these norms lead even to the ban of some AIS technologies (for example deep learning) in some communities? And in its turn this ban can lead to lag in technological development if those communities in all fields there AIS can be implemented including medicine. In its turn, it leads to situation than ethical norms will lead to technological differentiation of communities. I think this is important questions because for example today is existing different communities with ban on several medicine treatments (surgery or blood transfusion) and it is easy to expect strong ethical regulations in some communities for AIS.

Q5 p.41. This question is related with Q4 but implemented for business practice. How ethics norms can influence on technological perfection of AGI/ASI? Did the stronger ethics norms lead to limitations in AGI/ASI technology development? And in case of positive answer we can receive situation than some business companies will develop technologies with only visibility of ethics.

Q6 p.49. citation: "…we recommend that institution set up review board…" This review board can include members of several different institutions to make decisions more independent. I think this postulate can be presented in document.

Thank you very much for this document. I hope my questions and comments will be somehow useful for further development.

Sincerely

Pavel Gotovtsev
PhD tech. science

Development of synthetic organs in artificial intelligence

Autonomous University of Queretaro

González Ramírez Aldo Aarón
Ethical Aligned Design (22-35pp)

In this paper work I will develop the topic of synthetic bodies from the bioethical field, I will build on the relevant readings[1], in order to concentrate all the concepts acquired during the ieee pdf and apply them in my subject of interest: Artificial organs.

More than a great discovery, the development of artificial organs, is a precise necessity, since nowadays it is very requested a human organ which can satisfy vital needs within our organism.

It is also worth noting that, as we have already noticed, the appearance of new diseases are current, a fact that affects us too much if it is a degenerative disease, where we have no choice but to get rid of the affected area, which can be an any member or human organ, for that disease.

But not the whole picture is rough and gloomy, because while the obstacles are presented for humanity, this grows exponentially in the field of wits, as we well tells Tomasik[2] in its Artificial Intelligence and Its Implications for Future Suffering (2016, pp. 1-40), where he expresses to us with differential equations the behavior of human ingenuity, where human innovation grows exponentially on the basis of human ingenuity, which leaves us wondering if we are really far removed from a reality where degenerative diseases become a simple flu so to speak.

It should be noted that the last formula is applied by Tomasik only for artificial intelligence, but, in the formula, it expresses human ingenuity clearly, that in my opinion, the generation of new technologies in health, in this case synthetic organs, is a great hum example of ingenuity year. In fact, a human prosthetic, whether we speak of a rudimentary artificial kidney, is considered artificial intelligence.

In general, this idea of human organ replacement is very well positioned, since, in addition to complying with the moral statutes, we do not question procedures, since we do not need another human being, or rather, of an organ donor. I mention this because, somewhat surprisingly, there are religions that forbid organ donation[3], which should not be, because if it is true, most religions converge on the idea of prevailing life, as we tells Jonsen[4] in its "a history of religion and bioethics" (1988), where with the help of three main writers (McCormick[5] , Fletcher[6] and Ramsey[7]), who tells us about the "natural law , " which seeks to restore life At any cost; And in the case of an organ transplant, is vital to save the life of the patient, and even so is prohibited by the religions already mentioned.

Because of the above problem, it is necessary to create organisms that orient new technologies in the bioethical field, and with that to solve any religious controversy that hinder human development, something that we see very well explained in the text "Handbook of Global Bioethics" (2014), in the section of the country: Canada. Country that becomes the perfect example of technological, moral and social advance, being a pioneer in "bioethical"; I put in quotation marks bioethical because it was not developed as such in Canada, but its bases. This nation faces the system proposed by Fletcher, which

This type of emerging technology, and development takes time and research, which we see reported in the documentary "STEM CELLS: THE KEY REGENERATION"[8] . This documentary focuses politically, socially and economically on the problem of replacing an already useless human part.

It is important to highlight the regulation of these technologies, since as we will see in the research made, stem cells are a double-edged sword, so its arduous research, since its poor application can generate fatal tumors (carcinomas), giving as Resulting in a worsening of the patient's situation. Therefore we will see in the research test takes several individuals, who are we apply the experimental treatment, already approved in various deficiencies[9] in their bodies. Treatments that proved to be effective in its application, so much that it was possible to heal the experimental subjects.

To conclude this work, I would like to express my concern about the accelerated growth of human ingenuity focused on evil, which I agree with Tomasik, who tells us that the day will come when technology surpasses man, and not because of reference to films apocalyptic[10] cited by Tomasik, but be aware to investigate further and investigate new technologies, in order to better control them, and thus its proper application.

I do not say that artificial intelligence is bad, but it is dehumanizing, so much that it can lead us to replace our brain with simple algorithms and mathematical patterns, and our body in a metallic set of circuits.

[1] A history of religion and bioethics, 1988, Albert R. Jonsen, Oxford University Press, David E. Guinn.
· In Handbook of Global Bioethics, 2014, Henk AMJ ten Have Bert Gordijn Editors, Duquesne University.
· Pittsburgh, PA, USA, Springer Reference, (pp. 960-990)Artificial Intelligence and Its Implications for Future Suffering, 2016, Brian Tomasik, Foundational Research Institute, pp. 1-20.
[2] Tomasik, Brian. 'The Importance of Wild-Animal Suffering'. Foundational Research Institute.
[3] Judaism, Buddhism and Orthodox.
[4] Albert R. Jonsen [7]. Bioethics and writer, professor of ethics in medicine at the University of Washington.
[5] Richard McCormick (1923 to 2000). Jesuit professor of theological morality in the Jesuit seminary and editor of "Notes of theological morality".
[6] Joseph Fletcher (1905-1991). Professor of theological pastoral and Christian ethics at the Theological Episcopal School in Cambridge, Massachusetts.
[7] Paul Ramsey (1913-1988). A professor of religion at Princeton University, an American Christian ethic of the twentieth century, he was a Methodist and native of Mississippi.
[8] Documentary by Discovery Channel, which aired on February 12, 2012.
[9] Treatment research was used to regenerate the cornea to restore vision to blind patients; and how these cells can also be injected into the heart to clear the arteries and be used to treat type I diabetes.
[10] "I Robot" (2004) and "Terminator" (1984)

Dr. Christopher A. Tucker

Contribution in the form of feedback to ethically aligned design, Version 1

The document, while proposing a set of rules for the development of autonomous intelligent systems in terms of its responsibility and expectations during the course of human-machine interaction, freely admits that the means by which to root this ambition does not exist universally, rather, in competing societies and cultures. Under such a condition, the creation of punitive measures to limit undesirable norms contrarily to human cognition labors only within trial-and-error methodologies for value embedding. Intellectually, this implies that an a posteriori judgment is yielding an intuition to acquire knowledge about the AIS. We are hoping to understand what constitutes the scope of desirable behaviors in the context of human norms only by experience and not by cognitive reflection before hardware is designed and software is written. As it could be argued that research into AIS since its appearance in the 1950s has not closed the gap between what is known and unknown about AIS during runtime, therefore, a baseline philosophical framework is relevant to aid in understanding the limit of what are the range of possible behaviors for this AIS in context with human cognition, intuition, and judgment. It is therefore recommended, that Immanuel Kant's *Critique of Pure Reason* become the source wherein to generate this universal philosophical framework.

As the recommendation contained herein is unduly vast in context with the document, I propose additions to the text where mention of this concept is valid, given a further explanation in a resource citation. For example, on page 31, between the paragraph ending "…which suggests primarily a similarity structure" and "In addition, more concrete criteria must be developed…", I suggest the following addition for an example of a structure with a concrete criteria:

> Such an alignment of similarity structure is the introduction of a reasonable critique of human cognition [1] which could be applied as AIS norms, whereby creating a framework with which to guide implementation by the designer [2]. As presently admitted, one does not universally exist; therefore a framework, which consists of pure reason in the science of artificial intelligence, is required.

Thank you for your kind consideration.

Sincerely yours,

Dr. Christopher A. Tucker

References

[1] I. Kant. *Critique of Pure Reason*, Translation by J.M.D. Meiklejohn, London: George Bell and Sons, 1897.

[2] C.A. Tucker. "A proposal for ethically traceable artificial intelligence," arXiv: 1703.01908, 6 March 2017

NOTE TO THE EDITOR: The cited arXiv document for this feedback was made available just after the submission deadline for this recommendation. Before it is published, how can I send across the correct arXiv number when it becomes available?

Education and awareness about artificial intelligence

Quintanar Pozos José Eduardo.
The Autonomous University of Queretaro.
Pg: 21-23.
Biomedical Engineering.

For many years science fiction has looked to a future full of intelligent robots capable of helping solve complicated tasks or, in special session, our race was dominated by them (Tomasik, 2016) [1] . Today, science has managed to develop some systems based on science fiction, utilities that help man to perform his tasks in an optimal way. Nevertheless, this development has reached a rhythm that exceeds the speed of the social development of the man. From this same argument, religion considers that man loses its essence through the development of technologies, Artificial Intelligence represents the imitation of the supreme God, the creator of life, which puts mankind at risk and at the mercy of wrath of God, is as written in the scriptures that man should never try to play God (Jonsen a, 2006) [2] .

Therefore, in this document it is proposed to legislate on this type of scientific advances in addition, it becomes aware of the need in the need to create bodies in charge of social education for the discussion of the needs of society in the area of Systems Intelligent Where we talk about all the consequences of the application of these technologies to a society increasingly dependent on machines. It is important to create these bodies since the standards or guidelines to help human norms or values have not yet been set against autonomous intelligence systems.

Today it is hard to imagine how a man can give values to an artificial intelligence system because of the complexity surrounding human values. Despite what has been said, this is a totally achievable goal and, moreover, is already a reality.

It is important for the development of AIS to incorporate fully explicit moral standards. That for this type of intelligent systems would only follow and abide by a series of behavioral instructions for a given context. It is important that the artificial intelligence system reflects the values of the type of community to which it is programmed so that there is a good interaction between the man-machine.

To address this need The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems (EEI, 2016) [3] has set three main objectives for incorporating values to these systems:

1. Identify the rules and obtain the values of a community affected by specific AIS.
2. Implement the norms and values of that community within AIS.
3. I assess the alignment or compatibility of these standards and values between humans and the AIS within that community.

Obtaining these three objectives represents an active process for a good coexistence between the human and the machine within a specific community. Avoiding in this way that the intelligence system is overloaded with unnecessary values not proper to the community in which it is interacting. According to the EEI, this is defined as "moral overload" (EEI, 2016) [4] . That is why a focus should be placed on all stakeholders so that they can see that the systems are designed to give "transparent" results (such as explanations or inspection capabilities) about the specific nature of their Behavior towards the various actors within the community they serve.

As IEE mentions, this practice can not always eliminate the potential data bias present in many machine learning algorithms.

The values that must be incorporated into the Artificial Intelligence Systems should not be global, but specific so that the needs of each of the communities can be considered. It is important to propose the identification of these values. However, moral laws that govern a community are still difficult to identify). In addition, within these communities, there are subjects that differ from each other so that the same pattern of values can not be achieved.

It is important to know what kind of artificial intelligence system is going to be used and for what purpose, since, from this, the order of the values prioritization changes depending on the context of the intended community or even, Of time these values and norms become obsolete. Taking with it a great advantage where the scientist can create new systems capable of learning from their interaction with their environment, resulting in information about the user

The hope of the scientific community in favor of these technologies is the active inclusion of users and their interaction with the Artificial Intelligence Systems to increase the trust and general reliability of these systems. (EEI, 2016)

The history of conflicts between men has shown us that the problem often lies in the way in which these technologies are used.

It is important to take into account that we can not escape the impact that technology, in its various manifestations, has been reaching the world in which we are living. We have reached a point where human behavior can not be understood without adding to technology as an important factor in man's behavioral changes. Although technological development was once considered a source of dehumanization, it is now necessary to join this axis of life increasingly dependent on devices and machines.

[1] Tomasik, B. (2016). Artificial Intelligence and Its Implications for Future Suffering. En Foundational Research Institute (1st ed., pp. 3-9).

[2] Jonsen, A. (2006). A History of Religion and Bioethics. En D. Guinn, Handbook of Bioethics and Religion (1st ed., pp. 23-25). Nueva York: Oxford University

[3] The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems.
Ethically Aligned Design: A Vision For Prioritizing Wellbeing With Artificial Intelligence And Autonomous Systems, Version 1. IEEE, 2016.
http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html.

[4] The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems.
Ethically Aligned Design: A Vision For Prioritizing Wellbeing With Artificial Intelligence And Autonomous Systems, Version 1. IEEE, 2016.
http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html.

Interactive Exercise on Next Steps for the Academic Community (based on IEEE Ethically Aligned Design document) Workshop at AI for Social Good, Waseda University, Tokyo, Japan - 7 March, 2017

Autonomous Weapon Systems Group - Summary of Outputs

Members -

1. Danit Gal, Peking University/Tencent Technologies (gal.danit@gmail.com),
2. Julianne Chan, Digital Asia Hub (jchan@digitalasiahub.org)
3. Michael Baak, Digital Asia Hub (mikedbaak@gmail.com)
4. Udbhav Tiwari, Centre for Internet & Society (udbhav@cis-india.org)

Outputs

1. **Need for Auditable Protocols in Missile Defense Systems**

Missile defense systems that prevent targeted strikes by short, medium and long range attack missiles play a key role in the defense of strategically important locations, especially in areas of conflict. These systems are largely automated, from the initial steps of detecting a missile launch, tracking multiple missiles in their course and predicting their targets to finally deciding the best time and place to neutralise these attacking missiles. Such systems often have to take a call on which incoming missiles to prioritise for neutralising tactics in case of multiple incoming attacks and well as choosing the best location to perform the neutralisation. The average from from launch to hitting the target for such systems ranges from as less as 1 minute in case of short range missiles launched a few kilometers from the border to 10 minutes in case of ICBM launches. This incredibly short period almost mandates that the response times for such systems be measured in seconds, minimising human involvement. Therefore, considerations that go into designing the algorithms that operate in the software concerned with making decisions such as which possible targets to prioritise, the factors used for prioritisation (such as population or strategic value of possible targets), methods used for neutralization as well as the location of the neutralisation (in terms of damage on the ground) are of key importance. The ethical and principle factors that go into such such design decisions, both for the actions carried out by the missile system in case of an attack as well as auditing of such actions post such actions being carried out are of key importance to ensure maximisation of transparency, accountability and human rights concerns while minimizing loss of life or harm to property. The group thought that such decisions must be subject to review by humans when necessary and must also be periodically be reviewed to ensure modern standards of transparency, accountability and human rights are applicable to the entire system, automated or not, in the context of self defense and conflict.

2. **Need for Accountability Smart Projectiles**

Smart projectiles such as bullets, guided anti-personnel munitions (including grenades and missiles), etc. have the capability to both be guided post leaving their launching devices as well as exhibiting autonomous behaviour about how they may choose to neutralise their intended target. This includes choosing to incapacitate rather than creating lethal impact, rerouting their trajectory to avoid obstacles or to choose a new target, etc. The software and hardware routines that enable such decisions to be made, while designed by humans, are often subject to minimal or no human intervention once the projectile itself is launched. This leads to various questions about balance of culpability between the projectile and the human operating the launcher, especially in cases where the intended target may not be in the final destination of the projectile. The collateral harm that may occur in the day to day usage of such devices as well the long term impact impact of such events on the perception, usage and review of such smart projectiles can benefit significantly from consideration of ethical principles in their design from the the outset. The group decided that the primary steps in such cases would be those of rigorous testing prior to deployment according to international minimum standards, periodic review of how such devices are utilised in the field and a clear hierarchy of how responsibility will be delegated down the chain of command operating the device.

3. **Domestic Use of Autonomous Weapons**

The para-militarisation of domestic law enforcement has been on the rise in the past few decades, especially in regions considered vulnerable to armed conflict or strategic targets that can be as large as entire cities. In such cases, the use of automated weapons for crowd control, riot management, bomb defusal, non-lethal takedowns and even hostage rescue has been on the rise. In the context of them being used domestically, in civilian environments, with largely localised law and order breaches, the impact of such devices of civilians can lead to a variety of human right infringements, such as bodily harm, excessive force, lack of accountability, etc. Such weapons are largely domestically produced and rarely undergo the level of testing required by weapons used in battlefields. The group decided that having a separate set of more stringent guidelines, for such devices, which range from tear gas launchers to drones, to ensure minimal impact on human rights, accountability to democratic institutions and clearly defined use-case scenarios that can be audited post-facto are an urgent need to ensure sustainable and trustworthy use.

4. **Non Weapon Use of Autonomous Systems**

In both conflict and domestic usage of autonomous systems, there are a range of devices that rely on cameras, radar and other form of technology that have potential for abuse outside of weapon systems usage. Using an automated robot that can defuse bombs or even human targets, as a surveillance device with its camera and radar recording information for use beyond the mission, is one such possibility. Similarly, drones being used to capture personal and private information, including recordings inside one’s home, movements of personnel, etc. is also another scenario where there can be privacy, cyber security and proportionality considerations. Keeping this in mind, the group suggested that there be a clear mandate for the utilisation of such autonomous systems, a clear documentation of the outputs recorded by them and a secure chain of custody, both for storage and usage of such information be established prior to such systems being deployed in the filed, domestic of armed conflict. Ensuring evidentiary laws also account for the method is which evidence obtained by such devices was also discussed.

The following refers to the same Workshop as mentioned above. Transcriptions from this event to be provided soon.

AI for Social Good, Waseda University

Interactive Exercise on Next Steps for the Academic Community (with the IEEE Ethically Aligned Design document as a provocation), 7 March 2017

Participants and reporting groups

1. General Principles, Embedding Values into Autonomous Intelligent Systems, and Methodologies to Guide Ethical Research and Design

Michael Veale, University College London (rapporteur)

Ksenia Duxfield-Karyakina, Google

Nishant Shah, Artez University of the Arts, Arnhem, Netherlands

2. Safety and Beneficence of Artificial General Intelligence (AGI) and Artificial Superintelligence (ASI)

Masahiro Fujita (rapporteur)
VP, Head of Technology Strategy Department, System Research Development Group, Sony Corporation

Satoshi Kurihara
Professor of the University of Electro-Communications Graduate School

Kentaro Torisawa
Director General, Data-driven Intelligent System Research Center (DIRECT), Universal Communication Research Institute (UCRI), NICT

Shin Nomoto, Executive Principal Researcher, KDDI Research

Toshie Takahashi, Professor of Waseda University

3. Reframing Autonomous Weapons Systems

Udbhav Tiwari (rapporteur)
Centre for Internet & Society, India

Danit Gal
Peking University/Tencent Technologies

Julianne Chan
Digital Asia Hub

Michael Baak

4. Personal Data and Individual Access Control, Economics/Humanitarian Issues and Law

Bettina Berendt (rapporteur)
Department of Computer Science, University of Leuven, Belgium

Celina Beatriz
Instituto de Tecnologia e Sociedade do Rio de Janeiro (ITS), Brazil

Ryan Budish
Berkman Klein Center for Internet and Society, Harvard University, USA

Herbert Burkert Research Center for Information Law, University of St. Gallen, Switzerland

Andres Guadamuz
Intellectual Property Law, University of Sussex, England

Malavika Jayaram, Executive Director, Digital Asia Hub

Amy Johnson
History, Anthropology, and Science, Technology, and Society (HASTS) program, MIT, USA

Julian Thomas, Social Change Research Platform and RMIT University, Melbourne

Dear Sir/Madam ,

Thank you for the excellent work of your team. It’s really meaningful to make such a document for the development of AI/AS. I have some personal comments and further resources about the document that I hope will be useful for the update.

And may I ask to join the committee of methodology as a member, please？I’m really interested in the work and would greatly appreciate any of your favorable consideration of my application and looking forward to your reply. Thank you.

Some personal Comments on Ethically Aligned Design

1. *The document’s purpose is to advance a public discussion of how these intelligent and autonomous technologies can be aligned to moral values and ethical principles that prioritize human wellbeing.*

This is right. But the *moral value and ethical principles* are not right inherently. They will also change with the passage of time. What’s more, the document said in part 3 that there is a need to *differentiate culturally distinctive values embedded in AI design*. So which value and/or principle should the AI aligned to？How can we be sure that the *moral value and ethical principles* are reasonable?

2. What’s the difference between *ethically aligned design*, *values-aligned design* , *value-aligned system design* and *values-based design* (P36)？I noticed that the document uses the four terms to express the same/similar meanings. Keep one key term consistent would be better.

3. *It recognizes that machines should serve humans and not the other way around*(P36).

What does this sentence mean？Humans should be the master rather than the machine？If we want the machine to work well , we have to fund it , create it , maintain it , recycle it , or in one word , serve it. So I think the relationship of human and machines is not who serves who , but how they *accompany* one another (Peter-Paul Verbeek , 2010). We and machines should be friends , rather than master and servant.

4. *Innovation should be defined by human-centricity versus speed to market*(P36).

According to the normal ethics，*human-centricity* is a wrong idea，because it means humans only care themselves，see others，the environment，animals，plants，technology as tools only. And，*the discovery of an unsatisfied need does not necessarily justify the launch of an innovation*(Xavier Pavie 2014).

5. *Ethics is not part of degree programs* (P37).

To my personal understanding，this means ethics should not just be a course in the philosophy department，but should also be a course for other studies. So “Ethics is not ***only*** part of degree programs” would be better.

6. What’s the difference between *culturally distinctive values*(P39) and difficult values，please？Every value is a *distinctive* value even they belong to the same culture. Because they are ***different*** values，e.g privacy and safety. Different values include *culturally distinctive values*，rather than the other way round. So ***different values*** is a more suitable term here.

7. The document sometimes use *AI/AS*，sometimes *AIS* and sometimes *machines*，I think using the same term is better.

Further Resources for the Methodologies to Guide Ethical Research and Design

Issues：Ethics is not part of degree programs.

1.【edX course】Responsible Innovation: Ethics, Safety and Technology(How to deal with risks and ethical questions raised by development of new technologies)

2.【OZSW course】Philosophy of Responsible Innovation (2016)

3.【HEIRRI project】Higher Education Institutions & Responsible Research and Innovation

4. Jeroen van den Hoven, Pieter E. Vermaas, Ibo van de Poel eds. Handbook of Ethics, Values, and Technological Design[C]. Dordrecht : Springer , 2015

5. Richard Owen , John Bessant , Maggy Heintz. Responsible innovation : Managing the responsible emergence of science and innovation in society[C].Chichester : John Wiley&Sons Inc., 2013

6. Snow C P. The Two Cultures[J]. Leonardo, 1990, 23(2/3):169-173.

7. Kagan J. The three cultures: Natural sciences, social sciences, and the humanities in the 21st century[M]. Cambridge University Press, 2009.

Issues : **We need models for interdisciplinary and intercultural education to account for the distinct issues of AI/AS.**

1.【PARRISE project】Promoting Attainment of Responsible Research and Innovation in Science Education

2.【IRRESISTIBLE project】Engaging the Young with Responsible Research and Innovation

3.【ENGAGE project】Equipping the Next Generation for Responsible Research and Innovation

4.【EnRRICH project】Enhancing Responsible Research and Innovation through Curricula in Higher Education

Issues：The need to differentiate culturally distinctive values embedded in AI design.

1. Ibo van de Poel. Conflicting Values in Design for Values[A]. In Jeroen van den Hoven, Pieter E. Vermaas, Ibo van de Poel eds. Handbook of Ethics, Values, and Technological Design[C]. Dordrecht：Springer，2015：89-116

2. Wim Ravesteijn，Jia He，Chaohe Chen. Responsible innovation and stakeholder management in infrastructures:The Nansha Port Railway Project[J].Ocean&Coastal Management，2014(100)：1-9

Issues：Lack of value-based ethical culture and practices for industry.

1.【Responsible-Industry project】Responsible Research and Innovation in Business and Industry in the Domain of ICT for Health，Demographic Change and Wellbeing

2. Bernd Stahl.Foreword.In Konstantinos Iatridis·Doris Schroeder.Responsible Research and Innovation in Industry：The Case for Corporate Responsibility Tools[M].Dordrecht：Springer，2016

Issues：Need to include stakeholders for best context of AI/AS.

InterAction：How can academics and the third sector work together to influence policy and practice? ：https://www.rri-tools.eu/-/how-can-academics-and-the-third-sector-work-together-to-influence-policy-and-practice

Issues：**Lack of an independent review organization.**

Ethics review：http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/ethics_en.htm

Yours sincerely

LIU Zhanxiong

LIU Zhanxiong（刘战雄）

PhD Candidate on Philosophy of Technology in School of Humanities，Southeast University，China

Visiting Researcher on Responsible Innovation in Department of Values, Technology and Innovation ，Delft University of Technology，the Netherlands(2015-2016)

WeChant：liuzx199

Skype：liuzhanxiong19

Email：liuzhanxiong19@foxmail.com

liuzhanxiong19@outlook.com

Website：https://seu.academia.edu/liuzhanxiong19

https://www.researchgate.net/profile/Liu_Zhanxiong

Issues related to the IEEE Global Initiative for Ethical Considerations in AI and AS (version 1, 2016)

submitted by:

Jim Isaak (www.JimIsaak.com), no organizational affiliation, IEEE Senior Member, Computer Society President Emeritus, and past VP of the Society on Social Implications of Technology)

General issues:

[all pages] Include a date/version and URL for access to the document(s) on every page in the future number issues/recommendations in a referenceable way… for example: Part 4, section 1, issue 1, recommendation 4 can be paraphrased "resistance is futile" (but more importantly numbered 4.1.1.4)

Add "Harari, Y. N. (2016). Homo deus: A brief history of tomorrow" . As a reference (many sections)

Harari (above) makes a very useful distinction between AI and consciousness. Definitions for AI, AS, and AC (artificial consciousness), as well as AGI and ASI are needed – Since you ask for specifics, I will try to provide some—I know there are better experts at this, but it's grist for the mill:

- AI – a system that makes decisions or recommendations based on accumulated data/knowledge, past experience/learning that maximizes the probability of success in relation to a defined objective.
- AS – A system that uses AI to determine and execute physical actions with little or no concurrent human input.
- AC – A system that has sentient capabilities combines these with AI capabilities and accumulated objectives to make decisions reflecting its own interests.
- AGI – an AI system that can perform any intellectual task expected of a competent human. Such a system should regularly pass the Turing test in any context. Such a system may not qualify as an AC system.
- ASI – an AGI system capable of recursive self improvement. Such a system may not qualify as an AC system.

Shoot away, that's what straw men are for, but without some definitions, we are not providing the basis for critical distinctions. In particular, the recognition of consciousness (and perhaps self awareness) as not necessary elements of various levels of AI systems helps clarify that we are already dealing with some systems that qualify as AI's and probably AS's.

I realize that the IEEE Society on Social Implications of Technology has key leaders involved in this effort, as I hope is the case for ACM and other interested organizations. Where there are issues and/or recommendations that call for the involvement of professional societies, conferences, collaboration among diverse perspectives, etc. I would request that you explicitly include the appropriate societies, and encourage their engagement in addressing the recommendations. We look a bit silly recommending that some un-identified entities should take action, when we are one of the entities, and are hopefully engaging with many if not all of the others.

There are extracted phrasings from this report that should scare the bejibers [term of art, other words are often used] out of informed folks. Which is probably the correct response – recommendation 4.1.1.4 (Pg. 51 see above numbering proposal) – "Ensure that AI systems are corrigible", to which many folks respond "I'm sorry Dave, I'm afraid I can't do that." Is one of these. I suggest an annex that addresses the "real risks" of AI (AGI, ASI, and particularly ASI without AC) – but also addresses the "Crichtonization" of science. Fiction requires dramatic tension, and Michael Crichton was one expert at this. The public view of AI's is more likely defined by Terminator than by Asimov. We need to consider the cultural narratives that have developed in this area, and perhaps be prepared to develop some new ones.

