Ethically Aligned Design V2 (EAD) RFI Feedback response

EADv2 Feedback

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 on Ethics of Autonomous and Intelligent Systems* ("The IEEE Global Initiative") titled, *Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems.* The feedback contained in this document was requested regarding *Ethically Aligned Design*, version 2, released in December of 2017.

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 the next version 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 so that these technologies are advanced for the benefit of humanity."

Thank You.

Table of Contents

- Evan Lesmez, Adam