[Pg 15] In the context of adopting references such as the Universal Declaration of Human Rights, we must recognize that AI/AS systems will, in some situations, be designed for valid commercial reasons, even with strong policy and public support that will violate specific components of this document. This includes:

- Article 12 - No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honor and reputation.;

- Article 23 - Everyone has the right to work, to free choice of employment, to just and favorable conditions of work and to protection against unemployment.

The massive relinquishing of personal privacy to commercial (and in some cases governmental) interests is rapidly being targeted by AI applied to Big Data with no regard to Article 12, or perhaps lip-service in 8-point paragraphs a few pages deep in the "click to approve" terms and conditions. If the entities involved in this work are going to "walk the walk", then a few ethics challenges should probably be initiated against corporations, or individuals subject to the BCS/ACM or IEEE code of ethics. Or we can continue the policy of "Advancing Technology for the Benefit of Corporations". (I hate to be skeptical). If we cannot pursue this now and set some examples, there is no reason to recommend that such mechanisms will be any more practicable in the future.

Issues related to the IEEE Global Initiative for Ethical Considerations in AI and AS (version 1, 2016)

submitted by Jim Isaak (www.JimIsaak.com), no organizational affiliation, IEEE Senior Member, Computer Society President Emeritus, and past VP of the Society on Social Implications of Technology

1 General Principles observations

The concept of rights for sentient beings (New Zealand, and others have expanded this beyond humans), and then parallel to this the rights of Artificial Superintellegent conscious entities should be acknowledged, even if it is to state this document will not address these points.

Pg 16 - Missing recommendation: - Educational materials – videos, syllabi, games, etc. need to be created to assure engineers, management, policy makers and the public are aware of the human rights being considered.

Pg 18 – "Proving" is a very strong word in a technical community, and means something different in a court (even something different between civil and criminal courts)

Pg 20 – “flight data recorder” parallel – Access to such data must be transparent, not limited to manufacturer or the government. And the question of “ownership” of such data is very real (consumer, manufacturer, etc.) Finally, manufacturers may have a conflict of interest in how they implement such boxes – consider the recent VW/Diesel emissions testing.

A general comment on the general principles --- “it would be nice”. It is totally unclear to me why corporate, for-profit entities with short term stockholder demands, and protective legal advice will willingly adopt most of these recommendations.

Issues related to the IEEE Global Initiative for Ethical Considerations in AI and AS (version 1, 2016)

submitted by Jim Isaak (www.JimIsaak.com), no organizational affiliation, IEEE Senior Member, Computer Society President Emeritus, and past VP of the Society on Social Implications of Technology

2 Embedding Values in AIS (why did we change terms here?)

Pg 22 single AI/AS’s will affect multiple communities at once – many already span the globe, Google, Facebook, Siri, etc. and interact concurrently with a diversity of cultures.

Harari’s book (Homo Deus, already recommended as additional reference) points out three dramatically different “Humanist” value perspectives. One places the individual at the top – “liberal democracies” being an example, others place the community at the top (“communist” governments for example), and others place “survival of the fittest” at the top (Nazi Germany for example.) Variations on these exist today, perhaps not fully informed by the UN Human Rights declarations. Which should AI’s implement? What transparency is practical for systems built in one culture’s value system to implement the values in the consuming culture?

Pg 28 – (after numbering/separating recommendations above this point) – add a recommendation: Develop use case ethics scenarios, such as the "Trolley problem" applied to self driving vehicles, publish these on the web and encourage crowd sourced feedback that discloses cultural/national/religious etc perspectives – all to provide some insight to developers that care to become aware of these diversity of perspectives.

Pg 33 – "pro forma" Creative Commons like disclaimers of conditions of use may be useful to assure transparent disclosure of liabilities/limitations – pages of mouse time in a "shrink wrap" license (or web-click) are serious transparency issues.

3 Methodologies to guide ethical R&D

(My spell checker wanted to change this to Mythologies .. perhaps there is more AI in there than I suspected)

I hate to suggest this area is set for failure, but… Let me reword page 36 for you:

"Regressive companies, rejecting values based design, will benefit from: lower costs, shorter time to market, more rapid growth, and greater freedom in externalizing costs."

Now, how are we going to make the pitch for our perspective?

Ethics is not enforced or valued by policy, professional societies (when was the last publicly visible IEEE ethics challenge?), licensing, consumer awareness, purchasing processes, etc. In short, there are little or no incentives in most of the global economies to invest in, or defer growth or income to accomplish these objectives.

Pg 37 – Recommendations, add: The introduction of case studies, scenarios, role playing, online games, and quizzes can help support faculty in engaging students with ethics in degree programs. Extracurricular activities might also be promoted, encouraging the formation of "Futurist Clubs", or just student chapters of IEEE SSIT, that draw in technologists, but also liberal arts students, etc. Events like Socrates café's, suggested activities, or online webinar/discussion groups can facilitate focused consideration of the issues. Finally, forming social media interactions on major platforms could raise awareness on a broader basis. Perhaps we could draw IBM Watson into the discussion—to both learn from the dialog, and also engage students in a different way.

Pg 38 Recommendations, add: "IEEE Society on the Social Implications of Technology is one example of professional society hosted forum for education and dialog in this area. Building visibility for this and other such communities, and encouraging collaboration among them can facilitate education in this area. Such groups have often been the "step children" of their parent organizations, we are entering an era when these need significant investment, visibility and support to bring about the desired impact.

Pg 41 – Background should acknowledge that at times industry discourages investing in values based design explicitly or implicitly by putting pressure on function and time to market.

Pg 42/43 – Code of Conduct proposal. This only works if it is actually used in ways that are visible to and/or impact industry leaders and affected professionals. When was the last time BCS had a visible debate or action based on their Code of Conduct. Let me suggest one: the NYTimes and Wall St. Journal describe the use of big data and AI like capabilities to the disadvantage of the rights of 3rd parties and with explicit discrimination against US political parties by a UK organization. This was the "project Alamo" investment using Facebook (and other data sources) to influence the U.S. 2016 Election. It is an abuse that Hamari anticipated in his book Homo Deus (went to print before the election). It is controversial – but any real impact area of AI abuse is likely to be controversial. If ethnics charges are not raised, either in practice (directed at specific technologists) or as "moot court" examples from this type of situation, there is no expectation that future code of conduct/ethics will have any impact on future abuses.

· https://www.bloomberg.com/news/articles/2016-10-27/inside-the-trump-bunker-with-12-days-to-go

· https://www.wsj.com/articles/inside-donald-trumps-data-analytics-team-on-election-night-1478725225

· The Secret Agenda of a Facebook Quiz - The New York Times, Nov 19, 2016

Also, IEEE Code of Ethics is relevant as well (I hope)

Pg. 47 – use of black-box components. This has two meanings. Daniel Dennett, in his 2017 book "From Bacteria to Bach and Back" asserts that science is moving out of "intentional design", into areas of "Black Box Science", where genetic algorithms, and similar techniques are being used to 'discover' (evolve?) results that the researcher's could not directly design, and may not be able to explain. He refers to our entry in a "post intelligent design" world. This implies research and results that are not transparent, not traceable, not explainable by the human designers, and very likely to have unanticipated consequences. This section of the document addresses this meaning of "Black Box" – a reference to Dennett's book might be appropriate at least.

In other parts of the document, the concept of the "airplane black box recorder" is used in terms of monitoring the behavior of an autonomous system. Some clarity of distinction between these two is needed. It is interesting that the "monitoring" black box can probably not be incorporated in designs developed by the "evolutionary" black box.

(From Bacteria to Bach and Back: The Evolution of Minds; Daniel C. Dennett W. W. Norton: 2017)

4 Safety and Beneficence of AGI, ASI

As mentioned in my general comments, with straw-man candidates, definitions are needed here, including one for "consciousness" since it is important to observe that this characteristic may not be needed for either AGI or ASI.

Pg 49 "sufficiently capable" suggests a tipping point exists, may be reached and passed unobserved. This probably needs to be made more clearly.

Pg 50 – Other sections use the term "provable", here we need to encourage both the text, and researchers in these areas to use terms like "probable", and perhaps even try to assign some initial values to the probabilities. The "gain control over its reward channel" immediately raises awareness of the Kobayashi Maru scenario in Star Trek for some of us. A sufficiently capable AGI with a focused goal should, Kirk-like, be able to revise the rules of the game (perhaps this is Turing Test 2.0?)

Pg 51 – Corrigible systems – "*I'm sorry Dave, I'm afraid I can't do that*" (2001: A Space Odyssey) –right. Just after we develop corrigible humans? I'm disturbed by what I consider the "Crichtonization" of technology, where movies and novels need to create dramatic tension by casting technology into the role of a threat to humanity. At the same time we must also consider that possible paths and assign (and adjust) probabilities to these. I have suggested an annex/appendix to this document that assesses the real AI risks, and this section is the one that needs to have some of the concepts expanded in that section. If we can't document some of the concerns and discuss them, we cannot expect to have educated technologists, venture capitalists, etc. that are not blinded by profits, or informed by fiction.

Pg 52 – the concept of "span/impact of control" needs to be introduced here. An AI system that is involved in health and safety is a greater concern than one generating new novels. Part of the recommendations might be to expand Software Engineering licensing (well, encourage this for starters) to include a professional education and evaluated segment in this area. Blaming the C language is an error (and we should not have technical flaws in our examples), any C complier could create strings that are preceded by length indicators. The failure is in the read and assignment operations that do not include length limitations. (It is disturbing and informing to see how long this attack vector has existed without fairly minor changes in operating systems and or libraries that preclude buffer overflow.)

Pg 53 Recommendations – Expand professional licensing to include an introduction to the issues, and continuing education on the issues of ethical and safety issues for AS/AI systems. High visibility is needed for early violations of codes of ethics/conduct/licensing in these areas so that technologists, managers, investors realize that this is not going to "just float by" the way so many past violations have. (Did VW feel enough pain with the diesel air quality tests that it will discourage future abuse? .. is there a parallel in the AI world?)

Pg 54 – The "impacts" proposed here, which are valid IMHO, are inconsistent with the UN Human Rights declaration – (loss of jobs, privacy, etc.) – and to put it bluntly, we need to get a more realistic view of human value for the 21st century. Hamari's Homo Deus points out some of the issues with the rise of the "useless class", and possible transition to "evolutionary humanism" replaces "survival of the fittest" for "benefit to humanity". We need a new set of cultural narratives to accommodate the new age. The newly minted “anthropocene age” needs to be replaced with the “compupocene age” (and who/what names the next one is up for grabs.)

A defense and support fund will be needed for technologist whistle blowers. Retribution will happen, all past experience indicates this. If we expect technologists to risk their careers by disclosing the truth we must go beyond the IEEE Society on Social Implications of Technology Barus Award, and similar tokens, to real support.

5 Personal Data and Individual Access Control

Add to intro: “Big data in conjunction with AI techniques are being used to manipulate persons on the individual level with greater ‘knowledge’ about each individual than they may have about themselves.” (Reference Humani, Homo Deus; and the sources I listed related to section 3 on the use of such manipulation in the 2016 US election.) The Informed Consent horse has left the barn, what waiver in eight point type, twenty paragraphs deep in a “click accept” T&C is going to be informative about how 3rd parties may aggregate, integrate and abuse your personal data to inflict changes in your actions?

Pg 57: We need a new human right, “The right to know you are being manipulated” – I suspect a few AGI folks working with Facebook and other platforms can cause such a right to be adopted by the UN in a reasonable period of time, and they will think it was their own idea. (And would this be ethical?)

Pg 60 Recommendation addition: Laws must provide for individual and class action suits recovering significant incidental damages for the unanticipated abuse of personal data in manipulation of individual’s actions. This does create a constraint on marketing/sales and political campaigns, but it is hard to envision any “finding” and “penalty” that can actually affect abuses in this area.

Pg 63 – It is unclear how an individual can provide informed consent before an exchange of personal data takes place when the real power of personal data is obtained by aggregation from multiple sources, inference about identify, integration into manipulative action campaigns by third parties that may be non-obvious to the data collector and in jurisdictions remote from the transaction. (Consider Cambridge Analytica's use of Facebook psychological profile data in aggregation with many other sources to develop voter suppression campaigns.) See also pg 64 and the observation (accurate) that inferences can and will be made.

Pg 65, Add "Harari, Y. N. (2016). Homo deus: A brief history of tomorrow" As a reference, he suggests how AI's may provide better answers for your personal life than you can, from mates to selection of political candidates.

6 autonomous weapons

Pg 68 "ethical recommendations are needed to prevent these…" change "prevent" to "discourage" .. lets get real, many actors in this domain intend their actions to be covert and attributed to others.

Pg 69 – IEEE will need to consider if it will "Walk the Walk" here … how can any technologist have serious concern for a code of ethics that is not visible in its application? Even if there are not explicit actions visible about an individual, there can be Moot Court type discussions and educational programs that make the issues visible and provide some guidance.

Pg 71 – explicitly include in stakeholders and concerned third parties the persons/institutions that might be targets in authorization scenarios.

Pg 72 – identifiable systems. As pointed out, actors in this area may wish to not be identifiable. But, to minimize anonymous "re-use", of items created here there is a need for both "external" and "internal" standards/codes. For example, having a known standard code indicating that this system was created by XYZ corporation (date/model) is likely to be abused by covert players; so ethical manufacturers will need to include obfuscated 'fingerprints" as well, if they are willing to face the implications of unanticipated abuse.

Pg 77 new recommendation – Attention must be paid to the disposal/end-of-life treatment of devices to assure they do not pass into the control of unauthorized persons.

7 Economics/Humanitarian issues

Pg 80 – There needs to be recognition here that the "period of dramatic change" is not temporary, it is a continuous and accelerating thing. In effect, for those who have not encountered a problem with tracking changes yet, a time will come when they do.

Pg 80/1 the potential for great value by creating holistic solutions has a counter part of the value of creating differential advantage by creating exclusionary solutions.

Pg 82—need to define (as indicated elsewhere) AI, AGI, ASI, and artificial consciousness to facilitate public understanding. Much of the perception of these will be created by fiction, where dramatic tension tends to overplay some risks (and ignore others).

An online community of interested persons is needed (open to public, could be on social media), and the fact-checking, etc. needs to be actively involved in presenting and responding to public perceptions. This process must be sufficiently independent of, or include with checks and balances, representatives of industry, government and academia where folks have vested interests (of course almost all experts will have vested interests.)

Finally, this is a role that IEEE's Society on Social Implications of Technology has played in part, and is expanding – not that they are the only show in town, but providing funding and support for such groups is critical to their success.

Pg 83 – “Robots are taking jobs” – this message should be clear – Also, “new jobs created by robots are likely to require significantly different skills and educational background than those replaced by robots. Significant unemployment and under-employment can be expected.” Separate recommendation – national labor departments should be actively tracking job displacement by automation, and also jobs created by automation (along with required skills) to be able to provide a data-based evaluation of these assertions.

Pg 85 – “ensure equitable distribution of the benefits of …” I’m sorry, but this is balderdash. Income, wealth and other inequalities are growing rapidly, fed in part by the existent AIs, robots and technologies, and even more by the investment of those in the 1% to tip the scales further. With input from experts like Prof. Robert Reich, UC Berkeley (and past US Labor Secretary), you may be able to come up with a pragmatic recommendations. For now, the Background probably should read: “The benefits of AI/AS technologies world wide is driving increased inequality by almost every metric. It is unclear what path can lead to a more equitable application, benefit and impact from these technologies.”

Also – on personal information – “Harari, Y. N. (2016). Homo deus: A brief history of tomorrow”. As a reference. He points out that AI’s combined with big data can (will?) displace personal decision making for a number of reasons, not the least of which is the loss of any viable personal data.

Pg 87 – The advent of AI/AS can will exacerbate …. (Change word)

8 Law

Pg 90 – who owns logs, has access to them, etc. is critical – black boxes and logs are currently encouraged as an aspect of liability limitation, and also destroyed or not allowed for the same reasons.

New Committees

Pg. 96 – “black box science” issue – this is both real, unavoidable, and the first indication of the “Singularity” (which is a concept that should appear in this document even if not used as a working term)

Also: cultural bias – I suggest adding the term "Humanism" and reference Homo Deus related to this concept definition in a cultural context. (citation in prior materials)

Anthropomorphic approaches – uses the word "patients" I suspect this is not the intention, this paragraph has a "conclusion" which is inappropriate for an "issue"… "Is implanting human morality or emotions into an AI useful, effective, appropriate, …." Don't assume it is not.

The vocabulary, and perhaps more critically the "cultural narratives" (which is my take on Homo Deus, in part) are certainly not shared beyond select groups. The Science Fiction narratives have more visibility than the philosophic ones.

Pg97 – Mixed reality – does not require virtual reality – we have this in video games, and even our 'broadcast' TV at this point. The initials "MR" are used without the prior indication.

I love the "Eradicate the positive effects of serendipity", there are many relevant variations, some are faith based (God's plan for you as revealed…) some are Psych based (Tesla's revelation of the way AC generators might work) etc.

The connection between body/mind might as well admit that the issues here include the nature of "consciousness" and "free will" – Homo Deus, among other sources, suggests that humans will find "meaning" disrupted when the absence of both "Free will" and "self" become scientifically evident. (Of course they will have no choice …)--- Perhaps more relevant here is the issue of AI/Big Data to manipulate persons in subliminal and/or unconscious ways, over-riding what free-will they might have.

Issue 3 has a reverse question- what if anything allows AI to control a person?

Pg 98 – many of the issues raised in this section are addressed in various issues of "Technology and Society" from the IEEE Society on Social Implications of Technology – which should be acknowledged.

Pg 99 – concern for catastrophic loss of human autonomy – glad to see it acknowledged. Homo Deus, Spiritual Machines and other books should be references here. Issues like machine mis-representations, intimidations, etc. are important concepts to surface. People already assign computers with a degree of infallibility. … this can only get worse.

Pg 100, the word "inserted" might be better said "interact" … an AS (affective system – additional definition needed?) will interact across and between cultures, not just within them. We already have this with Siri, Google, Alexa, and my car's incompetent voice recognition.

Issue 6, systems manipulate emotions to alter human behavior. This should explicitly reference the "Project Alamo" activities in the 2016 US election. It is an abuse that Hamari anticipated in his book Homo Deus (went to print before the election). It is controversial – but any real impact area of AI abuse is likely to be controversial

- https://www.bloomberg.com/news/articles/2016-10-27/inside-the-trump-bunker-with-12-days-to-go
- https://www.wsj.com/articles/inside-donald-trumps-data-analytics-team-on-election-night-1478725225
- The Secret Agenda of a Facebook Quiz - The New York Times, Nov 19, 2016

Also, IEEE Code of Ethics is relevant as well (I hope)

Pg 101 – Policy making connection to EpicAI –this is a topic that IEEE SSIT is undertaking (in part, not as a primary focus, and not as an exclusive focus, but the interests of related groups should be recognized and hopefully we can collaborate to bring greater value to the process.

Issue 1 – how to help public service entities? --- a conference and tradeshow (revenue source) related to this would be one good idea … I can envision this as a collaboration between SSIT and other groups.

Pg 102 – Issue 3 – Help facilitate discussions between policy wonks and nerds. … some rewording might be warranted. While "influence" is a role for groups like IEEE USA, the "informing", tracking and ongoing dialog about emerging issues is at least as important, and It is not a US-centric problem – in fact, multi-national engagement on understanding of the issues is important to respect culturally specific approaches to addressing these as well as educating the technologists and other stakeholders.

Alexis J. Valentin,
The Secretary,
WhyFuture AI Concepts,
www.whyfuture.com

Designing a strong-AI is like having a vessel master of a ship of passengers, and whether that ship is on course to the passenger destinations or not is up to how we initially design that strong-AI.

There have been a lot of advancements that have been made with regards to artificial intelligence over the passage of time. As a matter of fact, the artificial intelligence research field has over time been coming up with massive features which are yet to be regarded as AI by the masses. Such features are inclusive of a number of existing online accomplishments such as the use of virtual agents, pattern recognition and targeted advertising as well (Martin, 2015). As such, it is a clear fact that AI plays a major role in today's society and therefore it is important to ensure that we are in a position to cope with all the advancements made in this sector through obtaining a deeper knowledge regarding the processes involved and their importance (Martin, 2015).

The most essential objective when it comes to the accomplishments made in terms of artificial intelligence is inclusive of the need to form an intelligent machine that has the ability to perform a number of functions. These functions are inclusive of their ability to be cognitive as well as rational, to plan, cracking glitches, grasping even the most intricate concepts, fast learner as well as being capable of learning from its past thus becoming better with each passing day. This amounts to the generally accepted description of the human intelligence (Martin, 2013).

The AI should be developed in such a manner that portray an extensive and profound aptitude to understand its environments for purposes of establishing what to do in the different situations that they are likely to come across. This further means that for the AI to be in a position to get a clear comprehension of its environment and how to respond to these different possible situations, then a need arises for it to be socially intelligent as well.

It also needs to have the ability to be creative since creativity comes in handy when it encounters a situation or situations that require its finest skills with regards to management of problems.

For purposes of realizing all the above mentioned attributes, it is important to take certain factors into consideration. The first of these factors is the need to look into the traits of altruism vis-à-vis those of psychopathy. It is important to look into the human altruistic behavior and make a thorough evaluation in order to be able to profile artificial intelligence around qualities that are considered as humane as well as philanthropic values.

This means that thorough studies ought to be conducted for purposes of exploring the deepest and most intricate foundations of human altruistic behavior. Other factors that ought to be taken into consideration are inclusive what is generally needed in order to conclude that a person is altruistic as opposed to a person who is not altruistic. In general therefore, when designing AI it is imperative that it is shaped around the best and most positive traits of people (Why Future Website, 2016). This therefore encouraging traits such as compassion, generosity and equality among others.

The second factor that should be taken into consideration is the aspect on ethical dilemma also referred to as the ethical paradox. This is where a situation presents itself and there is a need for the IA to choose what action to take between being diligently efficient or sticking to their moral obligation. This brings in the issue of psychopathy vis a vis empathy. In as much as artificial intelligence ought to be shaped in a manner that makes it very efficient, this should at no time beat the ability for it to be empathetic when the need arises.

AI ought to be designed in a manner that it can be able to instantly opt out of being efficient in order to have compassion towards someone or people depending on the situation at hand (Why Future Website, 2016). It is important therefore for persons designing AI to be able to structure AI’s in accordance with their defined moral systems as well as the manner in which they are supposed to position themselves depending on the different moral cases that they find themselves faced with (Martin, 2013).

The next factor that ought to be looked into is the one dimensional perception vis a vis the multidimensional perception. It is through our perceptions that we ultimately have the ability to critically evaluate the different situations that are presented before us. This subsequently influences the manner through which we behave thereafter. Perceptions are what determines the course of actions we take regarding different situations. It is important therefore for those charged with the responsibility of designing AI to shape it in such a manner that it will be able to properly determine when it is right to follow a one-dimensional perception as opposed to multidimensional perception and vice versa (Why Future Website, 2016).

References

Martin, (2015). Artificial intelligence: A complete guide. Retrieved from https://www.cleverism.com/artificial-intelligence-complete-guide/ last accessed on 16th February, 2017.

Why Future Website (2016). The Blueprints towards the development of good artificial intelligence. Retrieved fromhttp://www.whyfuture.com/single-post/2016/11/06/The-Blueprints-towards-the-Development-of-Good-Artificial-Intelligence last accessed on 16th February, 20

Charles H. Jones, PhD
C. H. Jones Consulting, LLC

1. P32 Trust. The discussion and recommendation focuses almost entirely on transparency. But this is not how most humans develop trust in things or people. Most of the results of this recommendation would be meaningless to most people. (Although they need to be done.) Trust is earned through consistency of action and not causing harm; the person or machine does what I want or expect every time. Perhaps some mention of the need for consistency and predictability should be included. Should have a psychologist and sociologist on the committee to address this.

2. There is an elephant in the room that this document ignores: sentient or self-aware AI. At best the analogy of the blind men and elephant might apply in that this issue is tangentially alluded to in different places. Most glaringly is P6 Section 3. The statement "Machines should serve humans and not the other way around." presupposes that AI will never reach the point of sentience deserving equivalents of human rights. This statement is also contradicted by one of the references "Sometimes it's hard to be a robot" by Whitby. Another place where the document dismisses this issue is P51 where "boxing" is mentioned. And the mention of "unintended behavior" on P49 trivializes this issue – both in terms of the difficulty of boxing and in terms of the ethics of putting a self-aware being in a box. The rights of sentient AIs is one of the biggest ethical issues of AI and this document dismisses it. This needs to be addressed long before it actually happens. This issue deserves its own major section.

3. P6, Issues - The statement "Values to be embedded in AIS are not universal" is too absolute. In fact, this statement is contradicted by references to the Common Good Principle and the Universal Declaration of Human Rights. There really are some universal values. Perhaps: "Not all values to be embedded in AIS are universal…" And the candidate recommendation (P24) needs to be rewritten to be more nuanced. For example, recognizing that not all AIS, such as task specific systems, need a complete set of values.

4. The following should be added to the issues under Section 4 P7: "Humans may (inadvertently) command AIs to do unsafe things. When should an AI say "no" to a human?" This issue is indirectly addressed in other sections and indirectly in the first issue of Section 4. However, the tone of those discussions regards autonomous actions emerging from the programming rather than from human commands. If this is not added as a separate issue, it should be addressed more directly in the discussion on P50.

5. P26 Issue regarding built in biases. Somewhere in this discussion their needs to be recognition that a built in bias can be *for* a particular group instead of against it. The recommendation should mention the need for balance. When does an action intended to aid one subgroup cause harm to the rest of society? Another aspect of this is the mention of "target populations". How does the AIS recognize when it is interacting with people outside this population?

6. There is no discussion on the use of avatars representing a person (e.g., using an AI to answer the phone or to automatically respond to an email.) A specific issue is when is it OK for an AI to lie? Humans do it all the time, is it OK for an AI to make a decision to follow such an example because the human it represents does? This might fit under subcommittee 2 on p99 but it deserves some discussion in this document.

7. There is no discussion of AI to AI interaction and how it might affect people. For example, what if someone has a robotic companion with all the emotions associated with such a relationship and another robot hurts the companion? We've already seen the human outcry when a human knocks a non-AI robot over. But this AI to AI interaction might very well include hidden activity such as financial exchanges or a host of other actions that directly affect humans.

8. P96 The statement "The attempt to implant human morality and human emotion into AI is a misguided attempt to designing value-based systems." is highly debatable. The reference to human morality contradicts most of the rest of the document. A fundamental assumption in this document is that we are trying to instill human values into AI. (This should probably be stated explicitly.) It is incorrect to suggest human values are not closely aligned with human morality. A growing body of science supports that human decision making is based on both mind and gut – both reason and emotion. Emotion generally provides a quick

response to something whereas reason allows us to verify and sometimes override an emotional response. The suggestion that a full set of human values can be developed without emotional input is naïve. This is also another example of the dismissal of self-aware AI. Do we truly want to develop beings that have no capacity for emotion? At best this discussion should center on when morality and emotion are and are not appropriate for AI design. But there is also a question as to what non-human values are appropriate for AI.

9. P5. The second general principle mentions the "natural environment". Yet this is never discussed again. It is appropriate to include this general principle but that implies it should be directly addressed in the document. An example where this might be added is P6 at the end of the summary of Section 3: "…for business, society and the natural environment." But it deserves to be addressed more completely and directly in general.

10. No glossary. I understand arguments for not having one, but this document does not even define AI or robot. So, in some sense, this document has not defined its scope. There are also a lot of technical and ambiguous terms. A truly ambiguous word is "honor". For example, several religions say to "honor" woman but then elsewhere imply woman have no rights.

11. We tend to criticize and not compliment. The document needs work, but overall this is an excellent draft that outlines many of the issues with some good recommendations.

Charles H. Jones, PhD

C. H. Jones Consulting, LLC

Comments for IEEE EAD V1

Jia He , IEEE Global Initiative China Committee member

https://www.linkedin.com/in/jia-he-54680018/

My comments about the Executive Summary

I was leading a workgroup to translate the executive summary of IEEE EAD V1 into Chinese, which aims to engage the Chinese community into our global initiative. From the translation work, I have the following takeaways to share with all of you:

1) The executive summary is very important for the people who are interested in the paper but have limited time to read the whole paper with hundreds of pages. The problem of the executive summary of EAD Version 1 is the imperfection of the contents. Actually the paper has two key parts of content for each section: issues and recommendations. While, only issues are included into the summary. So **my suggestion is to summarize the recommendations for each section into summary too.**

2) Two points of view need to be re-think about.

Part 3: Methodologies To Guide Ethical Research and Design said, The modern AI/AS organization should ensure that human wellbeing, empowerment, and freedom are at the core of AI/AS development. I agree that AI/AS development should respect the values of human society. However, different country has different values. Wellbeing, empowerment, freedom may not the common values which are broadly applied to all the countries. **Given culture and culture diversity, my suggestion is to do further research on it, and find out some common documents which have already been agreed universally around the world.**

Part 8: Law said, how can we ensure that AI is transparent and respects individual rights? For example, international, national, and local governments are using AI which impinges on the rights of their citizens who should be able to trust the government, and thus the AI, to protect their rights. The issue is fine, but the example has problems. First, there are four issues in the part 8, but only this issue gives an example. In order to keep the paragraph consistent, **my suggestion is to provide example in the main body of the paper, rather than putting it at the summary.** Second, one of the goals of this initiative is to raise up the awareness of people for ethics of AI. Government is one of the important stakeholders that we hope it can be engaged into the initiative. While this example actually is blaming governments to some extent. It's not necessary for us to blame any stakeholder at this paper. We can describe the risk and concerns instead. **My suggestion is to change the way of description for the example, and make it more friendly to engage the government and other stakeholders into this issues.**

3) A sentence is advised to be moved to another section.

The sentence of "There is a lack of access and understanding regarding personal information" in the Part 7 Economics/Humanitarian Issues should be moved to Part 5 Personal Data and Individual Access Control. Because it's about the personal data issues.

My comments for the Part 7 Economics/Humanitarian Issues

Part 7 raised up the issues of increasing of active representation of developing nations in The IEEE Global Initiative is needed. My suggestion is to add **3 recommendations.**

Why it's difficult for developing countries to participate in the discussion of our committees?

1) One of the reasons is that AI and autonomous technologies are not equally available worldwide. Developing countries are less developed on AI than developed countries. **My suggestion is to initiate a training program by which AI courses are provided by developed countries, or an expert exchange program by which the people between developed or developing countries have the opportunity to sit together to know each other.**

2) Second reason is the language problem. English is our working language. While many people from developing countries are not English native speaker, such as the people are from China, Japan, and Korea. They will take much more time to read and understand the paper. **My suggestion is to build up committees at developing countries for enhancing engagement.** For example, a Chinese committee was built up in March 2017. The committee aims to introduce IEEE EAD V1 in Chinese community and engage Chinese institutes or individual experts into this IEEE Global Initiative via translation, workshops, and online communication groups. The committee was organized with the similar principle as IEEE – inclusive, openness and professional. Obviously, we see the comments from Chinese community are increasing.

3) Another reason is limited budge to participate in the activities. Why not launch a Global Initiative Ambassador program. IEEE can select and support an ambassador to do outreach work in the country and pay the travel fee for the ambassador to the meetings such as Austin meeting in May. **Another suggestion is to convene those kind of meetings at developing countries.** IEEE can select some partners from developing countries to co-host those kind of meetings.

4) One more issue and recommendation could be added into Part 7 Economics/Humanitarian Issues.

The risk of unemployment for developing countries is more serious than for developed countries. You know that the industry of most developing countries is labor intensive. More and more jobs will be gradually replaced along with the development of robots or AI. This will not only happen at the manufacture industry, but also at the service industry. For example, if driverless car can service you well, drivers will loss the jobs, and if machine knows how to write news, the amount of the employment of journalists and editors will be reduced. The challenge of unemployment is even bigger for developing countries than for developed countries, which can exacerbate the economic and power-structure differences between and within developed and developing nations as we mentioned.

Actually it's necessary to have some researches made from now on. However, few organizations have the budge or motivation to do those kind of research because the benefits are far away from now. **My suggestion is to propose that the responsible AI companies should make some efforts on those kind of researches as CSR, because reducing the social problems of technology development should specially be done by a responsible AI companies. There are many methods to do the CSR work, including doing this kind of research inside of the company, or commissioning NGOs or third part research centers to do it.**

Comments & feedback

Version 1 of Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems

By Ansgar Koene, Senior Research Fellow at the Horizon Digital Economy Research institute, University of Nottingham, UK

Page 2 Executive Summary, 2nd paragraph, 2nd sentence: AI/AS have to behave in a way that is beneficial to people **and society** beyond reaching functional goals and addressing technical problems. – add "and society" to distinguish from system that exploit some people for the benefit of others. Also matches better with the fourth paragraph.

Page 6 Methodologies to guide ethical research and design, list of Issues: directly related to the issue of "The need to differentiate culturally distinctive values embedded in AI design" is an **additional issue of "Lack of clear identification and documentation of contextual scope limitation of AI/AS designs**".

Page 7 Safety and Beneficence of AGI and ASI, list of Issues: the final listed issue "Future AI systems may have the capacity to impact the world on the scale of the agricultural or industrial revolution" is simply a restating of the first sentence of the committee description text. It does not identify a clear issue. I would suggest removing this from the issues list.

Page 7-8 Personal Data and Individual Access Control: Two additional factors that should be mention in relation to this topic are, 1. **User trust, which is undermined when users can no know what personal information is help about them or whether is information is correct.** 2. **Trade in personal data, which is a powerful mechanism for acquiring necessary data for certain services, but which severely impact the ability of users to know who is holding what kind of data about them.**

Page 8-9 Reframing autonomous Weapon Systems, list of Issues: Important issues that is missing from the list are: 1. **How to deal with dual-use autonomous systems that can be used for both civilian and military purposes, e.g. chemical dispersal drones for crop-dusting.** 2. **Potential for rapid weaponization of civilian autonomous systems, e.g. mass hacking of autonomous vehicles**.

Page 18 Candidate Recommendations: Connected to the issue of diversity of cultural norms (recommendation 2), **designers should take into account the context dependent nature of norms and responsibilities, providing explicit boundaries for the scope of applications for which the AI/AS system has been developed.**

Page 21 Education and Awareness, Candidate Recommendations: This section needs an addition 4th paragraph recognizing that Education and Awareness is a two-way issues. Not only the public needs to be educated about AI/AS, but AI/AS developers also need to learn about the concerns and implications of their systems on citizens. This will require engaging with citizens using methods such as those developed for Responsible Research and Innovation (RRI) in ICT.

Page 23 Embedding Values Into AIS, possible addition: **A stakeholder-inclusive approach to the values embedded in the system will also help designers to become more aware of the implicit values they may be unconsciously embedding in their system.**

Page 28 Embedding Values Into AIS, Further Resources: Many of the important recommendations regarding inclusion of target populations in the design process are a core part of the Responsible Research and Innovation framework used by the RRI in ICT community, e.g. http://www.orbit-rri.org/

Page 29 Embedding Norms and Values in AIS, addition to end of third paragraph: It will be important for AI systems that learn human values and norms through bottom-up approaches to be able to **communicate which norms/values they have acquired in order to guard against undesired outcomes due to 'falling in with the wrong crowd' (e.g. Tay bot)**.

Page 39 Methodologies to guide ethical research and design, further resources: This might be another place where it would be appropriate to reference the RRI in ICT methodologies http://www.orbit-rri.org/

P46 Methodologies to guide ethical research and design, middle of first paragraph Tutt: "algorithm FDA" should include a reference to the paper by **Andrew Tutt "An FDA for Algorithms" published in Administrative Law Review, Vol. 67, 2016. Available at SSRN: https://ssrn.com/abstract=2747994**

Page 48 Methodologies to guide ethical research and design, additional candidate recommendation: **Developers of black-box components that are submitted into software libraries should provide clear documentation regarding the context/use-cases for which the system was initially designed and validated to go along with their code**. Where possible this should also include indications about expected limits of safe/ethical use.

Page 56 Personal Data and Individual Access Control, 2nd paragraph 2nd sentence: In addition to the process of data gathering there must also be more transparency and control over the trade in personal data.

Page 65 *Personal Data and Individual Access Control, Candidate Recommendations - possible recommendation to add:* **As part of the ability for individual to see which personal data is being held about them, they should also be able to get access to seeing which inferences have been made about them based on this data.**

Page 75 Reframing Autonomous Weapons System, left column: Paragraph 1 and paragraph 4 appear to be the same. Recommend removing paragraph 4.

Page 84 Economics/Humanitarian Issues, proposal for additional Issue: A need for new success metrics for AI/automation innovation.

Background – Successful innovation is primarily measured in terms of early 20th century metrics of time, energy or labour cost efficiency. The humanitarian dimension of innovation, especially important as AI/automation is applied to human-facing services, is often undervalued.

Candidate Recommendation – develop new human service experience related metrics, possible related to well-being or 'happyness index' metrics.

Page 91 Law, Background, possibly add to the last sentence: "**including decisions about filtering the information that a human is given for making the decision.**" I would like to propose this addition due to the risk of missing legal safeguard on AI transparency in system where the human is effectively 'rubber stamping' a decision that made by an

Submission by Eileen Donahoe, J.D., Ph.D. (Ethics)

Executive Director, Global Digital Policy Incubator

Stanford University Center for Democracy Development and the Rule of Law.

http://cddrl.fsi.edu/gdpi

eileen.donahoe@stanford.edu

1. General Principles: The articulated goal of the IEEE initiative to embody the highest ideals of human rights in AI/AS and to achieve maximum benefit to humanity is an excellent base for this endeavor. The question of HOW to ensure that AI/AS do not infringe human rights is the correct issue, as is the question of HOW to ensure accountability for effects of AI/AS on the enjoyment of human rights. (p.5) However, this core concept and goal is unintentionally undermined by the next section of the draft.

2. Embedding Values into AI Systems: The discussion on embedding values into AI systems subtly shifts to the search for "relevant human norms" that must be embedded. This move rests on a presumption that universal human norms do not exist. The issues section says this explicitly: "Values to be embedded in AIS are not universal, but rather largely specific to user communities and tasks;" "Moral overload: AIS are usually subject to a multiplicity of norms and values that may conflict with each other;" and "The need to differentiate culturally distinctive values embedded in AI design." (p.6)

3. Human Rights are Universal: The entire draft would benefit from better conceptual understanding of how universal human rights and international human rights law (IHRL) function. The IHRL framework rests on a well-established global norm that human beings have human rights by virtue of their humanity. Governments have legal obligations to protect and not violate the rights of citizens and people within their territory and jurisdiction. The existing body of IHRL articulates a wide range of substantive human rights, which are then made manifest across all different cultures with variation. While the IHRL framework does not require uniformity and homogeneity with respect to HOW these universal human rights are implemented, it does provide a concrete range of rights that are not optional, like freedom of expression and privacy. In addition, IHRL a provides basis for evaluating cultural norms and laws as they impact upon the enjoyment of

human rights. Ethically aligned design in AI/AS must be based on existing universal human rights which are already well articulated in international law. Furthermore, AI/AS design that reflects cultural norms and laws must be evaluated with reference to their impact on the enjoyment the universal human rights of people in those cultures.

4. Global Challenge of Protecting Universal Human Rights in a Global Digital Ecosystem: Governments, companies, international organizations and civil society organizations committed to protecting universal human rights in the global digital ecosystem are also struggling to articulate HOW to apply universal rights in this new context. Several dimensions of the IHRL framework are challenged by the digitization of everything. Two obvious challenges flow from the fact that digital technology facilitates instantaneous extraterritorial reach for everyone anywhere. This in turn challenges governments in their obligation to provide security, and in some cases, is leading even human rights-respecting governments in non-human rights respecting directions. The IEEE effort to incorporate universal human rights in AI/AS EAD may benefit from this larger conversation. The IEEE effort to articulate how to incorporate human rights may also contribute to these other communities struggling with make universal human rights real in the global digital ecosystem.

5. "Lack of values based culture and practices for industry:" (p. 6) Some industry efforts to incorporate human rights values already exist and may be helpful to IEEE's initiative. The IHRL framework places the primary obligation for protection of human rights on governments: Governments are responsible for providing security and protecting human rights of citizens. Yet, in the global digital ecosystem, private sector technology companies have taken on many governance responsibilities. For example, private industry own, operate and secure much critical internet infrastructure, and are now taking on responsibility for defense against "information operations" by foreign governments in the digital information realms. Big digital platforms also effectively govern the "public square" through terms of service, community guidelines and algorithms. A variety of efforts have been made to articulate voluntary responsibilities for private sector companies to respect human rights. The UN Guiding Principles on Business & Human Rights http://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf;

Global Network Initiative Principles to protect freedom of expression and privacy on ICT https://www.globalnetworkinitiative.org/; Voluntary code of conduct for digital platforms functioning in Europe to screen hate speech with AI http://europa.eu/rapid/press-release_IP-16-1937_en.htm. In addition, many private sector companies proactively commit to human rights responsibilities. Microsoft, for example, has engaged in a Human Rights Impact Assessment with respect to its own AI. Microsoft's proactive example could serve as a model for others.

Viola Schiaffonati

Artificial Intelligence and Robotics Laboratory, Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Italy

p. 22: I suggest to provide explicit definitions of the notions of value and norm and to clarify how they interact in this document. This might be helpful also for issues discussed later (see for example on p. 24, where norms and values are not differentiated in the Candidate Recommendations).

p. 30: I suggest to clarify the way in which the highly ambiguous notion of 'moral machines' is used in this context. Does this mean the possibility to build artificial moral agents? If this is the case, the arguments against the idea that ethical decision-making is possible only for human agents must be presented and discussed. If this is not the case, what does 'moral' mean in conjunction with 'machines'? Is it just a form of functional morality or is it different? For a recent debate on this see for example: K. Miller, M. Wolf, F. Grodzinsky (2017) "This 'Ethical Trap' is for Roboticists, not Robots: On the Issue of Artificial Agent Ethical Decision-Making", *Science and Engineering Ethics*, 23:389-401.

p. 62: When discussing Personal Data Access and Consent (Section 2), I recommend to consider not only privacy by design but also current alternative approaches to privacy. Interesting references on this are for example: J. Lane, V. Stodden, S. Bender, and H. Nissenbaum (editors) (2014) *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, Cambridge University Press; B. Roessler, D. Mokrosinska (2015) *Social Dimensions of Privacy. Interdisciplinary Perspectives*, Cambridge University Press.

p. 70/71: The potential for confusion about definitions is mentioned in particular with regard to Autonomous Weapons Systems and it is suggested that to clarify the term autonomy is important for understanding debates about AWS. Why not to extend this approach to the whole document by discussing, for example, the various meanings of autonomy in other potentially interested parts of the document (e.g., Committee Section 2 and 3)? This attitude toward conceptual and terminological clarification could be also important to overcome the issues related to the gross oversimplification of the media mentioned in Section 1 of Economics and Humanitarian Issues (p. 82).

Viola Schiaffonati, Ph.D.

Associate Professor of Logic and Philosophy of Science, Artificial Intelligence and Robotics Lab, Politecnico di Milano

http://www.deib.polimi.it/eng/people/details/70304

Alexandre Sacco Xavier

Organization that I represent: I´m a Master of Science in Information Systems by UFRGS (Federal University of Rio Grande do Sul, Brazil) but currently I'm an **autonomous researcher**.

I´ve listed my inputs by Chapter/Sub-Chapter to make it easier to whom is reading:

General Principles

Principle 1 – Human Benefit

- The chapter is about the Human Benefit underneath there are mentions to Human Rights, Rights of Child, Women, Person with Disabilities, but there´s no mention to the rights of the LGBT community. You should have it fixed since we are in a world in which all kind of diversity should be respected.
- I believe we should ensure in some way that we should not have any kind of religious prejudice in the way we develop technologies in the IA area.

In summary, the first item of this chapter could be re-written as below:

1. AI/AS should be designed and operated in a way that respects human rights, freedoms, human dignity, and all kind of diversities (cultural, sexual and religious).

Principle 2 - Responsibility

I suggest a creation of sub-item 1.1 to say: "All the players involved on the AI/AS production cycle must be aware that they can be considered co-responsibles in case of some harm generated by those products"

Principle 3 - Transparency

In the candidate recommendation, I suggest to add a fourth item as below:

(4) for tracking purposes, in case we need to evaluate why the AI/AS took some decision. On this case we need a very detailed system log, in which we can see each step and each algorithm that made the AI/AS took a such

decision. Since it´s a functionality that can consume too much resource from the system (memory/power) we would have to have a switch/setup to enable/disable it.

Principle 4 - Education and Awareness

As a first concern on this context, we should educate our children to use this kind of technology with ethic and responsibility to ensure we will have adults well educated. To do that we must have the subject each time more included on our schools curriculum.

Chapter 2 - Embedding Values Into Autonomous Intelligent Systems

Identifying Norms and Values for Autonomous Intelligent System:

Issue: Values to be embedded in AIS are not universal, but rather largely specific to user communities and tasks. On this context, I agree with the proposal, but it seems we are bring to the discussion a need of a Ethics Comitee in each company that will produce AI/AS to analyze the scenario in which those products will be inserted and so, take decisions on how to approach. We may need a kind of norms/values setup to be choosen in accordance with the country/region. In this case, that Ethics Comitee would act ensuring the norms/values don´t offend the market in which the AI/AS will be used because it can vary in accordance with many aspects of such culture/country/region.

Issue: Moral overload – AIS are usually subject to a multiplicity of norms and values that may conflict with each other. The prioritization on the multiple norms and values should be done for sure. The system should be configurable as possible to allow the user change that prioritization when it makes sense. For those situations in which this option cannot be allowed, the designer cannot take the decision by him/herself and need to be suported by a team (an Ethics Comitee for instance). Before to go to Production (or be implemented) the AIS should be submitted to a set of tests (functional tests) to ensure the AISs are in accordance with the context in which they are going to be inserted.

Issue: AIS can have built-in data or algorithmic biases that disadvantage members of certain groups. I agree with the recommendation taken and also I reinforce that designers/developer must be very careful to not assume that the AIS will answer the same way to everybody and so, they should submitt the system to all kind of tests with the many different users. This kind of consideration should be taken even for those people with some neural disability (by nature, illness or accident trauma) when the system should be carefully developed to not infer logic questions/answers from that kind of users.

Issue: Once the relevant sets of norms (of AIS's specific role in a specific community) have been identified, it is not clear how such norms should be built into a computational architecture. I agree with the approach recommended here but also I suggest that the AIS should be set to behave differently depending on the use it´s going to have. For instance, in those kind of functions in which the AIS is expected to protect, it should take decisions even with more responsibility than it´s user, which, in some cases, could be children, which from their side, are not expect to know everthing that´s dangerous for them, needing an adult (or an AIS) to protect them. On other hand, if the AIS is supposed to teach someone, it should not take decisions on behalf of the user but instead, help he/she in the learning process.

So far it´s all that I have. If time allows, I will send more inputs in another email.

Thanks for the work IEEE is doing on this area, and to do that on this democratic approach in which everybody can provide feedbacks to have a multiple hands/minds result. Really appreciated approach. Congratulations!

Thanks,

Alexandre Xavier

M.Sc in Information Systems / CSPO – Certified Scrum Product Owner

Frederike Kaltheuner, Policy Officer, Privacy International

Asaf Lubin, JSD Candidate, Yale Law School, Robert L. Bernstein International Human Rights Fellow, Privacy International

IEEE's Global Initiative for Ethical Considerations in the Design of Artificial Intelligence and Autonomous systems

May 15, 2017

Frederike Kaltheuner, Policy Officer, Privacy International

Asaf Lubin, JSD Candidate, Yale Law School, Robert L. Bernstein International Human Rights Fellow, Privacy International

Statement of interest

Privacy International is a non-profit, non-governmental organization based in London, the United Kingdom ("UK"), dedicated to defending the right to privacy around the world. Established in 1990, Privacy International undertakes research and investigations into government and corporate surveillance with a focus on the technologies that enable these practices. To ensure universal respect for the right to privacy, Privacy International advocates for strong national, regional and international laws that protect privacy. It has litigated or intervened in cases implicating the right to privacy in the courts of the United States, the UK, and Europe, including the European Court of Human Rights and the European Court of Justice. It also strengthens the capacity of partner organizations in developing countries to identify and defend against threats to privacy. Privacy International employs technologists, investigators, policy and advocacy experts, and lawyers, who work together to understand the technical underpinnings of novel surveillance technologies, and to consider how existing legal definitions and frameworks map onto such technologies.

I. Introduction

Novel applications and recent advances in Artificial Intelligence and Autonomous Systems have the potential to significantly affect the right to privacy. This is significant since privacy is the lynchpin of both indispensable individual values such as human dignity, personal autonomy, freedom of expression, freedom of association, and freedom of choice,[5] as well as broader societal norms.[6] Some commentators had noted that the right to privacy is in essence the "canary in our technological coal mine".[7] This is why Privacy International welcomes the IEEE's Global Initiative for Ethical Considerations in the Design of AI/AS,[8] which seeks to address this pressing issue. Focusing on the right to privacy, Privacy International wishes to provide, in the following, some general remarks on the initiative, followed by specific commentary on some of the report's key sections.

II. General Remarks

II.A. The Need for a Clear Definition of AI/AS

We very much welcome the initiative to develop principles of ethically aligned design in AI/AS, yet we noticed that the report lacks a clear definition of AI/AS. For instance, AI could be used to operate and control a component of a given system, while intelligent behaviour may be an emergent property of several interacting intelligent entities. It is unclear what the scope of this initiative includes. This lack of definitional clarity is a challenge, since different levels of abstraction and varying degrees of complexity and autonomy, along with the domains in which they are employed, raise specific ethical and regulatory issues.

[5] THERESA M. PAYTON & THEODORE CLAYPOOLE, PRIVACY IN THE AGE OF BIG DATA 1-5 (2014).

[6] ROBERT C. POST, THE SOCIAL FOUNDATIONS OF PRIVACY: COMMUNITY AND SELF IN THE COMMON LAW TORT, 77 CAL. L. REV. 957, 961-978 (1989). Summarizing Post *see* DANIEL J. SOLOVE, NOTHING TO HIDE: THE FLASE TRADEOFF BETWEEN PRIVACY AND SECURITY 50 ("As the legal theorist Robert Post has argued, privacy is not merely a set of restraints on society's rules and norms. Instead, privacy constitutes a society's attempt to promote civility. Society protects privacy as a means of enforcing order in the community. Privacy isn't the trumpeting of the individual against society's interests but the protection of the individual based on society's own norms and values").

[7] Payton & Claypoole, *supra* note 2, at p. 1.

[8] The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems. *Ethically Aligned Design*: A Vision For Prioritizing Wellbeing With Artificial Intelligence And Autonomous Systems, Version 1. IEEE, 2016. http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html.

For instance, the report's principles and examples cover a diverse range of applications and use-cases at different levels of complexity and abstraction - from machine learning to making inferences about individuals, and other domain-specific AI algorithms, to fully autonomous and connected objects. This lack of clarity is most apparent in the chapter on general principles, some of which seem to have been crafted with AS in mind, but do not necessarily translate to cases where, for instance, AI is employed to make inferences or decisions that affect individuals or groups.

We would encourage the Committee to clearly define the report's scope of applicability. Furthermore, we would welcome a more explicit discussion of AI/AS applications and use-cases where manufacturers, users or regulators do not know or cannot explain why a particular decision has been made. This would include a discussion of how such a situation can be resolved in different domains of application.

II.B. Ethics

We would like to echo the submission by Corinne Cath and Jon Crowcroft, which highlight the document's tension between incorporating 'ethics' or 'ethical values' and 'end-user values' and urge the committee to clarify its use.

Furthermore, ethically aligned design should not just address the way in which intelligent systems are built, but also discuss where AI/AS should be employed, and at what level of complexity and autonomy.

II.C. Domain-specificity of AI/AS

The Committee correctly recognises that a "universal set of norms/values that is applicable for all autonomous systems in not realistic" as these values are "largely specific to user communities and tasks."[9] Since the term AI/AS is not clearly defined, this raises the question of whether particular applications and use-cases require more specific guidelines, in particular in the context of domain-specific AI algorithms.

[9] *Ethically Aligned Design. supra* note 4, at p. 24

Take for instance the case of targeted online advertising versus government surveillance versus. Both use Machine Learning algorithms, however in the case of government surveillance, AI algorithms are used to identify suspects and targets, and could potentially inform decisions to use lethal force.[10] Any classification through machine learning is inherently probabilistic,[11] which in turn raises concerns about accuracy and efficacy. An exceptionally low false positive rate is remarkable in business applications, such as targeted advertisement. In the case of government surveillance, however, even an error rate as low as "0.008 percent of the Pakistani population" still corresponds to 15,000 people potentially being misclassified as "terrorists".[12] We urge the Committee to consider use-cases beyond autonomous weapons, where AI is used to make decisions about people that produce significant effects. Particularly as it related to sensitive decisions where bias or false positives can either determine (or significantly influence) life or death, or where outcomes severely impair an individual's fundamental rights, such as the rights to liberty, freedom of movement, privacy, etc.[13]

[10] Consider in this regard NSA's program "SKYNET" which collected in bulk the metadata communication of the entire Pakistani mobile phone network, and then used a machine learning algorithm relied on a machine learning algorithm codenamed "Random Forest" to try and rate "each person's likelihood of being a terrorist". Former director of the NSA and CIA, Michael Hayden, was later quoted as saying: "We kill people based on metadata". For more information *see* David Cole, *We Kill People Based on Metadata*, THE NYR DAILY (10 May 2014), *available at* http://www.nybooks.com/daily/2014/05/10/we-kill-people-based-metadata/; Christian Grothoff & J.M. Porup, *The NSA's SKYNET program may be Killing Thousands of Innocent People*, ARS TECHNICA UK (16 February 2016), *available at* https://arstechnica.co.uk/security/2016/02/the-nsas-skynet-program-may-be-killing-thousands-of-innocent-people/.

[11] *See* Jenna Burrell, *How the Machine 'thinks': Understanding Opacity in Machine Learning Algorithms*, 3(1) BIG DATA & SOCIETY 1 (2016).

[12] Christian Grothoff & J.M. Porup, *The NSA SKYNET Program may be Killing Thousands of Innocenet People*, ArsTechnica (16 February 2016), *available at* https://arstechnica.co.uk/security/2016/02/the-nsas-skynet-program-may-be-killing-thousands-of-innocent-people/.

[13] Note in this regard that even a Pentagon Research Chief acknowledged that Artificial Intelligence is "fundamentally limited". She specifically recognized that "the problem is that when they're wrong, they are wrong in ways that no human would ever be wrong... I think this is a critically important caution about where and how we should use this generation of artificial intelligence" (*see* Mark Pomerleau, *Pentagon Research Chief: AI is Powerful but has Critical Limitations*, Defense Systems (4 May 2016), *available at* https://defensesystems.com/articles/2016/05/04/darpa-chief-limits-of-artificial-intelligence.aspx). The Article also notes that the NSA has been one of the agencies pushing for "more use of automation and intelligent systems. Special Assistant to the Director of the NSA's Cyber Task Force, Philip Quade, is quotes as saying "we have organizations and machines that are capable of sharing information automatically, but... we need more machines to be able to automatically ingest it and act on it".

II.D. Privacy beyond data protection

The recommendations by the Committee addresses the right to privacy in chapter 5, on Personal Data and Individual Access Control. Central to this chapter is the definition of Personally Identifiable Information (PII). By organising the report's main chapter on privacy entirely around the concept of PII the report may inadvertently suggest that privacy harms can only occur if PII is involved. We would like to draw the Committee's attention to the growing role of AI algorithms in practices like profiling, where potentially sensitive information can be predicted or inferred from non-sensitive data.[14] Similarly, uses of AI in face recognition software has the potential to undermine anonymity in public space. While scenarios are about individuals, the data used or generated does not always fall within the definition of PII.

III - Remarks on Report Sections

P. 15 The committee mentions that it is developing principles for *all types* of AI/AS – mentioning this includes both robots and software AI. However, this still leaves unclear what exactly the committee holds AI/AS to be and whether principles and guidelines can be applied universally.

III.A. Remarks on Section: General Principles

P. 16 The list of treaties encompassed does not cover the full corpus of international human rights law nor international humanitarian law. We would expect to find references to additional key treaties, which should be taken into consideration when reviewing AI/AS policies.

These include such treaties as the International Covenant on Economic Social and Cultural Rights (ICESCR), the Convention Against Torture and Other Forms of Cruel, Inhuman and Degrading Treatment or Punishment (CAT), the Convention on the Elimination of Racial Discrimination (CERD), as well as the Martens Clause, the Hague Regulations of 1899, the Hague Regulations of 1907, and the Additional Protocols to the Geneva Conventions.

[14] Information Commissioner's Office. "Big data, artificial intelligence, machine learning and data protection." (2017), *available at* https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf

Principle 1 - Human Benefit

P. 16 The first principle only establishes a general obligation “to respect” human rights in the design and operation of AI/AS, and calls to establish undefined “governance frameworks" for such systems. We find this commitment to be insufficient and overtly generalized and urge the Committee to provide a far more detailed analysis of the way by which the human rights framework can offer more concrete guidance with regards to limitations on the development and use of AI/AS. In this regard we suggest the Committee considers the “Necessary and Proportionate Principles”, a set of international principles on the application of human rights to communications surveillance, launched at the UN Human Rights Council in Geneva in September 2013.[15] While the principles are tailored specifically for Governmental agencies engaging in communications surveillance, they can easily be applied, *mutatis mutandis*, to other forms of automated data gathering and data exploitation by AI/AS, even by private corporate entities. Consider for example companies such as Google, who seek self-learning AI machines to “psychologically profile and predict the behaviour of human consumers so that high-value ads can be delivered to them across Google’s search engine and content network”.[16]

According to various reports intelligence agencies have heavily invested in developed deep learning, quantum computers, and AI technologies.[17] Consider the following scenario. An intelligence agency in country X launches a covert program whereby self-learning AI machines are running on neural networks of quantum computers to swoop in the telephone, internet, and location records of whole populations. The machines than automatically break encryption, run queries on the data using a list of “selectors” (which itself is being routinely updated by the machines on the basis of algorithmic thinking), analyse the information, and make determinations based on the data. Such machines could then decide, independently

[15] For further reading see Necessary and Proportionate: International Principles on the Application of Human Rights to Communications Surveillance (2014), *available at* https://necessaryandproportionate.org/files/2016/03/04/en_principles_2014.pdf.

[16] Mike Adams, Skynet rising: Google acquires 512-qubit quantum computer; NSA surveillance to be turned over to AI machines, Natural News (20 June 2013), *available at* http://www.naturalnews.com/040859_Skynet_quantum_computing_D-Wave_Systems.html.

[17] *See, e.g.*, Max Smolaks, Snowden Reveals NSA’s Classified Quantum Computing Project, Silicon (3 January 2014), *available at* http://www.silicon.co.uk/workspace/snowden-reveals-nsas-classified-quantum-computing-project-134952; Dana Liebelson, *Why Facebook, Google, and the NSA Want Computers That Learn Like Humans*, Mother Jones (1 October 2014), *available at* http://www.motherjones.com/media/2014/09/deep-learning-artificial-intelligence-facebook-nsa; Christopher Steiner, Edward Snowden may be the Last of the Human Spies, The Guardian (29 June 2013), *available at* https://www.theguardian.com/commentisfree/2013/jun/29/edward-snowden-last-human-spies; See also, *Id.*

of any human, whom should be targeted by additional surveillance measures and what potential measures should be employed against the target (e.g. the placement of the target on a no-fly or economic sanctions lists).

A program, such as the one above described, would need to be assessed against the requirements laid out in the necessary and proportionate principles, and would ultimately never meet its requirements. A surveillance program must be reviewed and supervised at three different stages: before it is launched, while it is being carried out, and after it has been terminated.[18] Such program must be prescribed by primary legislation, that is both accessible and sufficiently clear to be foreseeable (the principle of legality); the program must serve a legal interest that is necessary in a democratic society (the principle of legitimate aim); the program is strictly and demonstrably necessary and adequate to achieve that legal interest (the principles of necessity and adequacy); the severity of the infringement must be reviewed taking into consideration whether less intrusive measures exist to achieve the aim (the principle of proportionality); determination regarding both the launch and later reliance on the program to engage in surveillance must be decided by an impartial, independent, and well-resourced judicial body (the principle of competent judicial authority); minimization procedures and procedural safeguards are put in place to prevent abuse at the collection, use, and sharing stages (the principles of due process, safeguards for international cooperation, and safeguards against illegitimate access); targeted individuals are notified, without delay, once it is deemed that such disclosure will not jeopardize the operation. Access to remedy should be provided to those who had the rights abused by the program (the principles of notification and access to remedy); the use and scope of the program should be made known to the public (the principle of transparency); there are independent and effective oversight mechanisms to ensure transparency and accountability (the principle of public oversight); finally, the program should not be used to limit the security of devices and networks, compel companies to assist in building “back-doors”, and reduce online anonymity (the principle of integrity of communications and systems). It is the responsibility of a country seeking to introduce advanced AI/AS into their surveillance operations to show if and how the usage of these tools would be in compliance with the above requirements.

[18] Roman Zakharov v. Russia, App. No. 47143/06, European Court of Human Rights, Judgment, para. 233 (4 December 2015) (“review and supervision of secret surveillance measures may come into play at three stages: when the surveillance is first ordered, while it is being carried out, or after it has been terminated.”).

International human rights courts and experts have significantly developed the understanding of privacy protections since the adoption of the International Covenant on Civil and Political Rights (ICCPR) in 1966, and the Human Rights Committee's adoption of General Comment No. 16, on the Right to Privacy, in 1988.[19] Amongst these accomplishments one can note a significant body of work on human rights and surveillance practices produced since 2009 by the U.N. High Commissioner for Human Rights and the U.N. Special Rapporteurs on Freedom of Expressions and Counter-Terrorism.[20] The repeated adoption by consensus of both U.N. General Assembly Resolutions, and U.N. Human Rights Council Resolutions on the right to privacy in the digital age, also marks a significant step forward.[21] The 2015 creation of a U.N. Special Rapporteur on the Right to Privacy is in itself a reaffirmation of the international privacy agenda, and his reports to the Council, further reaffirm his role as an international intelligence watchdog.[22]

The U.N. Human Rights Committee has begun to routinely address surveillance legislations and practices in its Concluding Observations to States beginning in 2014.[23] At the regional level the European Court of Human Rights, the Court of

[19] U.N. Human Rights, General Comment No. 16: Article 17 (Right to Privacy), U.N. Doc. HRI/GEN/1/Rev.1 at 21 (8 April 1988).

[20] *See e.g.* Report of the Special Rapporteur on the Promotion and Protection of Human Rights and Fundamental Freedoms while Countering Terrorism, U.N. Doc. A/HRC/13/37 (28 December 2009); Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, U.N. Doc. A/HRC/17/27 (16 May 2011); Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, U.N. Doc. A/HRC/20/17 (4 June 2012); Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, U.N. Doc. A/HRC/23/40 (17 April 2013); Report of the Office of the United Nations High Commissioner for Human Rights, The Right to Privacy in the Digital Age, U.N. Doc. A/HRC/27/37 (30 June 2014); Report of the Special Rapporteur on the Promotion and Protection of Human Rights and Fundamental Freedoms While Countering Terrorism, U.N. Doc. A/69/397 (23 September 2014); Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, U.N. Doc. A/HRC/29/32 (22 May 2015); Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, U.N. Doc. A/HRC/32/38 (11 May 2016); Report of the Special Rapporteur on the Promotion and Protection of Human Rights and Fundamental Freedoms While Countering Terrorism, U.N. Doc. A/HRC/34/61 (21 February 2017).

[21] U.N. General Assembly Resolution on the Right to Privacy in the Digital Age, U.N. Doc. A/RES/69/166 (18 December 2014); U.N. Human Rights Council Resolution on the Right to Privacy in the Digital Age, U.N. Doc. A/HRC/28/L.27 (24 March 2015); U.N. General Assembly Resolution on the Right to Privacy in the Digital Age, U.N. Doc. A/RES/71/199 (19 December 2016); U.N. Human Rights Council Resolution on the Right to Privacy in the Digital Age, U.N. Doc. A/HRC/34/L.7/Rev.1. *See also* U.N. Human Rights Council Resolution on the Safety of Journalists, U.N. Doc. A/HRC/33/L.6 (26 September 2016).

[22] *See e.g.*, Report of the Special Rapporteur on the Right to Privacy, U.N. Doc. A/HRC/31/64 (8 March 2016); Report of the Special Rapporteur on the Right to Privacy, U.N. Doc. A/71368 (30 August 2016).

[23] *See e.g.*, Concluding Observations of the Fourth Periodic Report of the United States of America, Human Rights Committee, U.N. Doc. CCPR/C/USA/CO/4 (23 April 2014); Concluding Observations on the Initial Periodic Report of Malawi, Human Rights Committee, U.N. Doc. CCPR/C/MWI/CO/1/Add.1 (19 August 2014); Concluding Observations on the Fifth Periodic Report of Sri Lanka, Human Rights Committee, U.N. Doc. CCPR/C/LKA/CO/5 (21 November 2014); Concluding Observations on the Seventh Periodic Report of the Russian Federation, Human Rights Committee, U.N. Doc. CCPR/C/RUS/CO/7 (28 April 2015); Concluding Observations on the Sixth Periodic Report of Canada, Human Rights Committee, U.N. Doc CCPR/C/CAN/CO/6 (13 August 2015); Concluding Observations on the Third Periodic Report of the Former Yugoslav Republic of Macedonia, Human Rights Committee, U.N. Doc. CCPR/C/MKD/CO/3 (17 August 2015); Concluding observations on the fifth periodic report of France, Human Rights

Justice of the European Union, and the Inter-American Commission and Court on Human Rights have developed considerable and authoritative jurisprudence on surveillance and privacy.[24]

The Report by the Committee fails to reflect on these developments both in Principle 1 and later in the later Section 8 titled "Law". We urge the Committee to include the core fundamental principles, enshrined in international human rights law, in their Report and apply them to AI/AS.

Committee, U.N. Doc. CCPR/C/FRA/CO/5 (17 August 2015); Concluding Observations on the Seventh Periodic Report of the United Kingdom of Great Britain and Northern Ireland, Human Rights Committee, U.N. Doc. CCPR/C/GBR/CO/7 (17 August 2015); Concluding Observations on the Fourth Periodic Report of the Republic of Korea, Human Rights Committee, U.N. Doc. CCPR/C/KOR/CO/4 (3 December 2015); Concluding Observations on the Second Periodic Report of Namibia, Human Rights Committee, U.N. Doc. CCPR/C/NAM/CO/2 (22 April 2016); Concluding Observations on the Initial Report of South Africa, Human Rights Committee, U.N. Doc. CCPR/C/ZAF/CO/1 (27 April 2016); Concluding Observations on the Seventh Periodic Report of Sweden, Human Rights Committee, U.N. Doc. CCPR/C/SWE/CO/7 (28 April 2016); Concluding Observations on the Sixth Periodic Report of New Zealand, Human Rights Committee, U.N. Doc. CCPR/C/NZL/CO/6 (28 April 2016); Concluding Observations on the Fourth Periodic Report of Rwanda, Human Rights Committee, U.N. Doc. CCPR/C/RWA/CO/4 (2 May 2016); Concluding Observations on the Sixth Periodic Report of Denmark, Human Rights Committee, U.N. Doc. CCPR/C/DNK/CO/6 (15 August 2016); Concluding Observations on the Seventh Periodic Report of Colombia, Human Rights Committee, U.N. Doc. CCPR/AZE/CO/4 (4 November 2016); Concluding Observations on the Sixth Periodic Report of Morocco, Human Rights Committee, U.N. Doc. CCPR/C/MAR/CO/6 (4 November 2016); Concluding Observations on the Seventh Periodic Report of Poland, Human Rights Committee, U.N. Doc. CCPR/C/POL/CO/7 (4 November 2016); Concluding Observations on the Sixth Periodic Report of Italy, Human Rights Committee, U.N. Doc. CCPR/C/ITA/CO/6 (28 March 20017); Concluding Observations on the Second Periodic Report of Turkmenistan, Human Rights Committee, U.N. Doc. CCPR/C/TKM/CO/2 (28 March 2017).

[24] *See e.g.*, Klass and Others v. Germany, App. No. 5029/71, European Court of Human Rights, Judgment (6 September 1978); Kopp v. Switzerland, App. No. 23224/94, European Court of Human Rights, Judgment (25 March 1998); Weber and Saravia v. Germany, App. No. 54934/00, European Court of Human Rights, Decision on Admissibility (29 June 2006); Liberty and Others v. The United Kingdom, App. No. 58243/00, European Court of Human Rights, Judgment (1 July 2008); Kennedy v. The United Kingdom, App. No. 26839/05, European Court of Human Rights, Judgment (18 May 2010); Uzun v. Germany, App. No. 35623/05, European Court of Human Rights, Judgment (2 September 2010); Szabó and Vissy v. Hungary, App. No. 37138/14, European Court of Human Rights, Judgment (12 January 2016); Digital Rights Ireland Ltd v. Minister of Communications, Marine and Natural Resources et al. (C-293/12); Kärntner Landesregierung and others (C-594/12), Joined Cases, Court of Justice of the European Union, Grand Chamber, Judgment (8 April 2014); Maximillian Schrems v. Data Protection Commissioner, Case C-362/14, Court of Justice of the European Union, Grand Chamber, Judgment (6 October 2015); Patrick Breyer v. Bundesrepublik Deutschland, Case C-582/14, Court of Justice of the European Union, Second Chamber, Judgment (19 October 2016); Tele2 Sverige AB v. Post- Och telestyrelsen (C-203/15); Secretary of State for the Home Department v. Tom Watson et. al. (C-698/16), Joined Cases, Court of Justice of the European Union, Grand Chamber, Judgment (21 December 2016); Garcia v. Peru, Inter-American Court of Human Rights, Case 11.006, Report No. 1/95, OEA/Ser.L/V/II.88 (17 February 1995); Tristán Donoso v. Panamá, Inter-American Court of Human Rights, Judgment (on Preliminary Objection, Merits, Reparations, and Costs), Series C No. 193 (27 January 2009); Escher et al. v. Brazil, Inter-American Court of Human Rights, Judgment (on Preliminary Objection, Merits, Reparations, and Costs), Series C No. 200 (6 July 2009); Ms. X and Y v. Argentina, Inter-American Commission on Human Rights, Case 10.506, Report No. 38/96 (15 October 1996).

Principle 2 - Responsibility

P. 18 While principle 2 rightly stresses the need to assure that AI/AS are accountable, and that culpability and liability are legislated, we do not find that the proposed recommendations sufficiently address how manufacturers/designers/users of AI/AS can explain "why a system behaves in certain ways". Any discussion of responsibility should address whether AI/AS can be responsible at all, if its manufacturers or external parties cannot sufficiently explain the system's behaviour.

Furthermore, we find that AI/AS raises more fundamental issues than merely "confusion or fear within the general public", one of the most important of which is the thorny question of how can we guarantee AI/AS' compliance with existing laws. Consider for example non-discrimination and privacy legislation, in the wake of learning systems that can result in unintentional discrimination,[25] or an AI that aggregates disproportionate amounts of data.[26]

Autonomous systems are equipped with sensors that collect data about the external world, including human behaviour. It is in this context that we welcome the recommendation put forward by the Committee to create systems for registration of producers/users of autonomous systems, in particular the inclusion of "sensors/real world data sources". We urge the Committee to stress further the importance of transparency about such data collection and the privacy invasions in this regard.

Significant literature has been produced with regards to the difficulties in determining liability facing the regulation of AI/AS (predominately the inadequacy of existing legal structures under contracts, criminal, and torts law).[27] These are mostly tied to the lack of causational links due to the unpredictability of AI/AS and to its self-updating code. We encourage the Committee to introduce more analysis of these issues in the Report.

[25] *See e.g.* Rosenblat, Alex, and Tamara Kneese. "Networked Employment Discrimination." (2014).
[26] See for instance Kashmir Hill, *This sex toy tells the manufacturer every time you use it*, Fusion, (8 September 2016), *available at* http://fusion.kinja.com/this-sex-toy-tells-the-manufacturer-every-time-you-use-1793861000.
[27] See, *e.g.* John Buyers, Liability Issues in Autonomous and Semi-Autonomous Systems, Osborne Clarke (2015), *available at* http://www.osborneclarke.com/media/filer_public/c9/73/c973bc5c-cef0-4e45-8554-f6f90f396256/itech_law.pdf; NEHAL BHUTA ET. AL., AUTONOMOUS WEAPON SYSTEMS: LAW, ETHICS, POLICY (2016); Jens David Ohlin, Machine Liability & The Combatant's Stance, *available at* https://www.law.upenn.edu/live/files/3916-ohlin-jens-machine-liability-and-the-combatants; Rebecca Crootof, *War Torts: Accountability for Autonomous Systems*, PENN. L. REV. (2015).

Principle 3 - Transparency

P. 19 We appreciate the working definition of transparency as “the ability to discover how and why the system made a particular decision [… or] acted the way it did.”[28] Secrecy and technical opacity often inhibits the ability of legislatures, judicial bodies, and oversight mechanisms to scrutinize these systems. “Open debate and security is essential to understanding the advantages and limitations” of AI/AS, “so that the public may develop an understanding of the necessity and lawfulness” of these tools.[29]

We would draft the Committee’s attention to the work of Jenna Burrell[30], who distinguishes between three forms of opacity: (1) opacity as intentional corporate or state secrecy (2) opacity as technical illiteracy, and (3) an opacity that arises from the characteristics of machine learning algorithms and the scale required to apply them usefully. The Committee’s recommendation lacks clarity and detail as to how these different sources of opacity can be mitigated in the vast variety of use cases that are mentioned throughout the Report.

In this regard it is also crucial to define what kind of transparency different stakeholders require. We find that individuals should be provided with sufficient information to enable them to fully comprehend the scope, nature, and application of AI/AS, in particular with regards to what kinds of data these systems generate, collect, process and share. In the case where AI algorithms are used to generate knowledge or make decisions about individuals, users of AI/AS, as well as regulators, do not just need to “determine and allocate responsibility when something goes wrong”, but should be able to determine how a decision has been made, and whether the regular use of these systems violates existing laws, in particular with regards to discrimination, privacy and data protection.

[28] *Ethically Aligned Design. supra* note 4, at p.19
[29] Report of the Special Rapporteur on the Promotion and Protection of Human Rights and Fundemental Freedoms while Countering Terrorism, U.N. Doc. A/HRC/13/37, paras. 54-56 (28 December 2009). While the Report focuses on surveillance techniques specifically, they are applicable to the AI/AS transparency debate more broadly.
[30] Jenna Burrell, *supra note* 7.

Governments and Corporations should publish, at a very minimum, aggregate information of the kind of systems being developed and deployed.[31] Finally, the Committee does not address the importance of whistle-blowing (and whistle-blower protections) in this sphere.

Principle 4 – Education and Awareness

P. 21 Public awareness should not just focus on "unscrupulous manufactures", but the public and users should also be educated about the way in which applications of AI and AS can affect fundamental rights, such as the right to privacy, as well as the mechanisms available for redress.

III.B. Remarks on Section: Embedding Values Into Autonomous Intelligent Systems

P. 24, see also p. 39 The Report takes the approach that privacy is a "culturally distinctive" concept. The report even goes as far as to suggest that different cultures might not consider privacy an issue at all, and engineers should take this fact into account in their designs. The Report further hints to the possibility that the right to privacy is a "western influenced ethical foundation". Privacy International completely rejects and opposes this position. The position not only has no merit but it threads a dangerous line, justifying privacy abuses by those Countries who will wish to argue that it is not an intrinsic and universal right.

Over 130 countries, in every region of the world, have constitutional statements regarding the protection of privacy.[32] Over 100 countries now have some form of privacy and data protection law.[33] The right to privacy is also articulated in all of the major international and regional human rights instruments, including Article 12 of the Universal Declaration of Human Rights and Article 17 of the International Covenant of Civil and Political Rights, as well as regional treaties covering every continent.[34]

[31] *Cf.* Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, U.N. Doc. A/HRC/23/40, paras. 91-92 (17 April 2013).
[32] https://www.privacyinternational.org/node/54
[33] *ibid*.
[34] Other treaties include the U.N. Convention on the Rights of the Child (Article 16); the U.N. Convention on Migrant Workers (Article 14); the European Convection on Human Rights (Article 8); the American Convention on Human Rights (Article 11); the American Declaration of the Rights and Duties of Man (Article 5); the Arab Charter on Human Rights (Articles 16, and 21); the ASEAN Human Rights Declaration (Article 21); the African Charter on the Rights and Welfare of the Child (Article 10); the African Union Principles on Freedom of Expression (Article 4).

In any event, and this is of particular importance for those Countries who have not adopted, some or most of the corpus of treaties of international human rights law, the right to privacy is part and parcel of customary international law. As noted by Rengel:

> "Given the extensive amount of recognition in international instruments of the right to privacy, the prominent place that the topic of privacy continues to occupy in wrings and commentary, and the treatment as binding norm that the right to privacy has received in both national and international legal systems, it can be concluded that there is a general fundamental right to privacy under customary international law. Although the need for protection of the right to privacy continues to expand, it appears that in certain contexts there is widespread recognition that the right to privacy protects individuals from the actions of the state and third parties infringing on that right."[35]

For these reasons, we urge the Committee to remove from the Report any statements that might allude to qualifications on the right to privacy or signalling of its insignificance.

III.C Remarks on Section: Personal Data and Individual Access Control

Data asymmetry vs. data exploitation

P. 56 The report identifies data asymmetry to be at the key ethical dilemma regarding personal information: AI/AS has "widespread access to our data", yet "we remain isolated from gains we could obtain from the insights derived from our lives". Data asymmetry, thus defined, rests on the assumption that inequality is primarily about unequal access to gains and assets.

[35] ALEXANDER RENGEL, PRIVACY IN THE 21ST CENTURY 108 (2013).

We would like to add, that users are also commonly faced with an informational asymmetry as to what kinds of data and how much data their devices, networks and platforms generate, collect, process or share in the first place.[36] This excessive nature of data processing frequently occurs without the explicit, informed consent or knowledge of the user. As we bring ever more connected devices into their homes, workplaces, public space and onto our bodies, we find that such data exploitation is a pressing concern.

Data minimisation begins with generation

P. 64; p. 56 While we agree with the argument that "new parameters must also be created regarding what information is gathered about individuals at the point of data collection", we would like to suggest that the problem originates earlier, at the point of data generation. Data exploitation begins with the excessive generation of data. Data generation occurs when a sensor turns information from the physical world into a signal and as such, generation is the precondition of data collection. A good illustration of this are microphones that are embedded in objects such as cars, under the guise of offering hands-free convenience. If the microphone is able to respond to our voice, does that mean is listening at all times? What inferences can be drawn from these data? And can we be sure that these data are not being shared or hacked? Many consumers are unaware about the fact that any microphone is able to constantly generate a signal, let alone whether this is being collected or even analysed or shared.

Definition of Personally identifiable information and personal data (Section 1 – Personal Data Definitions)

P. 60 While PII is central to data protection and informational privacy regulation around the world, the concept itself is not uniformly defined. The report rightly highlights the fact that different laws and regulation around the globe define PII differently, yet fails to explain why PII is so contentious. In many circumstances non-PII can be linked to individuals, and de-identified

[36] See Paul, Updated: Green Light or No, Nest Cam Never Stops Running, The Security Ledger (24 November 2015), *available at* https://securityledger.com/2015/11/green-light-or-no-nest-cam-never-stops-watching/ or Gibbs, S. Samsung's voice-recording smart TVs breach privacy law, campaigners claim (27 February 2015) https://www.theguardian.com/technology/2015/feb/27/samsung-voice-recording-smart-tv-breach-privacy-law-campaigners-claim.

data can be re-identified. Data that is initially PII can become PII in a different context or in different points in time, where AI/AS – in particular machine learning – increases the scope of non-PII data that can become PII. As a result, "whether information is identifiable to a person will depend upon context and cannot be determined a priori". [37]

In this light, it is important to highlight how the most recent regional data protection standards, the GDPR, will expand the definition of personal (from the Data Protection Directive 95/46/EC.) According to Article 4(1) of the GDPR, personal data means "any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person". The report merely mentions the GDPR in the context of new definitions for genetic and biometric data, which are now also clearly treated as sensitive personal data.

Owing to the fact that PII is becoming an increasingly fluid concept, this raises the important question whether data protection should be limited to situations that involve processing of PII[38]. It is important to highlight that AI/AS is at the heart of novel privacy challenges that cannot be reduced to PII. A good example is the use of sensors in smart cities, or emotional detection technology is public space[39].

[37] See Schwartz, P. M., & Solove, D. J. (2011). The PII problem: Privacy and a new concept of personally identifiable information. *NYUL rev.*, *86*, 1814.

[38] see Paul Ohm, Broken Promises of Privacy, 57 UCLA L. REV. 1701 (2010)

[39] see Andrew McStay, Now Advertising Billboards Can Read Your Emotions … And That's Just The Start, IFL Science (August 4, 2015), *available at* http://theconversation.com/now-advertising-billboards-can-read-your-emotions-and-thats-just-the-start-45519

Personal data as an asset

P. 60 The report defines personal data as "the sovereign asset of the individual" (p. 60). We believe that asset should not mean here as personal property[40], but rather as enable individuals meaningful control over their data. Data protection is seen as a way to protect fundamental rights, including, but not limited to the right to privacy. While certain human rights, such as the right to privacy, may be restricted for legitimate aims and may need to be balanced with other human rights, the very concept of fundamental rights is incompatible with the idea that these rights can be traded, sold, or purchased.

How AI/AS generates data

P. 61 This section is narrowly focussed on "sites or social networks" whereas different sections of this report refer to autonomous cars, and the use of AI algorithms to generate knowledge or make decisions about individuals or groups, an application of which could be the use of facial recognition in public space. We would urge the Committee to address the increasing prevalence of sensors and actuators in public space.

Individual choice, consent

P. 64, p. 57 The report recommends that individuals should "access, manage, and control how their data is shared" and be able to "choose how or whether to share their data with other individuals, businesses, or for the common good as they choose". This stands in contrast to the recommendation on consent, which the committee recommends to be both "conditional and dynamic". It's important that the recommendations reflect the highest standards established within data protection, whereby consent has to be freely given, as well as a specific, informed and unambiguous indication of an individual's wishes, by which she signifies through a clear affirmative action expresses an agreement to the processing of data relating to her. Consent, by its very nature requires awareness. Harm that may be caused without awareness requires safeguards. Under personal data protection laws (such as the EU General Data Protection Regulation) individuals have the right to erase, restrict and object to processing.

[40] See Schwartz, P. M. (2004). Property, privacy, and personal data. *Harvard Law Review*, 2056-2128.

III - Conclusion

In light of the above, Privacy International would like to reiterate the following four considerations:

- We would encourage the Committee to clearly define the report's scope of applicability. Where principles and recommendations are intended to only apply to particular domains or use-cases, this should be indicated.
- The privacy implications of AI/AS are not limited to Personally Identifiable Information (PII).
- Principles and recommendations should not offer weaker protection to individuals than the General Data Protection Regulation.
- The Committee should address challenges and possible solutions to the opacity of some AI/AS, in particular applications and use-cases where manufacturers, users or regulators do not know or cannot explain why a particular decision has been made. A minimum degree of transparency is the precondition for ethically aligned design.

We would be happy to engage with the IEEE formally and informally of these topics in the future.

Renato Opice Blum
www.opiceblum.com.br
Coordenador do Curso de Direito Digital do INSPER

It is a great pleasure and an honor for me to be able to collaborate with this important document.

- On page 89, I recommend adding "4. privacy and security" as an area of concerns, regarding Law & AI/AS theme. To be effective, AI and AS technologies depends on a massive data treatment, which may include personal and sensitive data. We also must observe that article 22 of the European Regulation (GDPR) sets that data subjects have a right not to be subject to a decision based solely on automated processing, which may represent an important challenge to AI designers and users. In this sense, privacy and security may represent an important challenge on the Law studies, considering regulation shall not represent a technologic development impeditive, but, on the other hand, privacy and security shall not be overlooked on this debate.

- On page 94, I recommend adding the following statement to the "Background" topic: "In addition, AI may address concerns regarding sensitive data treatment, as many of the AI systems are focused on improving health diagnosis and medical treatment.

- On page 94, I recommend adding the following subtopics on "Candidate Recommendation":

  > "3. Companies that use and manufacture AI should be required to provide clear information regarding personal data treatment procedures, warning data subjects and contractors about the risks that may be involved on the usage of this technology";

> “4. Designers should consider adopting the concept of privacy by design on the development of AI technologies, which means that privacy and security should be considered as a main aim on all steps of the system development”;
>
> “5. Data subjects may have the opportunity to appeal of an automated decision made by an AI system which is undertaken based on personal data treatment”.

- On page 94, I recommend adding the following references to the “Further Resources”: “Lodder, Arno R. and Wisman, Tijmen, Artificial Intelligence Techniques and the Smart Grid: Towards Smart Meter Convenience While Maintaining Privacy (January 13, 2016). Journal of Internet Law (Dec. 2015), Vol. 19(6), p. 20-27. Available at SSRN: https://ssrn.com/abstract=2714840” and “Kamarinou, Dimitra and Millard, Christopher and Singh, Jatinder, Machine Learning with Personal Data (November 7, 2016). Queen Mary School of Law Legal Studies Research Paper No. 247/2016. Available at SSRN: https://ssrn.com/abstract=2865811”.

Best regards,

Renato Opice Blum
www.opiceblum.com.br
Coordenador do Curso de Direito Digital do INSPER

Comments on Ethically Aligned Design, Version 1 submitted by Joachim Iden, TUV Rheinland Japan, email: joachim.iden@tuv.com

1. Regarding Section 1/ General Principles: suggestion to add the following principle

Principle: Awareness of Residual Risks and Preparedness for Failure

Even with safety-oriented design and multi-level countermeasures implemented, technology can fail and experience has shown that it does fail, even catastrophically so. Important is therefore to sufficiently understand

- the worst possible impact of a failure
- the probability of such an event
- the societal acceptability of the corresponding risk

Issues:

- how accurately can failure scenarios be determined?
- can probabilities actually be calculated?
- how can consensus be reached about acceptable residual risks when novel types of hazards are involved?

Equally important is to understand what responses will be required in case of a failure and the involved resources and their economic corresponding costs. These costs will not only be incurred in case of an actual failure, but also accrued by providing the capability to deal with such an event.

Society must decide whether it is willing to carry these expenses even in the light that certain stakeholders may be tempted to put the emphasis on the presumably extremely low probability of a failure event.

2. Regarding Section 1, page 19, Principle 3 – Transparency, suggestion to include the following considerations

Regarding Transparency

I. Transparency regarding specification, design and development should involve a clear reference to intellectual honesty when it comes to admitting that there may be unintended effects and specifically also when there are doubts and uncertainties regarding

- the assessment of the worst case scenario
- determination of associated probabilities of failure scenarios

Refraining from deploying a technology due to doubts and uncertainties must always remain a viable option. In order to maintain this option, education and training must emphasize self-critical thinking and methodologies must be developed and improved that foster the critical analysis of technological systems from diverse perspectives.

II. Discussing transparency with respect to the operation of a system must also involve a fundamental consideration of factors that may limit this kind of transparency in principle. Relevant questions may include e.g.

- does the requirement for transparency preclude the use of specific technologies ?
- are design approaches available ensuring that users can reliably infer imminent actions of the system from the observable interactions with it ?
- how to define and measure degrees of transparency ?

3) Regarding the Executive Summary, page 6 and the document in general

Comment: one stated issue on page 6 is “achieving a correct level of trust between humans and AIS”. The notion of trust is not defined in the document as a whole, nor what may be a correct level of trust and how it can be determined. If trust is considered a relevant notion, it must be formally defined and distinguished from seemingly similar notions like trustworthiness. A formal theoretical model will also be needed to discuss what may be meant by the expression “correct level of trust”.

An overview of approaches to the formalization of the notion of trust can be found in

M. Lahijanian, M. Kwiatkowska “Social Trust: A Major Challenge for the Future of Autonomous Systems”, The 2016 AAAI Fall Symposium Series: Cross-Disciplinary Challenges for Autonomous Systems, Technical report FS-16-03

Ilse Verdiesen MSc.
Officer in the Royal Netherlands Army
Master student TUDelft (graduation project on the ethics of Autonomous Weapons)

Dear members of the IEEE Global Initiative,

I have reviewed section 6 as it is closely related to my graduation project on the ethics of Autonomous Weapons. Many relevant and critical issues are raised in this section and I would like to offer the following remarks as feedback:

1. Introduction (p. 68): In paragraph three it is stated that: '*we would like to ensure that stakeholders are working with comprehensive shared definitions of concepts…*'. However, key concepts such as *Autonomous Weapons Systems* (AWS) and *Meaningful Human Control* are not defined in the introduction. Although I realize there is still no consensus on these definitions, I would like to suggest defining these two concepts in the introduction to provide clarity and guidance to the reader.

   A definition for AWS could be: *'A weapon that, without human intervention, selects and engages targets matching certain predefined criteria, following a human decision to deploy the weapon on the understanding that an attack, once launched, cannot be stopped by human intervention.'* (1: 11).

   The concept of Meaningful Human Control is in my opinion too vague and raises questions to me such as: *Is there also Meaningless Human Control? Or Meaningful Machine Control?* Therefore, I propose to use the term Human Oversight instead, meaning that an AWS is used under human supervision. The concept of Human Oversight is also used in the issue regarding inadvertent violation of Human Rights on page 76 and it can also be applied to other fields of AI as well, for example Autonomous Vehicles or AI in the medical domain.

2. The issue regarding codes of conduct (p. 69) is very broadly described and seem to apply to all forms of AI in general and not specifically to AWS. In my opinion all AI technologies should be developed with current legal frameworks and not only in regard of the International Humanitarian Law or International Human Rights Law apply. Therefore, I propose to remove it from the AWS section and place this issue in the section that covers Law (p. 89-94).

3. In the background description of the issue on page 72 it is stated that: '*The lack of a clear owner of a given AWS incentivizes scalable covert of no-attributable uses of force by state and non-state actors. Such dynamics can easily lead to unaccountable violence and societal havoc*.' In my opinion the use of an Autonomous Weapons System is equally attributable to an owner as conventional weapons and I do not see why there is a difference in accountability. An AWS will always be used by an actor based on a decision-making process and this will (in most cases) lead to an identifiable owner that deployed the weapon and that can be held accountable. The autonomous capabilities of the weapon will not change this. The description of this issue can benefit by providing a better motivation and additional literature as support. It either needs to be better substantiated or I propose to remove it as the ownership issue of an AWS is no different than that of conventional weapons.

4. The issue that '*By default, the type of automation in AWS encourage rapid escalation of conflicts*' (p. 77) is speculative in my opinion and references that substantiate this claim are missing. The causal relation between 'interaction of opposing AWS' and 'the increase of escalation' is not clear to me. Humans will decide when and how to deploy AWS and will take the risk of escalation into account. It appears to me that fear of lack of control and unpredictability are the underlying rationale for this issue. The description of this issue could also benefit of a better motivation and additional literature as support. In my view, it needs to be less speculatively described as it is now, or deleted from this document if the claim that 'interaction of opposing AWS will increase escalation' does not hold.

Please let me know if you have any questions or remarks regarding my feedback or if I can contribute more to this topic in the future.

Best regards,

Ilse Verdiesen MSc.

Officer in the Royal Netherlands Army

Master student TUDelft (graduation project on the ethics of Autonomous Weapons)

T: @IlseVerdiesen

Pages referenced: Section 6: Reframing Autonomous Weapons Systems p. 69-79

References

[1] AIV, & CAVV. (2016). *Autonomous weapon systems: the need for meaningful human control*. (No. 97, No. 26). Retrieved from http://aiv-advice.nl/8gr.

Pradyot Sahu; Senior Member, IEEE; Director, 3innovate

Dear Sir/Madam(s)

Here are my comments and suggestions regarding Version 1 of Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems.

A. On Page 16, Principle 1- Human Benefit instead of “Issue: How can we ensure that AI/AS do not infringe human rights? “

   I recommend "Issue: How can we ensure that each and every AI/AS does not infringe human rights?"

B. On Page 18, Principle 2- Responsibility instead of “Issue: How can we assure that AI/AS are accountable?”

   I recommend "Issue: How can we assure that each and every AI/AS made accountable?"

C. On Page 19, Principle 3- Transparency instead of “Issue: How can we ensure that AI/AS are transparent?”

   I recommend "Issue: How can we ensure that each and every AI/AS is transparent?"

D. On Page 21, Principle 4-Education and Awareness instead of “Issue: How can we extend the benefits and minimize the risks of AI/AS technology being misused?”

   I recommend "Issue: How can we extend the benefits and minimize the risks of each and every AI/AS technology being misused?"

E. The suggestions A, B, C, D are to make the principles more forceful and to include each and every AI/AS.

F. Each General Principle (page 15) may incorporate design, implementation and evaluation of each topic as shown below to implement General Principles in a better way. The issues of each topic may be the following: -

HUMAN BENEFIT – DESIGN, IMPLEMENTATION AND EVALUATION

Issue: How can we ensure design of human benefit for each and every case of AI/AS so that AI/AS do not infringe human rights?

Issue: How can we ensure implementation of human benefit for each and every case of AI/AS so that AI/AS do not infringe human rights?

Issue: How can we ensure evaluation of human benefit for each and every case of AI/AS so that AI/AS do not infringe human rights?

RESPONSIBILITY - DESIGN, IMPLEMENTATION AND EVALUATION

Issue: How can we assure design of responsibility for each and every case of AI/AS made accountable?

Issue: How can we assure implementation of responsibility for each and every case of AI/AS made accountable?

Issue: How can we assure evaluation of responsibility for each and every case of AI/AS made accountable?

TRANSPARENCY - DESIGN, IMPLEMENTATION AND EVALUATION

Issue: How can we ensure design of transparency for each and every case of AI/AS?

Issue: How can we ensure implementation of transparency for each and every case of AI/AS?

Issue: How can we ensure evaluation of transparency for each and every case of AI/AS?

EDUCATION AND AWARENESS - DESIGN, IMPLEMENTATION AND EVALUATION

Issue: How can we design the extension of the benefits and minimize the risks for each and every AI/AS technology being misused?

Issue: How can we implement the extension of the benefits and minimize the risks for each and every AI/AS technology being misused?

Issue: How can we evaluate the extension of the benefits and minimize the risks for each and every AI/AS technology being misused?

G. Section 1- Automation and Employment (page 82) may add a new principle or issue.

AI-Assisted New Job Creation Principle

Issue: How can we ensure AI/AS create enough new jobs while eliminating some of it?

Background – There seem to be enough rational and irrational fear in the press and the public that AI/As will take out almost all the jobs and the world will be left almost without any employed human beings. In a fast-changing environment of AI/AS implementations, any job losses due to AI/AS are required to be compensated by AI-assisted new job creation.

Thanks

Pradyot Sahu, Senior Member, IEEE

Director, 3innovate

pradyot.sahu@3innovate.net

pradyot.sahu@gmail.com

Submission by: Christina Demetriades
Deputy General Counsel,
Sales & Delivery,
Accenture

Response to call for comments:

IEEE's Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems *"Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems" of 13 December 2016.*

We welcome the opportunity to comment on the detailed proposals and recommendations made by the IEEE's Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems in its December report.

The overarching aim of this report – to encourage Ethically Aligned Design of Artificial Intelligence and Autonomous Systems – is laudable. As we create more sophisticated and effective Artificial Intelligence and Autonomous Systems that are increasingly pervasive across government and industry, the desire to ensure that those AI and AS should in some way mirror human values and ethics is fundamental. Fundamental because, without alignment of AI and AS to value systems and ethical guardrails, there is a real risk that AI/AS will perpetuate and/or exacerbate ethical challenges arising today (for example bias, breach of privacy etc), leading to a trust deficit which could undermine the adoption of those technologies. In deterring adoption we may be deprived of opportunities to unlock trapped value and solve other societal problems offered through use of the emerging technology.

There is a broad public awareness of the potential social and economic upheaval that the uses of such technology may bring. Driven by concern about this impact and questions about how best to address that challenge, there are various bodies looking at this issue. The IEEE report seeks to address those concerns by suggesting a framework of thinking around the ethical issues arising out of the use of the technology. We are aligned to those general recommendations, and provide our detailed comments below. We see the IEEE paper as a very positive contribution to the debate.

However, in this note we would like to pose some challenges to the approach in the report, to test the resolve of our collective thinking.

The general assumption being made is that if we create ethical frameworks, then we can design AI/AS in such a way that eliminates bias, handles data sensitively and in a way which cedes control of personal data to the individual and is transparent about the decisions and actions taken by the relevant AI/AS.

What if the simple truth is that you cannot trust Artificial Intelligence?

What if it is that the human intervention, in the process of commencing[41] the design, build and run of these systems necessarily means that we will intrinsically always create systems which reflect our own human failings? Our own value systems and ethics have not, to date, managed to entirely eliminate these very same issues without AI/AS, even though it is desirable to do so. Indeed, even as society has developed more and more sophisticated tools over generations, we still have seen a rise in the nature of problems that we have in exactly similar spheres to those that the IEEE report raises.

We do not for a moment suggest that we should not be trying to address the issues raised by the IEEE, or that the detail of this group's recommendations are not worthy aspirations, deserving of detailed consideration and adoption. However, we think these aspirations should be considered in light of challenges both known and unknown.

For example, do we honestly think that humans can eliminate or avoid similar failings in AI/AS merely because we focus more on the possibility of them arising or because we educate the technologists or enterprises as to the potential risks? Pragmatically, do we instead need to acknowledge that the technology will be as flawed as we are as a species?[42]

[41] Here we focus on humans at the beginning of the design process given that the expectation is that AI/AS will, in due course, routinely be capable of re-designing itself and evolution.

[42] http://www.sciencemag.org/news/2017/04/even-artificial-intelligence-can-acquire-biases-against-race-and-gender

If that is the case, then our proposition is that we need to change our own sphere of reference such that we consider that:

- Humans are not reliable ethical beings.
- We need to be practical and realistic in our expectations of AI/AS. There will be bias in those systems, given we cannot hope to entirely avoid it. Rather, we need to acknowledge that it will arise and seek to minimize its impact by creating governance structures that allow sufficient transparency to detect flaws and course correct when discovered.
- We have crossed the Rubicon in terms of personal data. In the digital age, where as individuals we are engaging digitally across all dimensions of our lives, it is inevitable that enterprises, governments and other organizations will collect more information and inferences about us than we knew existed or even appreciated. Informed consent will be difficult to achieve and, given the fact AI/AS will evolve without human intervention, is ephemeral. A desire for "data sovereignty" of the individual is very likely unachievable. If that is right, then what type of control over personal data is appropriate and should be brought to life in the form of privacy regulation?
- We are at a crossroads where the expectation of privacy varies widely by demographic group.[43] Does the current legislative framework need fundamental rethinking in this context?
- Transparency is not the same as honesty or fairness. Being transparent about how one gets to a decision does not automatically lead in the human sphere to more honesty, necessarily, or more fairness.
- Complete transparency may very well be unachievable and undesirable. In social interactions between humans, there are some matters which are left unsaid and, in being unsaid, are helpful to the social contract. In regular human interactions, there are certainly some relationships that would not benefit if we were all completely transparent all the time. Why do we think that all human to AI/AS interactions would benefit from complete transparency?

[43] While adults between the ages of 18-29 share more personal information online, they also are more likely to employ more strategies to be less visible online. http://www.pewresearch.org/fact-tank/2016/09/21/the-state-of-privacy-in-america/

- In the age of Big Data, more data will lead to better inferences. All social interaction, whether human to human or human to robot, involves the making of assumptions and calculations made on data both known and inferred. We need to consider whether as a consequence AI/AS may have a significant advantage to human to human interfaces by relying less on inference and more on data, with many attendant benefits.
- Solid cybersecurity practices are an essential underpinning to ethical design to ensure that consumer trust is established and maintained throughout the entirety of an AI deployment cycle.

In the end, we "overcome" or succeed as a race, despite our many failings, because there is something about our humanity which allows us to moderate or militate against them. This points to a fundamental need to ensure that our systems are human centric, by which we mean there is a human at the center of the processes, design and appeals process which govern them.

I would like to thank the following people for their contributions to our submission - Sean Burke, Jack Calderwood, Rumman Chowdhury, Jennifer Handa, Nijma Khan, Christopher Lynn, Lisa Neuberger-Fernandez, Deborah Santiago, Lina Su, Louise Townsend, and Barbara Wynne, with special thanks to our Chief Technology Officer, Paul Daugherty, our AI Group Chief Executive, Nicola Morini-Bianzino, and our Global Labs lead, Marc Carrel-Billiard. Their contributions reflect our alignment with the IEEE's approach that AI/AS is not merely about technology, but that a successful and ethical AI deployment requires a multi-disciplinary collaboration across the enterprise.

Detailed Comments

Accenture has reviewed the IEEE's consultation document with some interest and include below our comments on the following sections of the IEEE's document, attached as Appendix A to this cover letter:

- Section 2 – Embedding Values into Autonomous Intelligent Systems
- Section 3 – Methodologies to Guide Ethical Research and Design
- Section 5 – Personal Data and Individual Access Control
- Section 7 – Economics/Humanitarian Issues
- Section 8 – Law

I respectfully submit Accenture’s comments to the IEEE paper on behalf of Accenture.

Warm Regards,

Christina Demetriades

Accenture, Deputy General Counsel - Sales and Delivery

Appendix A

Accenture IEEE Response

Structure of our response

For ease of reference we have included the text from the IEEE document in the left column of our table below and marked the relevant page reference. Accenture's contributions, by way of corresponding comments or drafting notes, are set out in each row on the right hand side of the table. We have divided the table into parts addressing the various Sections listed above.

Key to our comments

Our comments on the text of the IEEE document are marked as follows:

- our comments and suggestions are marked with *italics*,
- suggested additions are marked in **bold**, and
- suggested deletions are ~~struck~~.

A. This part A sets out Accenture's comments and suggested revisions to Section 2 "Embedding Values Into Autonomous Intelligent Systems" and Section 3 "Methodologies to Guide Ethical Research and Design".

Original version	Accenture Contributions
Page. 33 The second level of transparency, as stated above, is needed to evaluate a system as a whole by a third party (e.g., regulators, society at large, and post-accident **investigators**).	*AI is defined by its models that learn and self-evolve over time. As a result, one aspect of oversight is a continual and evolving process of evaluation, as appropriate for the level and complexity of the model*. *It is our view that transparency is a very important aim and, in a private sector enterprise setting, must be part of a broader compliance and governance framework if it is to influence outcomes in practice and to reinforce structures of accountability.*

Page. 37 Ethics and ethical reflection need to be a core subject for engineers and technologists beginning at University level and for all advanced degrees.	We suggest the following edit: Ethics and ethical reflection need to be a core subject for engineers and technologists beginning at University level and for all advanced degrees. **Human-centric design (as defined in Section 2) emphasizes the need for ethics and ethical reflection. Key to any course curricula around ethical AI considerations are not just theoretical discussions, but project-based ethical implementations as part of hands-on training.** Comment: *As a matter of best practice, when designing compliance systems for organizations, it is important to see training on key topics such as this as an ongoing educational need and an area where ongoing investment is required.*
Page. 39 A responsible approach to embedded values (both as bias and as value by design) in ICTs, algorithms and autonomous systems will need to differentiate between culturally distinctive values (i.e. how do different cultures view privacy, or do they at all? And how do these differing presumptions of privacy inform engineers and technologists and the technologies designed by them?). Without falling into ethical relativism, it is critical in our international IEEE Global Initiative to avoid only considering western influenced ethical foundations. Other cultural ethical/moral, religious, corporate and political traditions need to be addressed, as they also inform and bias ICTs and autonomous systems.	A comment on "Background": *Ethical discussions, including demographic-based considerations, organically occur in settings with a diverse and equally empowered participatory audience.* *Currently, representation of women and minorities is low in all levels of STEM fields within education and in industry. Similarly, non-Western regions, particularly regions of lower income and economic mobility, are as a consequence sometimes left out of AI solutions and considerations.* *An imperative to include diverse approaches must also create inclusion of these parties and data sets that are reflective of the societies from which AI is deployed.* *Accenture has multiple AI-driven initiatives to reduce barriers to entry for these under-represented groups, outlined in the Section 7 discussion.*

Page. 41 Technology leaders give innovation teams and engineers too little or no direction on what human values should be respected in the design of a system. The increased importance of AI/ AS systems in all aspects of our wired societies further accelerates the needs for value-aware leadership in AI/AS development.	We suggest the following addition to "Background": **Developers and integrators of AI technologies operate in the context of client instructions and client/industry priorities. In order for industry to fully embrace ethical integration into AI solutions, the utility of these goals must be clearly understood.** **Precedence exists for a utility-based justification for ethical and moral imperatives – by way of example, LEED certification motivated corporations to invest in sustainable design, based on cost estimates for long-term savings.**
Page. 43 There is a divergence between the values the technology community sees as its responsibility in regards to AI/AS, and the broader set of social concerns raised by the public, legal, and social science communities. The current makeup of most organizations has clear delineations between engineering, legal, and marketing arenas. Technologists feel responsible for safety issues regarding their work, but often refer larger social issues to other areas of their organization.	A comment: *Responsibility and ownership of outcomes are aligned. The difficulty in creating a sense of 'responsibility' for unethical AI outcomes lies in a mechanism of communicating consequences back to the implementers (whether management or technologists), and including an aspect of responsibility for those consequences to that feedback.* *Other parts of organizations (Legal staff included) will need to shift how they think about their roles, their interactions with other teams, and their responsibilities to account for these shifting needs. And completely agree with our comment back on need to communicate consequences, which is something we routinely struggle with as an organization in many situations.* *Given concerns about liability, litigation, privacy, etc. companies may need to rethink their feedback loops to get this right.*

Page. 45 The algorithms behind intelligent or autonomous systems are not subject to consistent oversight. This lack of transparency causes concern because end users have no context to know how a certain algorithm or system came to its conclusions.	A comment on p. 45 “Background” (that also applies to p. 33 “Background”): *AI is defined by its models that learn over time. As a result, one aspect of oversight is a continual and evolving process of evaluation, as appropriate for the level and complexity of the model.* *Second, there are three moving components to an AI solution: data, algorithms, and people. Each of these three are subject to different types and levels of oversight at different stages.* *Suggest including a discussion or enumeration of what oversight over time for these aspects of an AI solution would entail.*

A. This part B sets out Accenture's comments and suggested revisions to Section 5 "Personal Data and Individual Access Control" & Section 8, "Law", as it regards Personal Data.

Original version	Accenture comments and revisions
Page 56 A key ethical dilemma regarding personal information is data asymmetry. Our personal information fundamentally informs the systems driving modern society but our data is more of an asset to others than it is to us. The artificial intelligence and autonomous systems (AI/AS) driving the algorithmic economy have widespread access to our data, yet we remain isolated from gains we could obtain from the insights derived from our lives.	Comment: *We would recommend adding a short definition (in a footnote, glossary or link) of what is generally meant by "data asymmetry" (e.g. an imbalance of power caused by one party having more control over personal information than the other – or in this case one party being able to derive more benefit than the other).* *Is it possible to give some examples of gains that could be made for individuals?*
Page 56 To address this asymmetry there is a fundamental need for people to define, access, and manage their personal data as curators of their unique identity. New parameters must also be created regarding what information is gathered about individuals at the point of data collection. Future informed consent should be predicated on limited and specific exchange of data versus long-term sacrifice of informational assets.	Suggested edit: To address this asymmetry there is a fundamental need for people to **be informed of the use of their personal data (including contextualizing the use so that people can understand why it matters and what kinds of impacts the use could have) and to** define, access, and manage their personal data as curators of their unique identity. Comment: *It would be helpful to elaborate on what kind of new parameters must be created regarding data collection to protect personal data including specific security technologies.*

Page 56 There are a number of encouraging signs that this model of asymmetry is beginning to shift around the world. For instance, legislation like The General Data Protection Regulation (GDPR) is designed to strengthen citizens' fundamental rights in the digital age and facilitate business simplifying rules for companies by unifying regulation within the EU. Enabling individuals to curate their identity and managing the ethical implications of data use will become a market differentiator for organizations.	Suggested edit: The **EU** General Data Protection Regulation (GDPR) is designed to strengthen citizens' fundamental rights in the digital age and facilitate business ~~simplifying rules for companies~~ by **attempting to unify** regulation within the EU. Comment: *In practice derogations are nevertheless possible in some areas*. *As a consequence it is likely that some level of disparity will remain across EU Member States, notwithstanding the Regulation coming into force.*
Page 56 While some may choose minimum compliance to legislation like the GDPR, forward-thinking organizations will shift their data strategy to enable methods of harnessing customer intention versus only invisibly tracking their attention. We realize the first version of The IEEE Global Initiative's insights reflect largely Western views regarding personal data where prioritizing an individual may seem to overshadow the use of information as a communal resource. This issue is complex, as identity and personal information may pertain to single individuals, groups, or large societal data sets.	A comment in relation to: "While some may choose minimum compliance to legislation like the GDPR, forward-thinking organizations will shift their data strategy to enable methods of harnessing customer intention versus only invisibly tracking their attention." *It may be helpful to include an example of what is meant here – are you suggesting that AI would detect customer intentions?*

Page 58 The following definitions, resources, and candidate recommendations are provided to realign the systematic tracking, distribution, and storing of personal data to overtly include individuals and their predetermined preferences in the process. Issue: How can an individual define and organize his/her personal data in the algorithmic era? Background Personal data needs to embrace an individual's definition and clarification of his/her identity, mirroring unique preferences and values. Candidate Recommendation Where available, individuals should identify trusted identity verification resources to validate, prove, and broadcast their identity.	Suggested edit: Personal data needs to embrace an individual's definition and clarification of his/her identity, mirroring unique preferences and values **whilst nevertheless being accurate and verifiable where necessary.** A comment: *While we agree that individual's own definitions of identity and their preferences and values are important, there should be some balance/alignment with generally accepted values, which also will change over time*. *Separately, as nearly every facet of a person's life is or moves online, we may need to shift how we think about personal data and what constitutes personal data to the point that it is not just about considering how to curate personal data, but also that one needs to curate one's own "data persona." When one considers AI, algorithms and all of the various ways in which data gets collected and can be combined then everything from a person's apple preference to driving habits to health records could fall into the expansive definition. Does part of the dialogue need to be around new terminology and concepts to address that?* Suggested edit: Where available, individuals should identify trusted identity verification resources to validate, prove, and broadcast their identity. **Governments, standards organizations, industries and others should be encouraged to further develop such resources where appropriate.**

Page 58 The following are two examples of identity programs along these lines: • eIDAS Work is underway to explore extending the U.K. Verify Program to commercial applications and not just government. This aligns to the implementation of the eIDAS scheme throughout the European Union, known as Regulation (EU) N°910/2014. Adopted by the co-legislators in July 2014, the eIDAS scheme is a milestone to provide a predictable regulatory environment that enables secure and seamless electronic interactions between businesses, citizens, and public authorities. It ensures that people and businesses can use their own national electronic identification schemes (eIDs) to access public services in other EU countries where eIDs are available. The aim is to create a European internal market for eTS—namely electronic signatures, electronic seals, time stamp, electronic delivery service, and website authentication—by ensuring that they will work across borders and have the same legal status as traditional paper-based processes. With eIDAS, the EU has provided the foundations and a predictable legal framework for people, companies, and public administrations to safely access services and do transactions online and across borders in just "one click." Rolling out eIDAS means higher security and more convenience for any online activity such as submitting tax declarations, enrolling in a foreign university, remotely opening a bank account, setting up a business in another Member State, or authenticating for internet payments.	A comment: *eIDAS: Recommend commenting on or linking to current status of roll out.*

Page 59-60 Issue: What is the definition and scope of personally identifiable information? Background Personally identifiable information (PII) is defined as any data that can be reasonably linked to an individual based on their unique physical, digital, or virtual identity. As further clarification, the EU definition of personal data set forth in the Data Protection Directive 95/46/EC defines personal data as “any information relating to an identified or identifiable natural person.” The Chairwoman of the United States Federal Trade Commission has also suggested that PII should be defined broadly. The new GDPR legislation also provides definitions for genetic and biometric data that will become even more relevant as more devices in the Internet of Things track these unique physical identifiers. Candidate Recommendation	We suggest the following edits: Personally identifiable information (PII) is **generally considered** ~~defined~~ as any data that can be reasonably linked to an individual based on their unique physical, digital, or virtual identity. As further clarification, the EU definition of personal data set forth in the Data Protection Directive 95/46/EC defines personal data as “any information relating to an identified or identifiable natural person.” The Chairwoman of the United States Federal Trade Commission has also suggested that PII should be defined broadly. The new GDPR legislation also provides **a broad** definition~~s~~ **of personal data by reference to identifiers such as name, identification number, location data, online identifier or other factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of a natural person as well as** definitions for genetic and biometric data **and for profiling** that will become even more relevant as more devices in the Internet of Things track these unique ~~physical~~ identifiers.

Page 60 PII should be considered the sovereign asset of the individual to be legally protected and prioritized universally in global, local and digital implementations. In the U.S., for instance, PII protection is often related to the right of the people to be secure in their persons, houses, papers, and effects, pursuant to the fourth amendment to the Constitution (e.g., the Supreme Court's ruling in US v. Jones from 2012, 565 U.S.).lviii In the EU, PII protection is commonly framed in terms of informational self-determination and defense of human dignity. In both cases, (See generally United States v. Jones, 565 U.S. 400 (2012)) the aim should be to tackle key ethical dilemmas of data asymmetry by prioritizing PII protection universally in global, local, and digital implementations.	We suggest the following edits: PII should be considered the sovereign asset of the individual to be legally protected and prioritized universally in global, local and digital implementations **whilst nevertheless recognizing and balancing the wider interests and needs of other stakeholders such as public bodies, business and society.**
Page 61 Issue: What is the definition of control regarding personal data? Background Most individuals believe controlling their personal data only happens on the sites or social networks to which they belong. While taking the time to update your privacy settings on a social network is important, the logic of controlling your personal data is more holistic and universal in nature. Instead of individuals having to conform to hundreds of organization's terms and conditions or policies, in a world where people control their own personal data, those organizations would conform to an individual's predetermined requirements. Candidate Recommendation Personal data should be managed starting from the point of the user versus outside actors having access to data outside of a user's awareness or control.	Comment on p. 61, "Candidate Recommendation" *Perhaps this recommendation should be augmented by a reference to the education and awareness of users and how that could be achieved. On page 57 it states "Provides for future educational programs training all citizens/individuals regarding the management of their personal data and identity" but this theme does not seem to be further explored in Section 5 and arguably many of the candidate recommendations will to some extent be dependent on user take-up which can only be achieved by awareness and understanding.*

Page 63 If you cannot access your personal data, you cannot benefit from its insights. Also, you will not be able to correct erroneous facts to provide the most relevant information regarding your life to the actors you trust. Multipage agreements written to protect organizations must also quickly and genuinely inform users of their choices for trusted consent in the algorithmic era.	Regarding p. 63 "If you cannot access your personal data" a comment: *Perhaps include a reference to the GDPR including the concepts of* *Privacy by Design, privacy by default, accountability and Data Privacy Impact Assessments?* *What is meant by "trusted consent" as opposed to "informed consent"? A definition may be helpful here.*
Page 63 Candidate Recommendation Practical and implementable procedures need to be available in order for designers and developers to use "Privacy-by-Design"/Privacy-by-Default methodologies (referring to the practice or business philosophy of privacy embedded in the development of a service).	A comment regarding "Privacy-by-Design/Privacy-by-Default methodologies (referring to the practice or business philosophy of privacy embedded in the development of a service)": *We recommend a preference for procedures be developed at a global/industry level in combination with organizations.*

Page 63 In order to realize benefits such as decision enablement and personalization for an individual, open standards and interoperability are vital to ensure individuals and society have the freedom to move across ecosystems and are not trapped by walled gardens. In order to safeguard this freedom, for example, Article 20 of the EU regulation on data protection (Right to Data Portability) sets up the right to receive PII that individuals have provided to a data controller, in a structured, commonly used and machine readable format and have the right to transmit those data to other controllers without hindrance from the controller to which the personal data have been provided. lix Paradigms like “differential privacy” may also allow for designers and developers to bake privacy into the design and development of services. lx Differential privacy shifts the focus from “your data” to finding general usage patterns across larger data sets. Differential privacy is not about anonymization of data, as that can be easily de-anonymized through intelligent cross-referencing. Instead differential privacy uses hashing, sub-sampling, and noise injection techniques to obfuscate personal information about individuals. However, while differential privacy may provide a methodology for better usage of private or public data, it should be implemented in complement to tools and methodologies empowering individuals to manage and control their data. As a tool for any organization regarding these issues, a good starting point is to apply the who, what, why, and when test to the collection and storage of personal information: 1. Who requires access and for what duration—is it a person, system, regulatory body, legal requirement “or” input to an algorithm?	A comment regarding: “In order to realize benefits such as decision enablement and personalization for an individual, open standards and interoperability are vital to ensure individuals and society have the freedom to move across ecosystems and are not trapped by walled gardens” *It would be useful to mention awareness, education and training on these concepts for developers and designers.* In addition we suggest the following edits on page 63: As a tool for any organization regarding these issues, a good starting point is to apply the who, **how,** what, why, and when test to the collection and storage of personal information: 1. Who requires access and for what duration—is it a person, system, regulatory body, legal requirement “or” input to an algorithm? 2. **How is the data accessed/used** ~~What is the purpose for the access~~—is it read, use and discard or collect, use and store? 3. **What is the purpose for the access and** ~~W~~**w**hy is the data required—is it to fulfill compliance, lower risk, because it is monetized, or in order to provide a better service/experience? 4. When will it be collected, for how long will it be kept, when will it be discarded, updated, re-authenticated—how does duration impact the quality and life of the data?

2. What is the purpose for the access—is it read, use and discard or collect, use and store? 3. Why is the data required—is it to fulfill compliance, lower risk, because it is monetized, or in order to provide a better service/experience? 4. When will it be collected, for how long will it be kept, when will it be discarded, updated, re-authenticated—how does duration impact the quality and life of the data?	
Page 64 Issue: Data that appears trivial to share can be used to make inferences that an individual would not wish to share. Background How can individuals be sufficiently informed to give genuine consent? Candidate Recommendation While it is hoped AI/AS that parse and analyze data could also help individuals understand granular level consent in real-time, it is imperative to also put more focus on the point of data collection to minimize long-term risk. Further Resources As analysis becomes more autonomous, not even the analysts will necessarily know what conclusions are being drawn and used in the process. This means that informed consent could become too complex for companies to ask for or consumers to give. This is why we need to move focus away from the consent of the user to the point of data collection. Too much data is collected for no immediate purpose. There needs to be limits and exact purposes for the collection of personal data. Use limitations are also important and may be more feasible than collection limitations. Organizations should commit not to use data to make sensitive inferences or to make important eligibility determinations.	A comment regarding "This means that informed consent could become too complex for companies to ask for or consumers to give" a comment: *How does this fit with legal requirements such as GDPR – is the recommendation that those are changed or that the focus on data collection fits within these legal frameworks?* And regarding "This is why we need to move focus away from the consent of the user to the point of data collection" *Do you mean the amount and/or type of data that is collected at the point of collection? Is the recommendation to move away from reliance on consent? Or to improve consent mechanisms?* *Separately, a discussion re: combinatorial effects of data may be of interest here. As it becomes ever easier to combine data, one can back into a person's identity from information that on its face seems like it would be anonymous. Once we no longer have visibility/control over what combinations occur, we also will not know when a combination could lead to a new "data persona". Can we find a way to program the AI/AS to recognize when a consent might be required from an unintended combination/consequence?*

Page 65 Issue: How can data handlers ensure the consequences (positive and negative) of accessing and collecting data are explicit to an individual in order for truly informed consent to be given? Background It is common for a consumer to consent to the sharing of discrete, apparently meaningless data points like credit card transaction data, answers to test questions, or how many steps they walk. However, once aggregated these data and their associated insights may lead to complex and sensitive conclusions being drawn about individuals that consumers would not have consented to sharing. A clear issue, as computational power increases with time and algorithms improve, is that information that was thought private can be linked to individuals at a later stage in time. Furthermore, as data is stored in terms of summaries rather than as raw observations, and may be key to training algorithms, keeping track of data usage and potential risks to privacy may be increasingly complex. Candidate Recommendations To guard against these types of complexities we need to make consent both conditional and dynamic. Safeguards are required to surface the downstream impact of data that appears to be trivial that can be later used to make inferences that an individual would not wish to share. Likewise, resources and legislation should be afforded to an individual so they can retract or "kill" their data if they feel it is being used in ways they do not understand or desire.	Regarding p. 65 Candidate Recommendations, a comment: *How does this fit with the earlier recommendation to move focus away from consent and on to point of collection?* And regarding "Likewise, resources and legislation should be afforded to an individual so they can retract or "kill" their data if they feel it is being used in ways they do not understand or desire" *This is reflected in GDPR for example, where consent can be withdrawn and where purpose limitations and non-compatible uses must be considered.* *Note however that withdrawing consents becomes harder to do the further downstream one is from the original collection point, especially when data appears innocuous at the point of consent versus when data is combined. Also, how does one back out or "kill" data that by itself appears innocuous until combined? Presumably that is where the guardian platforms come in, but this also emphasizes why this has to be done wholesale – piecemeal will be hard to make effective enough to establish trust.* *Lastly, as discussed previously, there may be situations that may arise in which an individual's desire to retract or "kill" their data also must be balanced against the wider interests and needs from a public policy perspective (e.g., the law enforcement context).*

Page 66-67 Candidate Recommendation Algorithmic guardian platforms should be developed for individuals to curate and share their personal data. Such guardians could provide personal information control to users by helping them track what they have agreed to share and what that means to them while also scanning each user's environment to set personal privacy settings accordingly. The guardian could serve as an educator and negotiator on behalf of its user by suggesting how requested data could be combined with other data that has already been provided, inform the user if data is being used in a way that was not authorized, or make recommendations to the user based on a personal profile. As a negotiator, the guardian could negotiate conditions for sharing data and could include payment to the user as a term, or even retract consent for the use of data previously authorized for a breach of conditions. Nonetheless, the dominant paradigm for personal data models needs to shift to being person based and away from system and service-based models not under the control of the individual/human. Personal data cannot be controlled or understood when fragmented and controlled by a myriad of entities in legal jurisdictions across the world. The object model for personal data should be associated with that person, and under the control of that person utilizing a personalized AI or algorithmic guardian. Specifically: • For purposes of privacy, a person must be able to set up any number of agents/guardians or profiles within one agent with different levels or types of personal data associated. • During the handshake/negotiation between the personal agent and the system or service, if the required data set contains elements the personal agent will not provide, the service may be unavailable.	Regarding "Algorithmic guardian platforms should be developed for individuals to curate and share their personal data." A comment: *This could be tied into the earlier suggestion on education for individuals. Are there any examples (or comparisons from other areas) that could be included here?*

If the recommended data set will not be provided, the service may be degraded. • Default profiles, to protect naive or uninformed users, should provide little or no personal information without explicit action by the personal agent’s owner. Further Resources • We wish to acknowledge Jarno M. Koponen’s articles on Algorithmic Angels that provided inspiration for portions of these ideas. • Companies are already providing solutions for early or partial versions of algorithmic guardians. Anonyome Labs recently announced their SudoApp that leverages strong anonymity and avatar identities to allow users to call, message, email, shop, and pay—safely, securely, and privately. • Tools allowing an individual to create a form of an algorithmic guardian are often labeled as PIMS, or personal information management services. Nesta in the United Kingdom was one of the funders of early research about PIMS conducted by CtrlShift.	

A. This part C sets out Accenture's comments and suggested revisions to Section 7, "Economics/Humanitarian Issues".

Original version	Accenture comments and revisions
Page 83 Technological change is happening too fast for existing methods of (re)training the workforce.	We recommend adding: Technological change is happening too fast for existing methods of (re)training the workforce **and technology is at risk of outpacing human adaptation and causing a disconnect between the opportunities to take advantage of technology and the human ability to evolve without support.**
Page 84 While there is evidence that robots and automation are taking jobs away in various sectors, a more balanced, granular, analytical, and objective treatment of this subject will more effectively help inform policy making, and has been sorely lacking to date.	We recommend adding: While there is evidence that robots and automation are taking jobs away in various sectors, **the evidence also suggests that technology has the potential to augment and humanize other aspects of work and create new jobs in the future.**[44] **A** more balanced, granular, analytical, and objective treatment of this subject will more effectively help inform policy making, and has been sorely lacking to date. A comment: *Seeing AI, a Microsoft research project, uses AI to provide support for the visually impaired. It uses computer vision, image and speech recognition, natural language processing and machine learning to describe a person's surroundings, read text, answer questions and identify emotions on people's faces.*

[44] Cite Accenture research here

	Joonko is an AI powered diversity and inclusion coach that uses a data-driven approach to ensure diversity in recruiting, empowerment, retention and promotion. It analyzes decisions, actions and events ensuring that the data collected is unbiased and fair to all employees.
Page 84 In order to properly understand the impact of robotics/AI on society including those related to employment, it is necessary to consider both product and process innovation as well as wider implications from a global perspective.	We recommend adding: It is necessary to consider both product and process innovation as well as wider societal global perspective. **There are other non-market related AI implications – on the way people live their lives – on physical and mental health, on personal and social interactions and in the fundamental sense of identity not to mention the positive impacts for people with disabilities and its potential to reduce gender bias if designed properly.**
Page 85 AI and autonomous technologies are not equally available worldwide.	A comment: *We recommend adding a reference to an additional digital disadvantage that a lack of access to AI could create (those who are already vulnerable are put at further disadvantage due to lack of access to AI). Also with only 40 percent of the world population connected to the internet and nearly 20 percent unable to read or write, there are additional barriers to ensuring that disadvantaged individuals and communities are not left behind.*

	We recommend adding under the recommendations: **– Actions to build access to AI should take a systemic approach that reflect the foundational requirements (such as access to the internet, literacy) that access to the benefits of AI entail.**
Page 87 **Empowering Developing Nations to Benefit from AI** It is imperative that all humans in any condition around the world are considered in the general development and application of these systems to avoid the risk of bias, classism, and general non-acceptance of these technologies.	A comment: *While the intent to focus on empowering developing nations to benefit from AI is the right one it is too simplistic to apply a one-size-fits all approach. There are nuances by nation within the group of developing countries e.g. India we could argue that India is at the forefront of some of the AI development because of the national policy focused on Digital First and the outsourcing industry, yet it has some of the most impoverished/under skilled people, with high levels of skills inequality – whilst other countries that fall under the umbrella of developing have different profiles and types of challenges.* *Also if we use mobile technology as a proxy there could be opportunities for developing nations to leapfrog to new AI solutions if the supporting infrastructure internet access, energy, access to data) is stable. So the issue is less about access to AI as it is about broadband and energy.*

A. This part D sets out Accenture's comments and suggested revisions to Section 8, "Law".

Original version	**Accenture comments and revisions**
Page 89 The early development of artificial intelligence and autonomous systems (AI/AS) has given rise to many complex ethical problems. These ethical issues almost always directly translate into concrete legal challenges—or they give rise to difficult collateral legal problems. Every ethical issue, at some level of generality, implicates some related legal issue. For instance, the classic "trolley problem" from philosophy has translated into the very urgent need to decide what is legally defensible when an autonomous vehicle is faced with an accident that might harm human beings. Certain decisions which would be acceptable for a human being would not necessarily be tolerated by society when taken by AI or embedded in AIs. In this sense, the recommendations of the Law Committee should be understood as an important complement to the ethics recommendations provided by other Committees. Additionally, we are concerned that some humans are particularly vulnerable in this area, for example children and those with mental and physical disabilities. The development, design, and distribution of AI/AS should fully comply with all applicable international and domestic law. This obvious and deceptively simple observation obscures the many deep challenges AI/AS pose to legal systems; global-, national-, and local-level regulatory capacities; and individual rights and freedoms.	One general comment on this section, which also applies to the entire publication: *We agree that lawyers are professionals trained in ethics and are well-placed to be involved in setting out a framework for ethically aligned design and what Accenture calls "Responsible AI".* *However, we also believe that both the Legal and Compliance functions within organizations are critical to ensuring that AI systems operate in accord with ethical standards and regulations.* *Robust governance structures and compliance approaches are imperative so that we can ensure that ethically aligned design is maintained appropriately as the technology evolves.* *Accenture operates under a set of "Core Values" that we believe can help in setting out approaches to ethically aligned design. Some of these are:* - *Stewardship* *Fulfilling our obligation of building a better, stronger and more durable company for future generations, protecting the Accenture brand, meeting our commitments to stakeholders, acting with an owner mentality, developing our people and helping improve communities and the global environment.*

Our concerns and recommendations fall into three principal areas: 1. Governance and liability 2. Societal impact 3. “Human in the loop” There is much to do for lawyers in this field that thus far has attracted very few practitioners and academics despite being an area of pressing need. Lawyers should be part of discussions on regulation, governance, and domestic and international legislation in these areas and we welcome this opportunity given to us by The IEEE Global Initiative to ensure that the huge benefits available to humanity and our planet from AI/AS are thoughtfully stewarded for the future.	- *Respect for the Individual* *Valuing diversity and unique contributions, fostering a trusting, open and inclusive environment and treating each person in a manner that reflects Accenture’s values.* - *Integrity* *Being ethically unyielding and honest and inspiring trust by saying what we mean, matching our behaviors to our words and taking responsibility for our actions.* *https://www.accenture.com/us-en/company-ethics-code* *Each organization can in fact reinvigorate their core values to adapt to the impact of AI/AS.* *On page 89, “concerns and recommendations” we suggest adding points 4 and 5:* **4. Transparency – end to end understanding of how a particular instance or type of AI works and how to trace and audit decisions** **5. Fairness – mitigating against the possibility of AI incorporating bias or discrimination or otherwise infringing ethical values of fairness, including a right to appeal errors or questionable decisions**

Page 90 Candidate Recommendations Although we acknowledge this cannot be done currently, AI systems should be designed so that they always are able, when asked, to show the registered process which led to their actions to their human user, identify any sources of uncertainty, and state any assumptions they relied upon. Although we acknowledge this cannot be done currently, AI systems should be programmed so that they proactively inform users of such uncertainty even when not asked under certain circumstances. With higher potential risk of economic or physical harm, there should be a lower threshold for proactively informing users of risks and a greater scope of proactive disclosure to the user. Designers should leverage current computer science regarding accountability and verifiability for code. Lawmakers on national, and in particular on international, levels should be encouraged to consider and carefully review a potential need to introduce new regulation where appropriate, including rules subjecting the market launch of new AI/AS driven technology to prior testing and approval by appropriate national and/or international agencies.	*On page 90, we have the following comment:* *When referring to "users" does it include consumers of business services or citizens of governments? Perhaps the term requires a definition.* *We also suggest the following edits:* Most users of AI systems will not be aware of the sources, scale, and significance of uncertainty in AI systems' operations. The proliferation of AI/AS will see an increase in the number of systems that rely on machine learning and other developmental systems**.** ~~whose~~ **AI** actions are not pre-programmed **but should be written to** ~~and that do not~~ produce **traceable** "logs" of how the system reached its current state. **Without auditability**, ~~This~~ **AI processes** will create~~s~~ difficulties for everyone ranging from the engineer to the lawyer in court, not to mention **impeding** ethical issues of ultimate accountability **and the collateral impact on consumer trust**. Although we acknowledge this cannot be done currently, AI systems should be designed so that they always are able, when asked, to show the registered process which led to their actions to their human user **(without requiring companies to disclose proprietary systems in a way that negatively impacts their intellectual property rights and investments)**, identify any sources of uncertainty, and state any assumptions they relied upon **to provide ultimate traceability and an audit trail**. […]

	Lawmakers on national, and in particular on international, levels should be encouraged to consider **and carefully review with industry partners whether there is** a potential need to introduce new regulation where appropriate, including rules subjecting the market launch of new AI/AS driven technology to prior testing and approval by appropriate national and/or international agencies**, over and above the protections provided by industry standards**. **Ethically Aligned Design also requires that both industries developing AI and business stakeholders using AI take concerted action to establish and promulgate best practices and industry standards.**
Page 91 *Issue:* *How to ensure that AI is transparent and respects individual rights? For example, international, national, and local governments are using AI which impinges on the rights of their citizens who should be able to trust the government, and thus the AI, to protect their rights.* *Background* *Government increasingly automates part or all of its decision-making. Law mandates transparency, participation, and accuracy in government decision-making. When government deprives individuals of fundamental rights individuals are owed notice and a chance to be heard to contest those decisions. A key concern is how legal commitments of transparency, participation, and accuracy can be guaranteed when algorithmic- based AI systems make important decisions about individuals.*	

Candidate Recommendations 1. Governments should not employ AI/AS that cannot provide an account of the law and facts essential to decisions or risk scores. The determination of, for example, fraud by a citizen should not be done by statistical analysis alone. Common sense in the AI/AS and an ability to explain its logical reasoning must be required. All decisions taken by governments and any other state authority should be subject to review by a court, irrespective of whether decisions involve the use of AI/AS technology. Given the current abilities of AI/AS, under no circumstances should court decisions be made by such systems. Parties, their lawyers, and courts must have access to all data and information generated and used by AI/AS technologies employed by governments and other state authorities. 2. AI systems should be designed with transparency and accountability as primary objectives. The logic and rules embedded in the system must be available to overseers of systems, if possible. If, however, the system's logic or algorithm cannot be made available for inspection, then alternative ways must be available to uphold the values of transparency. Such systems should be subject to risk assessments and rigorous testing. 3. Individuals should be provided a forum to make a case for extenuating circumstances that the AI system may not appreciate—in other words, a recourse to a human appeal.	With respect to: All decisions taken by governments and any other state authority should be subject to review by a court, irrespective of whether decisions involve the use of AI/AS technology. Comment: *We agree this kind of court review may be needed, but given the earlier discussion about needing to protect the most vulnerable is this setting too high a bar for relief for those who might be most affected by those decisions?*

Policy should not be automated if it has not undergone formal or informal rulemaking procedures, such as interpretative rules and policy statements. 4. Automated systems should generate audit trails recording the facts and law supporting decisions. Audit trails should include a **Page 92** comprehensive history of decisions made in a case, including the identity of individuals who recorded the facts and their assessment of those facts. Audit trails should detail the rules applied in every mini-decision made by the system. Issue: How can AI systems be designed to guarantee legal accountability for harms caused by these systems? Background One of the fundamental assumptions most laws and regulations rely on is that human beings are the ultimate decision makers. As autonomous devices and AI become more sophisticated and ubiquitous, that will increasingly be less true. The AI industry legal counsel should work with legal experts to identify the regulations and laws that will not function properly when the "decision- maker" is a machine and not a person. Candidate Recommendations Any or all of the following can be chosen. The intent here is to provide as many options as possible for a way forward for this principle.	*On page 92, we suggest the following additional point 5:* **5. To fulfil the promise of AI, the private sector must lead in establishing best practices that ensure accountability, integrity and trust. There is a need to educate government policy makers to ensure that they develop public policy which is designed to embrace AI for humans and the world. The private sector can implement a number of options as part of this objective:** ● **Educational forums including business, government and citizens** ● **Review of policy constructs and existing social safety nets to adjust for AI** ● **AI design incentive programs** ● **Assist government programs to create affordable access to data sets**

1. Designers should consider adopting an identity tag standard—that is, no agent should be released without an identity tag to maintain a clear line of legal accountability. 2. Lawmakers and enforcers need to ensure that the implementation of AI systems is not abused as a means to avoid liability of those businesses and entities employing the AI. Regulation should be considered to require a sufficient capitalization or insurance guarantee of an AI system that could be held liable for injuries and damages caused by it. **Page 93** In order to avoid costly lawsuits and very high standards of proof that may unreasonably prevent victims from recovering for damages caused by AI, states should consider implementing a payment system for liable AI similar to the worker's compensation system.	*And the following edits:* One of the fundamental assumptions most laws**, standards** and regulations rely on is that human beings are the ultimate decision makers. As autonomous devices and AI become more sophisticated and ubiquitous, that will increasingly be less true. The AI industry legal counsel should work with **business, technical and** legal experts to identify the **standards,** regulations and laws that will not function properly when the "decision- maker" is a machine and not a person. Candidate Recommendations Any or all of the following can be chosen. The intent here is to provide as many options as possible for a way forward for this principle. 1. Designers should consider adopting an identity tag standard—that is, no agent should be released without an identity tag to maintain a clear line of legal accountability. 2. **Industry, ~~L~~l**awmakers and enforcers need to ensure that the implementation of AI systems is not abused as a means to avoid liability of those businesses and entities employing the AI. ~~Regulation~~ **Standards** should be considered to require a sufficient capitalization or insurance guarantee of an AI system that could be held liable for injuries and damages caused by it.

The standard of evidence necessary to be shown to recover from the payment system would be lower: victims only need to show actual injury or loss and reasonable proof that the AI caused the injury or loss. But in return for easier and faster payments, the payments would be lower than what might be possible in court. This permits the victims to recover faster and easier while also letting AI developers and manufacturers plan for an established potential loss. 4. Companies that use and manufacture AI should be required to establish written policies governing how the AI should be used, who is qualified to use it, what training is required for operators, and what operators and other people can expect from the AI. This will help to give the human operators and beneficiaries an accurate idea of what to expect from the AI while also protecting the companies that make the AI from future litigation. 5. States should not automatically assign liability to the person who turns on the AI. If it is appropriate to assign liability to a person involved in the AI's operation, it is most likely the person who oversees or manages the AI while it operates, who is not necessarily the person who turned it on. 6. Human oversight of AI should only be required when the primary purpose of the AI is to improve human performance or eliminate human error. When the primary purpose of the AI is to provide for human convenience, like autonomous cars, requiring oversight defeats the purpose of the AI.	With respect to this: 5. States should not automatically assign liability to the person who turns on the AI. If it is appropriate to assign liability to a person involved in the AI's operation, it is most likely the person who oversees or manages the AI while it operates, who is not necessarily the person who turned it on. *Comment:* *Do we need to build in additional emphasis to distinguish manages operation of the AI from managing the AI?* With respect to this: When the primary purpose of the AI is to provide for human convenience, like autonomous cars, requiring oversight defeats the purpose of the AI. *Comment:* *Is it really that there is no need/desire for human oversight or just that the type of oversight changes? While we agree that direct human oversight of all decisions/actions would defeat the purpose, presumably there needs to be human oversight during the set up and testing to confirm accuracy/safety before the AI is given a green light and there should remain some sort of oversight should a troubling trend or change in circumstances occur.*

7. Intellectual property statutes should be reviewed to clarify whether amendments are required in relation to the protection of works created by the use of AI. The basic rule should be that when an AI product relies on human interaction to create new content or inventions, the human user is the author or inventor and receives the same intellectual property protection as if he or she had created the content or inventions without any help from AI. Further Resources • Weaver, John Frank. Robots Are People Too: How Siri, Google Car, and Artificial Intelligence Will Force Us to Change Our Laws. Praeger, 2013. **Page 94** Issue: How can autonomous and intelligent systems be designed and deployed in a manner that respects the integrity of personal data? Background AI heightens the risk regarding the integrity of personal data. As consumers, we are worried about privacy but also about the integrity of our data, including the danger of our data being hacked, misused, or even falsified. This is not a concern that is unique to AI, but AI heightens it.	1. Generally, encourage research/measures/ products aiming to ensure data integrity; clarify who owns which data in which situations **and who has responsibility for correcting it**. With respect to this: 2. Discuss regulation and the pros and cons of regulation of data ownership by individuals and companies. Comment: *Re: integrity of personal data, we suggest more of a discussion around accountability with respect to cybersecurity (which may be the point around regulation) and data privacy as opposed to regulation around data ownership – we would argue that existing IP ownership frameworks should be robust enough to incorporate principles around data "ownership."*

Candidate Recommendation 1. Generally, encourage research/measures/ products aiming to ensure data integrity; clarify who owns which data in which situations. 2. Discuss regulation and the pros and cons of regulation of data ownership by individuals and companies. Further Resources • Pasquale, Frank. Black Box Society. Harvard University Press, 2014. • Artificial Intelligence, Robotics, Privacy, and Data Protection, 38th International Conference of Data Protection and Privacy Commissioners, 2016.	

Thomas Dandres, Ph.D.

Research Officer / Agent de Recherche

CIRAIG, Polytechnique Montréal, dép. génie chimique

Ethically Aligned Design

Comment summary:

Introducing AI/AS into the society is expected to make the life of at least some people more pleasant. My concern is that the development of ethical rules should probably not be restricted to human beings. Indeed, introducing AI/AS into the society may also affect the environment in a broader sense. Thus, I recommend to expend the idea that AI/AS should be beneficial to the planet as a whole, including living ecosystems (and human beings) and possibly management of mineral resources. In my vision, AI/AS should globally improve the life of people but also the environment. At least, the development of AI/AS should not harm the environment.

Detailed comments:

Page 2 :

« AI/AS have to behave in a way that is beneficial to people

beyond reaching functional goals and addressing technical problems »

Maybe AI/AS should also benefit to all living species and more broadly to the planet.

Page 5 :

« The modern AI/AS organization should ensure

that human wellbeing, empowerment, and

freedom are at the core of AI/AS development. »

Maybe the planet as a whole should be considered instead of focusing only on the human wellbeing.

« This ethically sound approach

will ensure that an equal balance is struck

between preserving the economic and the social

affordances of AI, for both business and society. »

The environment should also be included in the affordances of AI.

« the key drivers shaping the human-technology

global ecosystem and address economic and

humanitarian ramifications, and to suggest

key opportunities for solutions that could

be implemented by unlocking critical choke

points of tension. »

The environment should also probably be included.

Page 16 « Principle 1 – Human Benefit »

Maybe this should be « global benefit » to include all living species and their environment

Page 24 « Values to be embedded in AIS are not universal »

Maybe the protection of the planet and the environment, and the sustainable use of resources and ecosystems should be considered as a universal value that must be implemented in AIS.

Page 36 « In order to create machines that enhance human wellbeing, empowerment and freedom,

system design methodologies should be extended to put greater emphasis on human

rights, as defined in the Universal Declaration of Human Rights, as a primary form of human

values »

Maybe the machines should enhance more than the human wellbeing : protect the planet and the environment, support the sustainable development, etc.

« It aims to create sustainable systems that are thoroughly scrutinized for social costs and

advantages that will also increase economic value for organizations by embedding human

values in design. »

Sustainable systems mean the environment has to be considered somewhere.

Page 47

« When systems are built that could impact the safety or wellbeing of humans »

Maybe impacts on the environment should also be considered.

Page 49

"should be developed only for the benefit of all humanity and in the service of widely shared ethical ideals."

It should benefit to the all humanity and the environment (to include natural ecosystems, plants and animals).

Page 55

«It is imperative that the pursuit and realization of capable AI systems be done in the service of the equitable, long-term flourishing of civilization.»

Maybe it should be explicitly written that long-term flourishing of civilization implies the protection of the environment (including natural ecosystems, plants and animals).

Page 80

«Dialogue about the effects of technology on people is needed with respect to those technologies that can have a longer term, chronic effect on human wellbeing»

Maybe it should be extended to effects on the environment (including people, ecosystems, plants and animals).

David G. Hunt,
WhyFuture AI Concepts,
Alexis J. Valentin,
The Secretary,
www.whyfuture.com

Designing a strong AI is akin to having an experienced and capable captain navigate a ship of passengers and, whether that ship is on course to the passengers' destinations or not will depend on the strength of the captain's training – how that strong AI is initially designed.

1. Introduction

There have been numerous advancements that have been made regarding artificial intelligence over the passage of time. As a matter of fact, the artificial intelligence research field has been prolific in introducing new and innovative features that have yet to be recognized as AI advancement by the masses despite widespread use. The most familiar of such features include a number of existing online accomplishments such as the use of virtual agents, pattern recognition, and targeted advertising (Martin, 2015). While it is clear that AI already plays a major if understated, role in modern society, ensuring that society is in a position to cope with all these advancements by obtaining a deeper knowledge regarding the processes involved and their importance is vital (Martin, 2015).

The primary objective of computerized reasoning attempts is to create a discerning machine that is fit for planning, thinking, arranging, taking care of issues, thinking dynamically, appreciating complex thoughts, taking in rapidly, and always learning. This amounts to the generally accepted description of human intelligence (Martin, 2015)

2. Concerns for a value-based ethical design

In attempting an ethically-aligned design (EAD), the EAD's principle priority, rightly so, begins with with the consideration of universally accepted concepts of human benefit and "do no harm" (EAD, pp. 16-17) and further specifies a value-based framework for embedding ethical design into AI as outlined in Section 2, (EAD, pp. 22-35).

However, there is yet another concern that is missed out when choosing a value-based framework in attempting an EAD. A set of values that comes without empathic connection, and without prior learned rationale may result in an AI with actions that would merely imitate ethics – rather than actions as a result or intent of true human benefit. While this may still serve the purpose of a functioning AI, it cannot claim to be ethically-aligned, merely ethically compliant.

To illustrate, if a society originally adopts a norm against consuming meat on the basis of ethics, then it can be said to have adopted an ethical value. But if generations of families continue to accept and embed that value into their children to the point that even very young children are socially trained to do the same, then the abhorrence for meat is presented as purely psychological and the avoidance simply imitation. Without the underlying empathy to rationalize this preference, this non-meat value in young children cannot be considered ethically aligned.

3. Altruism as a concept of ethical alignment in favor of value-based AI with perceptions of psychopathy (Candidate recommendation)

The AI should be developed in such a manner that it portrays an extensive and profound aptitude to understand its environments for the purpose of establishing what to do in the different situations that it is likely to come across. This further means that, for the AI to be in a position to comprehend its environment clearly and understand how to respond to these different possible situations, it needs to be socially intelligent as well.

It also needs to be creative since creativity comes in handy when encountering situations that require the management of problems.

For the purpose of realizing all the above-mentioned attributes, it is important to take certain factors into consideration. The first of these factors is the need to look into the traits of altruism vis-à-vis those of psychopathy. It is important to look into human altruistic behavior and make a thorough evaluation in order to be able to profile artificial intelligence around qualities that are considered humane, as well as philanthropic values.

This means that there is a need for thorough research to discover more about the deepest and most intricate foundations of human altruistic behavior. Other factors that ought to be taken into consideration are inclusive of what is generally needed to conclude that a person is altruistic as opposed to selfish. Therefore, in general, when designing an AI, it is imperative that it be shaped around the best and most positive traits of people (Hunt, 2016). This encourages attributes such as compassion, generosity, and the pursuit of equality, among others.

A suggestion would be to complement an existing Candidate Recommendation for prioritization under the Moral Overload issue (EAD, p. 25) with a "lessons learned" database which would allow the AI to compare present conditions with a searchable repository of similar or relevant conditions that also records rationale as well as outcomes or repercussions of actual human decision-making. This is not a perfect method of teaching a machine altruism but it provides a blueprint for the why and if of required decision-making and may not necessarily affect the AI decision.

4. Efficiency kill-switch (Candidate recommendation)

The second factor that should be taken into consideration is the ethical dilemma known as the ethical paradox. This refers to a situation in which there is a need for the AI to choose which action to take: Being diligently efficient or staunchly keeping to its moral obligation. This brings up the issue of psychopathy vis-à-vis empathy. Inasmuch as artificial intelligence ought to be shaped in a manner that makes it efficient, this should, at no time, beat the ability for it to be empathetic when the need arises.

An AI ought to be designed in a manner that allows it to instantly opt out of being efficient in order to show compassion toward someone or people according to the situation at hand (Hunt, 2016). The recommendation of a "lessons learned" database would also help the AI learn this, as it would present historical evidence of what was deemed to be "correct" ethical decisions made by humans.

Consider the following from Foot (1967, p. 3): An airplane pilot has lost nearly all control of the plane. This pilot is presented with a dilemma. The pilot can either steer the plane and crash into a less populated area or do nothing and allow the plane to crash into a more populated area.

According to Hunt (2016), one can see that if the pilot chooses to steer the plane to a less populated area, he or she is more so acting on empathy rather than efficiency. However, in the fat man trolley problem (Thomson, 1985, p. 1409), as explained by Hunt (2016), pushing the fat man over the bridge to save more lives is choosing efficiency over empathy, and most people would reject the notion of pushing the fat man over the bridge as they prefer to place empathy above efficiency. Furthermore, it points to the fact that individuals choosing efficiency over empathy in such a situation correspond with more of a psychopathic mind (Singer, 2005, p. 341, as cited in Greene 2002, p. 178).

There is a different example from Hunt (2016) of an efficiency-over-empathy situation: A doctor is in urgent need of vital organs to save five patients. A person happens to arrive at the clinic with the exact needs of these five people. Should the doctor sacrifice this individual against his or her will and save the five patients in need? Most people would say, “No, it would not be morally permissible for this doctor to proceed” (Thomson, 1985, p. 1396). However, what if the doctor did decide to do this, and this was considered standard practice at the clinic? The clinic would be a place that almost everyone would avoid, and people would not trust the clinic. Thus, there is a need for an AI to place empathy over efficiency.

Therefore, it is important for persons who design AI to be able to structure it in accordance with their defined moral systems as well as the manner in which they are supposed to position themselves depending on different situations that they may face in the future where they will have to make moral decisions (Martin, 2015).

Perception is another important factor to consider when designing AI, since it is through perception that people have the ability to critically evaluate the situations that are presented before them. Therefore, it is important to factor in a dimension of context with a dimension of actions. The lack of an in-depth analysis of context can lead to a conclusion that seems to defy common sense in certain situations. As an illustration of this, suppose an AI is given a task to judge and weigh the positives versus the negatives of a person. One may conclude that the AI is justified when it tallies the negative attributes of a person, such as thievery, and the positive attributes of a person, such as occasional charity. However, consider the following: A person's house burns down. This person becomes frustrated, emotional, and utters foul language in an expression of emotional distraught. This person punches a tree and sobs in the corner over losing everything he or she owns. Another person can understand why he or she is acting like this via the dimension of context, as a tragedy had just befallen to this person. However, the AI's flaw in this situation would be apparent. It would tally the person's actions, such as punching the tree and uttering foul language, and label him or her an undesirable person against evidence that is devoid of context when, in reality, he or she might be an awesome person. This goes to show that it is crucial to design an AI while factoring in an understanding of both the context and actions of a given situation, as this can lead to an AI with a more compatible perception, which is better for humanity (Hunt, 2016).

5. Conclusion

In summary, the concern is that a value-based framework will result in AI that is circumstantially ethically compliant instead of AI that is deliberately ethically aligned.

As such, it is recommended that an AI should be designed with the ability to look at why and how particular systems, beliefs, codes, and values are the way they are for humans and make a decision based on how each particular of these relates to priors. Upon doing this, it can implement its decision based on all of these facts (Hunt, 2016).

References

Foot, P. (1967). *The Problem of Abortion and the Doctrine of the Double Effect*. Oxford Review, (5), 1-7. Retrieved April 21, 2017, from http://pitt.edu/~mthompso/readings/foot.pdf

Greene, J. D. (2002). *The Terrible, Horrible, No Good, Very Bad Truth About Morality, and What to Do About It*. Ph.D. Dissertation. Department of Philosophy, Princeton University. Retrieved from http://emilkirkegaard.dk/en/wp-content/uploads/Joshua-D.-Greene-The-Terrible-Horrible-No-Good-Very-Bad-Truth-about-Morality-and.pdf

Hunt, D. G. (n.d). *The Blueprints Towards the Development of Good Artificial Intelligence.* Retrieved February 16, 2016, from http://www.whyfuture.com/single-post/2016/11/06/The-Blueprints-towards-the-Development-of-Good-Artificial-Intelligence

Martin. (2015). *Artificial Intelligence: A Complete Guide*. Retrieved February 16, 2017, from https://www.cleverism.com/artificial-intelligence-complete-guide/

Pan, L. (2016). *Why China Isn't Hosting Syrian Refugees*. Foreign Policy. Retrieved from http://foreignpolicy.com/2016/02/26/china-host-syrian-islam-refugee-crisis-migrant/

Singer, P. (2005). *Ethics and Intuitions* [Abstract]. The Journal of Ethics, 9, pp. 331-352. doi:10.1007/s10892-005-3508-y

Thomson, J. J. (1985). *The Trolley Problem* [Abstract]. The Yale Law Journal Company, Inc, 94(6), 1395-1415. doi:10.2307/796133

Thank you for this initiative, which is a good cause and not too early. I would like to suggest the following additions to the selection of papers and names that are found in Ethically Aligned Design V1.

1. The works of Ivan J. Jureta. Reason: For the kind of work you envision, a rsolid foundation in requirements engineering will be needed. Jureta's formal, ontology-based approach may just be the right tool.
2. This additional work by Jeroen van den Hoven: "Information technology, privacy, and the protection of personal data", in: Information technology and moral philosophy, e.d van den Hoven, Jeroen (2008). His listing of the different kinds of "information-based harm" could be especially useful.

Best regards,

Kurt Thomas
Bonn, Germany

Ariella Berger
www.unboundedresearch.co
May 2017

Response in consideration to the IEEE's paper "Ethically Aligned Design"

It is with great pleasure that I present comments for consideration to the IEEE, based on your commendable paper "Ethically Aligned Design; A Vision for Prioritizing Human Wellbeing with AI and AS."

I am working an initiative to provide an inclusive platform for ancient traditions to have a "third voice" in AI ethics considerations, alongside companies and governments. Our aim is to provide continuous exposure of AI ethical dilemmas as they emerge to a learned and influential community of industry leaders, thinkers and practitioners of religious and spiritual traditions and to nurture interfaith and philosophic dialogue for the good of humanity in the age of AI. A Q4 17 Jerusalem event ("AI and the Sages") is planned.

Enquiries regarding further ongoing work, speaking, public and private discussions are welcome.

"Berger Thought Experiment" *Referred to in the comments below*

Imagine two training autonomous vehicle (AV) fleets, one in Israel- a land of exceedingly unpredictable drivers - and one in Switzerland- a land of exceedingly predictable drivers. Following the training period, the Israeli trained AV fleet is released on Swiss roads and the Swiss trained AV fleet is released onto Israeli roads. What happens next?

- AV fleets trained alongside unpredictable Israeli drivers should have no issue co-existing with predictable Swiss drivers
- When the Swiss trained AV fleet in Israel, there are two probably outcomes:
- Swiss AV fleet crashes and burns, defeated by Israeli drivers, who go on driving badly.
- Swiss AV fleet survives despite driving very correctly; with enough Swiss AVs on the roads, Israelis begin (grudgingly) to drive predictably, like the Swiss.

Specific Comments on the IEEE Report

p16 How can we ensure that AI/AS do not infringe human rights?
Principle 1 – Human Benefit

p18 How can we assure that AI/AS are accountable?
Principle 2 – Responsibility

Consider euthanasia:

- Holland and Switzerland see euthanasia as a Universal Human Right. The basis is an interpretation of death is an absence of life, making the UHR "right to life" equivalent to the "right to non-life", and so euthanasia.
- Other countries reach an opposite conclusion. Their understanding is more geared towards rights implying obligation. The UHR "right to life" is more conservatively interpreted as an "obligation to protect life."

1. UHRs are not truly universal -they are open to interpretation. If the lack of absolute universality is unrecognized, the task of verifying human right infringement /accountability is vulnerable to manipulation.
2. Despite the semantic distinction, in practice there is a "fuzziness" that blurs principles and cultural bias

p25 Moral overload – AIS are usually subject to a multiplicity of norms and values that may conflict

- An AV uses Waze to drive an optimal route. Say on-board Mobileye system swerves an AV to avoid a collision, taking it off-route. Waze then recalculates the route; the AV then continues its path. An internal AI system manages a hierarchy and enabling Mobileye's collision avoidance to prevail over Waze.
- Emerging behavioral economics has shown the "predictably irrational" nature of human decision-making. Yet human "irrationality" is only so through a linear lens. The reality is, it is an adaptive response to meeting multiple and conflicting human needs.

1. A potentially rich moral system is likely to have a multiplicity of norms and values in conflict with each other, until the right "Macro Moral AI" can oversee it (assuming such an AI at adequate quality can exist).
2. Without patience to await such a "Macro Moral AI", humanity is vulnerable to an inadequate (and therefore ultimately immoral) AI/AS system
3. Whereas an adequate "Macro Moral AI" is the human interest, commercial impetuses would not naturally be geared to paying this burden. There is an argument to be explored surrounding market subsidy.

p84 AI policy may slow innovation
Section 1 – Automation and Employment

Berger Thought Experiment: The more Swiss AVs enter the Israeli fleet, the less complexity the Minimum Viable Product (MVP) of the AV need have. When the Swiss AVs dominate the entire vehicle fleet, drivers will have less freedom to drive unpredictably.

Swapping freedom to drive unpredictably (freedom giving moral autonomy), and reading the AVs as AIs, and complexity of the MVP of the AI as the moral sophistication of the AI; I suggest this illustrates:

1. AI increased penetration correlates with lower minimum standard for AI moral sophistication. The level of AI moral sophistication influences humans, if the two are to co-exist and humans adapt to the AI.
2. The essence of human freedom is based on having choice, implying a certain inefficiency. Robust technology and so AI is marked by maximal achievement of a stated goal. Thus- we must recognize a fundamental friction between the inefficiency of the human experience and the efficiency of the machine.
3. The more technology and AI penetrates human life - particularly spaces where there previously freedom to take moral decisions, the less humans practice moral autonomy.
4. If effective AI policy, aimed at ensuring more sophisticated moral AI - has a price of slowed down innovation, this should be both anticipated and supported.
    - What flag would prompt a deliberate slow-down of innovation, if moral enough AI is not yet there?

- How can these flags be defined when future AI moral dilemmas are to be undiscovered?
- How can necessary innovation slowdown by safeguarded when the global economy is competitive?

p96 Addressing Cultural Bias in the Design of AS. Classical Ethics in Info & Com Technologies

Reconsider the Berger Thought Experiment- where Swiss AVs are good enough to survive Israeli drivers, they impact human drivers to act more predictably and with less freedom to be unpredictable.

1. Cultural bias - as well as being dispersed over different peoples- is also a phenomenon born from how humanity interacts with machines. Technology influences culture and humanity.
2. Moral codes culture and technology co-evolves, in a shared feedback system. With neuroplasticity, I suggest, there is an active feedback system. ("Cyborg Dance")

Proposal to extend AI policy to encompass system redundancy

Berger Thought Experiment	=> Taking => • AVs as a metaphor for AIs • The level of technical sophistication of the AVs as the level of morality of the AIs	Apply metaphor to Berger Thought Experiment, giving insights about morality and AI:
The greater the proportion of AVs in the fleet. the more the technical complexity of the MVP decreases.		A. At a certain penetration of AIs in human use, the minimal operational level of moral sophistication decreases.
When the fleet is 100% AV, the technical complexity of the MVP is the lowest.		B. When AI penetration is 100%, the minimal operational level of moral sophistication is at the lowest.
Once AVs are good enough to survive the human fleet, at a certain tipping point, AV driving behavior influences human driving behavior		C. Once AIs is good enough to co-exist with humans, at a certain tipping point, level of morality in the AI influences levels of human morality

Once AI becomes the Meta Paradigm, it is impossible to roll back. And- as (B) and (C) above show- once the tipping point passes where AI is the Meta Paradigm, AI's moral sophistication drives down, influencing and exposing co-evolving humanity (doing what I coin the "Cyborg Dance").

This strongly suggests a strategic consideration for AI policy to encompass, in addition to a constraints hierarchy, a redundancy based approach.

To understand the impetus, consider the work of Austrian economist Schumpeter, who saw economic systems shaped by a dominant technological paradigm that set the tone for institutions and interests of the entire system, including cultural and moral norms. AI did not exist in his time.

Schumpeter's unspoken assumption, therefore, was that humanity controlled the technological paradigm. To account for this assumption, I suggest re-configuring Schumpeter's model by describing humanity the Meta-Paradigm adopting a Technology Paradigm.

(A), (B) and (C) show the negative impact of AI on human wellbeing; the more AI penetrates humanity, the lower the minimal level of moral sophistication of the AI. At (B) and (C), AI is so pervasive that the morality of AI influences the morality of co-existing humans. The tipping point - where the damage to the wellbeing of humanity is done and is irreversible- is when the Meta Paradigm flips from Humanity to AI.

Thus- I suggest that the IEEE considers its AI policy also encompassing discussion of redundancy, as a defense from this harmful tipping point.

Systemic Redundancy - Layman Explanation
Imagine two connected water reservoirs supplying a town though a single pipe. This system has poor redundancy- a burst pipe cuts of water supply. Imagine a single reservoir with two pipes to the town. Redundancy is also sub-optimal - a poisoned reservoir means no water supply. But if two reservoirs (disconnected from the other) each has pipes to town, redundancy will be superior.

The Biblical Sabbath constraint, preserving one day in seven as the minimal domain of human autonomy with some but not full use of technology, inspires this redundancy strategy.

Implementation of an AI policy related to redundancy should of course be pro AI and in no way neo-Malthusian. Specifics of AI policy design with redundancy considerations merits further exploration and discussion, in my opinion, amongst peers and is beyond the scope of this brief. The means of ensuring and defining redundancy is wide open. Such considerations could match wider societal policy planning such as UBI implementation.

The IEEE Global AI Ethics Japan Committee
Workshop Organizer: Arisa Ema / Katsue Nagakura

Comments from IEEE Workshop in Japan Attendees

Workshop background
Purpose
The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems (“The IEEE Global Initiative”) is a program of The Institute of Electrical and Electronics Engineers, Incorporated (“IEEE”).
Ethically Aligned Design represents the collective input of over one hundred fifty global thought leaders in the fields of Artificial Intelligence, law and ethics, philosophy, and policy from the realms of academia, science, and the government and corporate sectors. The IEEE Global Initiative’s goal is that Ethically Aligned Design will provide insights and recommendations from these peers that provide a key reference for the work of AI/AS technologists in the coming years.
A second goal of The IEEE Global Initiative is to provide recommendations for IEEE Standards based on Ethically Aligned Design. IEEE P7000™ – Model Process for Addressing Ethical Concerns During System Design was the first IEEE Standard Project (approved and in development) inspired by The IEEE Global Initiative. Six further Standards Projects have also been approved, demonstrating The Initiative’s pragmatic influence on articles of AI/AS ethics.

Workshop Purpose
The IEEE Global Initiative realizes the key role Japan plays in global thought leadership in the realms of Artificial Intelligence, Autonomous Systems, robotics and ethics. It is the goal of The IEEE Global Initiative to involve more Japanese colleagues into the process of creating "Ethically Aligned Design, Version 2" as Version 1 of the document had a largely Western perspective.
However, discussion on "Artificial Intelligence and Society" was also being held in Japan in academia, government and companies including the Ethics Committee of the Society for Artificial Intelligence since 2014.
There are a number of perspectives overlapping with what is discussed in this document. On the other hand, it is important to listen to diverse values and at the same time express our opinions on a unique viewpoint from Japan, in order to create rules and standards based on Ethically Aligned Design.

Therefore, we will hold workshops in Nagoya, Kyoto and Tokyo to compile the feedback from Japan that will be presented as part of The IEEE Global Initiative's meeting where discussions for updates of this document will be held.

Workshop program and attendees

2017/04/21 15:30-17:30 @Nagoya 8 people

2017/04/28 16:00-18:00 @Kyoto 11 people

2017/05/02 13:00-15:00 @Tokyo 19 people

2017/05/02 15:00-17:00 @Tokyo 8 people

2017/05/03 10:00-12:00 @Tokyo 8 people

2017/05/03 13:00-15:00 @Tokyo 5 people

Total 44 people

Special Thanks

❖Konstantinos Karachalios, Managing Director of The IEEE Standards Association and Member of the IEEE Management Council

❖Iwao Hyakutake, IEEE Director, Japan Office

❖Raja Chatila, Chair, The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems

❖Kay Firth-Butterfield, Vice-Chair, The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems

❖John C. Havens, Executive Director, The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems

❖Danit Gal, Director, Outreach Committee of The IEEE Global Initiative

❖Ryota Kanai, Founder & CEO of ARAYA

❖Yutaka Matsuo, Chair of the Ethical Committee of Japanese Society for Artificial Intelligence (JSAI)

❖Hiroko Kamide, Chair of Special Interest Group on "Philosophy and Practice for Robotics", The Robotics Society of Japan.

About this comment paper

- This paper aims to introduce comments from various viewpoints by listing comments by each participant rather than summarizing various opinions.

- In writing this paper, we first wrote out the remarks at each workshop and gathered 1) opinions on the "Ethically Aligned Design, Version 1" and its background (general remarks), and 2) the eight chapters. It was released to all workshop participants from May 6th to May 15th, 2017, and received corrections and additions. After that, the organizers reorganized it.

- Contents are written in bullet points. Comments with indentation downwards are statements related to the superior comments.

- The attributes of the speaker are shown in each comment. The comment's attributes are categorized into 6 categories: Academia’s Information System (INFO) (11 people), Academia’s Humanities and Social Sciences (SSH) (12 people), Public Sector (6 people), Industry (5 people), Media (5 people), and others (5 people).

General Remarks
About the title

- Why is it only “Ethic” instead of ELSI (Ethical, Legal and Social Implications)? (Academia SSH)

- Does this document use the term “Aligned” as aligned to “humans” by designing it to harmonize with human beings and the environment? Or used as aligned to “moral value and Ethical principle?” (Academia SSH)

- “Ethical” is a modifier, and so is it rather important to be “Aligned” = “harmonize” to human beings? Is it correct to think “aligned” as "familiar design (najiminoaru)” in Japanese? (Academia SSH)

- What is AI?

- Please indicate what are the problems that are peculiar to AI, that is different from information ethics (Academia SSH)

- There is an issue on “black box”, but it is the same in other machines as well (Academia INFO)

- The characteristic of AI is that the creator of programs will shift from human to IT. Then the problem is that humans can no longer understand nor predict the programs (Public Sector)

- AI has several stages (Accademia INFO)

    1. The stage where the input and the output are clear as humans write the program code
    2. The stage where the parameters are changed by reflecting the learnt data into codes (Provided data becomes important)
    3. The stage where parameters are recreated by themselves from feedbacks from the environment.
    4. The stage where they duplicate themselves from feedbacks from the environment. When it reaches this point, it is singularity. We have entered a little bit on this fourth stage

    - Although the stability of the robot is guaranteed, AI cannot confirm that the answer is "correct". It can only say, "It seems to be correct" and therefore, the stability is not guaranteed (Academia INFO)
    - Impossibility of verification increases for AI. Reproducibility such as behavior in the database disappears. Human ability could not catch up with it. Humans cannot understand the significance of weighting array of deep learning (Industry)
    - The important view point is that AI is networked and is not independent. Self-driving cars are also communicating with each other. Discussions are required on the premise that IoT and etc. are also connected (Academia INFO)
    - The machine that processes data is AI. Discussion on data such as how to collect data is also important (Other)

- Because AI technology is developing, it is difficult to define it. Instead we have to do parallel discussion of ethics and development of technology (Public sector)
- The way of thinking is different, depending on whether you see AI as something "not yet seen, developing" or something "already existing". The report of the Ministry of Internal Affairs and Communications states AI as "already existing", but if including Artificial General Intelligence (AGI) as in this report, the view point of AI becomes something "not yet seen" (Academia SSH)

About the position of EADv1 and sustainability

- It is necessary to have consensus of politics and society to make standards. How to create such a scheme (Are we going to create a third-party institution such as standard institution?) (Academia INFO)
- Isn't it difficult in terms of sustainability on the point that the people contributing to the standardization of IEEE are working voluntarily? (Public Sector)
- The Japanese Association for Artificial Intelligence also issued guidelines, but by issuing such guidelines, the responsibility of the academic community will increase. It is difficult to create a mechanism to ensure that academic societies adhere to the standards such as at the Editorial Committee (Academia INFO)
- Who will detect ethical problems in the first place, what organization can account for responsibility and investigation when problems occur, and who possess such authority? (Academia INFO)
- I would like to have an explanation on where this activity is heading. I would like the organization to disclose easy-to-understand documents such as video messages for late comers. It is difficult to understand the message "Anyone can be involved" when making a standard (Public Sector)

Are there incentives for engineers to comply with EADv1?
Skeptical opinion

- There is no incentive to participate in creating this for engineers (Academia INFO)
- There is a feeling of distrust towards politics from engineers regarding the fact that there is a political consensus first. However, I think standards should be made later (Academia INFO)

- As the private sector cannot participate with too much regulation, we want it to be something more positive (Industry)
- We need guidelines to promote innovation while responding to social anxiety. It is hard to adopt as an industry unless it is made a little more loose (Public Sector)
- Participating in this will be an incentive for researchers in the Humanities and Social Sciences (Academia SSH)
- It is easier for the development side to have whitelists, blacklists, and collections of cases. With too many gray zones, the development side will by daunted (Academia INFO / Public sector)
- Japanese companies have a trauma that they failed technically at the time of the second AI boom, so it may be difficult to come together like the Partnership on AI (Public Sector)

Positive opinion

- Creating possible soft law such as the "regulation sandbox" in the finance industry would be beneficial (Academia SSH)
- Japan's "Special Zone (Tokku)" was written in EADv1. It is also important to build such a mechanism (Academia SSH)
- The model for this is in the privacy principle of OECD. As long as one abides by the rules, they receive an endorsement, therefore it will be an incentive for companies to comply with the rules (Academia SSH)
- It is important to show presence as a group of engineers while the European Parliament, the US and Japanese governments, etc. have issued guidelines and reports already (Academia SSH)
- In the consulting industry, it is easier to proceed with work if there is such an endorsement or a fixed frame beforehand rather than considering framework individually with customers. There are times when it is easier to work with framework and laws (Industry)
- When a standard is created at a tough level, one can be ahead of competition by stating that one has the only technology that can secure the tough level standard (Industry)
- It makes it easier for companies to advance its development by incorporating ethical principles created by IEEE and others, into their company's competence (Industry)

Need explanation of created and excluded articles

- The eight articles should be reasonably ordered. It is unclear what kind of logic this order is in (industry)
- "Ethical norms that AI should have" and "ethical norms that AI developers should have" should be considered separately. Is this the difference between 2 and 3? (Academia INFO)
- Of the topics currently discussed, I want a list of discussions of "It was discussed but did not make it into the 8 chapters (Media)

Need for organizing the target reader

- The discussion becomes distracted unless it is discussed after clarifying to whom (individual, industry, country, human race, etc.) these rules and suggestions are targeting and then discussed (Industry / Academia SSH)
- I want to divide the actors into, for example, development / operation (business), user, policy related doctor, legal person, media, general citizen. Even with the same topic it is easier to understand if it is arranged according to viewpoints of different actors. (Academia SSH)
- Since there are problems that can be resolved by technology and problems that cannot be solved by technology alone (regarding operation and use), please separate them. Such as technical evaluation and user's impact assessment (Academia INFO)
- On the other hand, if you divide them, the responsibility may become ambiguous (Academia SSH)
- The target is stated as AI developer, but who is a developer. Is it the person who writes the algorithm or is it the person who teaches the data? (Public Sector)
- Microsoft's TAY is one example where the given data was not good (Academia INFO)
- I want a fail-safe concept not only from the standard that engineers need to protect but also from the society side. For example, like in the case of self-driving cars where there are compensation and lawsuits. There is a need of a mechanism and standards that has a responsibility of accountability but do not have a burden on users, which allows users to just try using it with no anxiety. Otherwise it is too scary for users as well to use self-driving cars (Academia SSH)
  Discussion of openness and maluses (abuse) should be discussed

- Open source is not mentioned in this document. Because the progress of AI is linked with the Internet, we should also discuss the open culture of the Internet (Academia INFO)
- There is also a possibility that it can be misused by making it open, but we need the viewpoint that it is necessary to make it open to increase the number of people who will protect the technology, in comparison to the people who may use it out of malice (Academia INFO)
- To use it out of malice is not a problem of ethics for development but of the users (Academia INFO)
- This document is in the position of belief that human nature is fundamentally good. Targeting those who will obey the rules and follow them. Therefore, it cannot stop creating a malicious AI system through banning by rules. It is not the idea of binding with ethics, but the idea of getting trust and obtaining allies (supporters) that is constructive. AI vision is necessary for agreement to protect people who protect the AI vision (Others)
- IEEE should protect engineers. There probably would not be an incentive to abide by this code of ethics, unless it is incorporated that they would be protected even if they are misused or if they make careless mistakes (Academia SSH / Media)
- For the message in this document, there is a premise it will succeed if everyone abides by these rules, but it is not so. There is a need of viewpoint on how to make people who do not share the vision, AI is useful for humans, into consideration when creating such rules (Academia SSH)

Other Missing Viewpoints

- Fairness of technical analysis is advanced in Europe but this point is not included. (Academia INFO)
- Discussion on HCI (Human-Computer-Interaction). Affective computing and other story will probably come in to the discussion next, but it is important to talk about the interface that connects AI to the environment (Academia INFO / SSH)

The feedbacks to the Comment

- We would like to have the feedback to this comment and the reason how each topic discussed in Japan were accepted (Public Sector)

1. General Principles

    - The value that general principles lists should be strictly selected. You should write “these are principles that you have to follow” instead of “the issue” (Others)
    - I don’t agree to put the detailed rules like” you cannot do this or that” on General Principles. It is difficult to set the mediate principles between principles that help technologies develop and principles that restraints the technologies’ risks (Public Sector)
    - Considering ‘human rights’, does the state with low awareness of the importance of human rights follow these principles?
    - However, we share the same presupposition that the Universal Declaration of Human Rights is highly valued. Thus, we share the presupposition that we align with these principles to some extent, though each of us has different value (Academia SSH)
    - Does it contradict to set a common standard with considering the diversity? Standard narrow the range of the diversity.

2. Embedding Values into Autonomous Intelligent Systems
    - We can create the artificial Intelligence that make people addicted into something to take the money from them. What kind of rules or norms can prevent the artificial intelligence from taking anti-social actions that is embedded to do so?
    - There are several opinions about the discriminative artificial intelligence like insurance assessment. Even if embedded values at the first stage are business-oriented or social, it could turn out to be anti-social. (Media / Academia INFO)
    - The discriminative artificial intelligence means differently according to the cultural value. (Media / Academia INFO)
    - Though accountability is the important concept that is in accordance with transparency, is it difficult to explain the process of the system in a text level with the current technology? (Academia INFO)

- In a medical field, they judge whether the medicine is effective or not according to the correlation between the medicine and its effectiveness. In the human society, the medicine advances because they trust its correlations. Thus, is it important to consider how the 'trust' structure functions rather than thinking about accountability that does a top-down explanation. Couldn't we move beyond the fear often associated with the AI unless we observe it from a bottom-up viewpoint? (Academia INFO)
- It is impossible for the current technology to lead causation from the correlation between input into and output from the AI. (For example, about the Tesla's accident, people could not understand the reason why the AI on Tesla mistook to judge and made an accident even if they look at the code.) Therefore, there are mainly two ways to manage this problem; (1) budgeting the research for leading causation from the correlation or (2) giving up the research due to its difficulty and discover the other way (like statistical data) to gain the trust. On the current stage, it might be better to pursue both, each government should budget the research that lead (1) and (2). (Public Sector)
- Even if it is difficult to collect the data from the real world, there are some ways to do so by simulating the accident or incident like AlphaGo. It is necessary to make the most of the fact from outside the world. (Academia INFO)
- However, is it different between a cyber-physical (Go) system and a real world (Car).
- It is necessary for self-driving system to set a mechanism that collects data by introducing the system into where the technology is needed such as depopulated area or long-distance track and gains trust. (Academia INFO)
- Basically, it is difficult to maintain the transparency. This document emphasizes accountability, but it is the fact that the industry keeps a certain distance from following such rules because accountability could prevent the industry from advancing the technology or he benefits. (Public Sector)

- Is it possible to create the program that could discriminate or not? It cannot be possible to create such program that is able to deal with all problems. Thus, what we can do is to write or clarify technology limit and adaptation limit at least. (Academia INFO)
- System that stereotypes the real discrimination should not be created. There need some system that could be prevented by human intervention. For example, if there are a people who want to exclude certain organization, it should be warned by the systems. (Academia SSH).
- For example, combination of criminal prediction system and conspiracy charges bill will be very dangerous. Could we prevent developing such kind of systems? (Academia SSH)
- Some states still have female discrimination. Is it necessary to set the rule that avoids creating the artificial intelligence with such the cultural biased data? At least, it is necessary to introduce mechanisms that can always feedback about whether the technology has biased data or not.
- This kind of problem is not a technological problem but a political problem. (Academia INFO)
- How do we think about “people who do not want to use the AI”? Is there any consideration for not have-not but wants-not? There might be some people who do not want to use a nursing care robot.

3. Methodologies to Guide Ethical Research and Design

- Education is necessary to help people be aware that the value of technologists is unconsciously embedded into the design of the AI. It is important to make technologists notice that ‘your value is embedded into the program even if you create the system for a certain purpose or the company’. (Academia INFO)
- It should be written that people in charge of a scientific communication and ethical education is highly valued (Academia SSH)
- Is ethical education included in the curriculum of training a data scientist? (Academia SSH)
- It is difficult to create an educational standard because the requirement of data scientists, for example, is different depending on each case. Is it possible to create an educational standard of ethics? (Academia SSH)

- IEEE did a great work on the ethical education for technologists and it is reflected in Japanese textbooks. However, the value of the ethical education of IEEE sometimes mismatch with that of Japan. (For example, American insider's accusation does not match with Japanese culture.) (Academia SSH)

4. Safety and Beneficence of Artificial General Intelligence (AGI) and Artificial Superintelligence (ASI)

- Why AGI topics comes on 4th? Wouldn't it be OK to be listed at the end? (Media)
- When the word, AGI, comes up, it feels like that people discuss the future rather than the present issues. This may reduces the performance capabilities or binding force of this document, and it feels like not thinking about the AI in the current industry. (Public Sector / Academia INFO)
- On the other hand, other topics are not only specific about the AI. With this topics, it appeals like principles on AI (Academia SSH)
- This topic might be dealing with the fear of general public. (Academia SSH)
- There are fears such as creating "Terminator." They are exemplified by how to manage malicious people or what happens if the human gives the AI a wrong direction. (Media)
- Some threats have happened because the AI was out of control like an economic crisis in Greece, or Tokyo stock exchange. It is dangerous to insist that there is no threat so you don't have to be afraid, rather you should show and announce the threat (Academia INFO / Media)
- AGI is about the future, but like this document, the topic of AGI should be discussed separately from the issue of machine learning in a current business field. Industries we would like to introduce machine learning as soon as possible, so drawing a line between AGI and not AGI is beneficial for the company that want to introduce the technology. (Industry)

- The content of this topic is too shortsighted and near-futuristic. Originally AGI and ASI is much close to the human, but this document has a realistic perspective. I would expect more AGI-like contents. (Academia INFO)

5. Personal Data and Individual Access Control

- The topic related to privacy and personal data has been discussed through several international conferences for 20 years because GDPR and each state has been dealing with the issue. It does not seem that the topic on this document is new. A discussion of the AI and personal data should narrow its focus more. Other international conference related to personal data have started discussing the issue of the AI. Therefore, it might be necessary to corporate with such groups. (Academia INFO / SSH)
- We should consider seriously that the government is collecting personal data (for example, Snowden revealed). Rather that AWS, that is more important to discuss. Personal data was collected contrary to the intention and that could become the targeting data of the autonomous weapon systems. We need to consider new rights for example right of veto over the profiling was admitted at EU GDPR (Academia INFO/SSH)

6. Reframing Autonomous Weapons Systems

- The term ‘Reframing’ about weapons gives the impression that we already accept autonomous weapons. (It sounds strange especially from Japanese perspectives). Should it be better not to give such impression? If there are some risk that is not acceptable, we should put some embargos on the weapon. (Academia SSH)
- We should draw a line of the limit and danger rather than keeping the bottom line (Academia INFO)
- It is prohibited from using the indiscriminate weapon (because the reason why international treaty bans the use of ABC weapons and antipersonnel landmines is not only its excess injury but also indiscrimination). It is necessary to

discuss how to manage the risk because there is a possibility that self-control systems and indiscriminate weapons might become indiscriminate when the accident happens. For example, it supposes to refer to the possibility of malicious people's use for indiscriminate attacks or the risk of indiscriminate attacks due to the trouble of the program. (Academia SSH / public Sector)

- Can we manage the technology when the fight starts? It is written that autonomous weapons easily break out a war. (Academia INFO)
- Ethics and the technology on weapons are still in progress. Though it is ideal that humans can control the weapon, there is no such trustworthy technology. (Public Sector)
- AI distinctive technological discussions should be written down more. Weapons described on this document seem old. Currently it is possible to attack by profiling with personal data. It is close to the discussion of 5 Personal Data. (Academia INFO)
- The offensive weapon like Stuxnet should be discussed (Academia INFO)
- The idea of "Military weapon" is stereotyped (Industry)
- Homing missile is automatic and it is not AI specific issues. There are a lot more other new military missiles that should be discussed in this chapter. (Industry)
- We need to discuss whether it is a problem of technology; what procedure is acceptable; and what kind of reality is being done on situations where physical rights are lost by autonomous weapons (Others)
- How should we consider about the boundaries such as how technically it can be automated whether the AI itself can target danger or whether to input what the human thinks is dangerous? Combination with conspiracy is dangerous regardless of theft or predicton.(Others)
- Dual-use issues should be discussed more.
- There is no way to separate weapons and weapons for civilian uses (Academia INFO)
- It is an economic problem and it is not military problem if you can sell it. After all, it enters in 6 (security) or 7 (economic industry) depending on which is its objective. if it enters in economic industry, talks on weapons can also be

told in the framework of 7. (Media/Academia SSH) Definition that dual-use technology cannot be separated for military uses and for civilian uses. Therefore, we need to think about what purpose it is rather than technology. (Media)

- I would like to know basic design philosophy such as why it is necessary to discuss autonomous weapons. Why is it not simply "prohibition" but "reconstruction"? I would like to know the basic principle first, for example “it is ok to use autonomous weapons for humanitarian purposes”. There is a story that "although it is natural to fight war, because it is humane to make your soldiers not suffer from trauma and stress when they are killed, we use autonomous weapons”.(Media)
- We Japanese did not discussed these issues. That's why, we want opinions and basic principles of why and when autonomous weapons are OK. We can start discussion from the point of whether we agree or disagree with, but since they are not shared, it is difficult for Japanese to discuss. And is the story of autonomous weapons consistent with the story of human rights of general principle (1) ? Or are there fundamental principles such as it is OK that the military uses them but not that terrorists use them. (Media)
- It will be a story of politics but not of technology. (Academia INFO)
- I do not know how much and to what extent engineers can commit to this issues. How much engineers can control? Does the range of technological influence spread beyond what engineers think? (Others)
- There is a premise that human soldiers are autonomous. (Industry)
- I feel that in this chapter, anxiety about being a black box is written rather than anxiety about military weapons. (Industry)

7. Economics/Humanitarian Issues

- There is not much written about employment. (Academia INFO)

- If we are not going to talk about "basic income" on problems of technology and employment, I think that they are not specific to AI.(Media)
- Since the story of employment is a matter of social influence, I think that you should not write about it in this book for engineers. (Media)
- The argument combining with the problems of values in 2 is important; whether it is alright if it has economically profit. WELQ Problem (which is copyright infringement and ethical issue on curation sites that released doubtful medical information) became a problem in Japan because it handled medical information. However, it will happen commonly if people want to make money economically. Will it be a problem if people become addictive to it? (Media)
- Like the discussion in 2, technology becomes anti-moral regardless of the intention of technology design. For example, how much is it allowed to use AI for monitoring and personnel management of specific people? (Academia SSH)
- Will another scale such as well-being is necessary instead of economic rationality? Even if you make other measures like that, is it worthwhile as a business model? (Media)
- How to deal with the calculation of insurance which is "reasonable but inconsistent" calculated by AI. The insurance premium may become unreasonably high due to cancer risk etc. which people are not conscious of. (Others)
- AI also make mistakes. It may be possible to inappropriately raise the charge by interactions with the environment. We tend to have a sense that AI is all-purpose, but it is necessary to be able to set alarm bells ringing and invoke a veto right. (Academia INFO)
- Advocating that AI has uncertainty itself may create new anxiety and distrust. (Academia SSH)

  It is better to show comparisons between what AI can do and what humans can do. It is better to show evidences such as AI makes fewer mistakes than humans do. You need to discuss how to create trusts. (Academia INFO)
- It is said that media must not give incorrect information. It is media and public policies that reach public rather than engineers, so is it engineers to give correct information? (Academia SSH)

- Although you say that you want many kinds of people to participate in discussions, I don't know who are they. (Others)
- Deciding itself whether it is scientifically correct may hinder discussions when you create groups to check media. Do not contradict various discussions? To avoid doing so, it is better to keep it to an extent that let the engineers inform responsibly. (Others)
- Fact check will narrow the discussions? (Public Sector)
- Even if you do a fact check, everyone will not obey it, so it may not be a problem. They think that if it is obviously wrong, they correct it. (Academia SSH)
- It is difficult to understand what is "utopia" and "fact", so I want concrete examples. (Academia SSH)
- Discrimination may be an obvious mistake. (Academia SSH)
- Where will literacy education to uses written? (Academia SSH)
- Is it necessary for engineers to do so? (Academia SSH
- Advertisements can resolve media education for security. (Academia SSH)
- Does literacy education adversely affect the anxiety? Literacy education arises when social problems comes out. That was the same with the Internet. There is no incentive for engineers to embark unless problems happen. (Media)
- Until then it has become an incident base such as responding by self-regulation. (Academia SSH)
- I think that problems such as filter bubbles are peculiar to AI. The filter bubble itself has long been, however, we should argue about it because the speed of AI and the Internet is different. (Academia INFO)

8. Law
   - Why is "law" included in ethics? Cases included in this chapter must have been considered as being better to be dealt by law, but how was that decision made? How were cases divided in to those that can be resolved with law/ethics and those that can be resolved using technology? (Academia SSH)
   - Is there a consensus that "trolley problems" should be considered as a matter of delineation in law or of social decision-making, rather than an ethical issue? (Media)
   - It is possible that the "trolley problem" is considered as a legal issue rather than an ethical one because it is now an urgent

problem. However, in any cases, the author should have stated it clearly. (Academia SSH)

- There is a section where it is mentioned that the issue should be dealt by lawyers. Just like this, the author should first clarify which section s/he wants who to refer to, as well as who are the main actors of the issue. (Academia SSH)
- There must be some issues that should be self-restraint rather than by law. (Academia SSH)
- Despite its appearance at the very beginning, why is the word “liability” hardly included in the rest of the text? How do accountability, legal accountability, and liability relate with each other? (Academia SSH)
- If liability is defined as the obligation of clarifying who is responsible for what, the author should explain why it is included in this text, for it is not a matter that technicians should think about. (Academia SSH)
- Artificial intelligence is not perfect, and there is possibility of causing accidents. In case of such accidents, the public would want to know why that happened, but it is unlikely that sufficient explanation will be provided. Arbitral institutions or investigation systems for medical accidents might become necessary. There are some points where third-party committees are mentioned, but it should be written more specifically. (Media/Industry)
- Standardization of the process is necessary rather than that of the content. For instance, the process of going through an ethical review could be made a standard. (Academia SSH)
- It is essential to clarify the locus of responsibility by, for instance, creating a recursive system which can re-establish the framework. (Academia SSH)
- Is it possible to add more freedom to the framework and make it more abstract like introducing the concept like “by-design” rather than standardizing the process. (Media)
- In that case it would be impossible to know the conditions in case of accident investigations and to see how tight the causal relationship is. (Academia SSH)
- The evolution speed of artificial intelligence technology may be too fast to complete the conventional PDCA cycle each time. An annual inspection would no longer be enough. (Industry)

- It would become more and more difficult to verify technology. However, there are people who nonetheless demand for absolute safety in Japan. (Media)
- On the other hand, Japan is a country which carries out BSE blanket testing for "zero-risks," where individuals can appreciate trust. It may be good to have some countries like this. (Industry)
- Standards should be applied to testable fields where "safety" could be verified. Subjective elements such as the sense of relief should be guaranteed within another scheme. For example, there is a debate over regulation and certification of trust in the movement towards international standardization. It is possible to technologically consider an electrical system which verifies the traceability of trust. (Industry)
- Who is going to do the ex post facto evaluation in contrast to the assessment beforehand? For example, investigation for plane accidents are led by a team of experts appointed by the government. It would be better for the developers as well if there was a fixed system. The system must be able to separate the pursuit of liability from investigation of the cause. (Academia SSH)
- Debate over intellectual property rights and copyrights are also important. (Academia INFO)
- Is it possible to hold artificial intelligence responsible when considering whether your own work is infringing someone else's? (Others)
- Japan is a country profiting from content creation. Systems that permits data collection for AI analysis or makes copyrights for AI creation flexible should be arranged. (Academic INFO)
- The issue of the integrity of personal data had better been discussed in chapter 5. Likewise, the discussion over data protection might be better if it was included in chapter 8, for the law regulation part is the most integral in that section. (Academia SSH)
- I hear the voice from the engineer's side that we cannot predict what will happen by AI. However, if so, does not the law know what to put discipline / regulation subjects on? Although it is said that regulations are delayed, it seems that there is only the way to think out problems later as individual technologies are developed. (Academia SSH)

Attendees List

Naonori Akiya
Daichi Amemiya
Tatsuo Arai
Naoteru Asakawa
Teruyoshi Ehara
Arisa Ema
Yoko Furukawa
Tadao Gen
Hiromitsu Hattori
Kojiro Honda
Iwao Hyakutake
Ryutaro Ichise
Kudo Ikuko
Koji Ikuta
Akiko Kajikawa
Hiroko Kamide
Noritsugu Kanzaki
Takayuki Kato
Makoto Kawai
Kazunori Komatani
Tora Koyama
Minao Kukita
Wataru Mito
Yuko Murakami
Katsue Nagakura
Hiroshi Nakagawa
Kei Narihara
Motoki Ono
Hirotaka Osawa
Takushi Otani
Takumi Oyama
Jun Ozawa
Hiroki Sato
Hase Satoshi
Masatoshi Shimizu
Takashi Sugimoto
Hideaki Takeda
Shinji Takeda

Ikuya Tanaka
Kazuya Tanaka
Etsuko Tane
Rika Wakao
Akio Yajima
Mamoru Yoshizoe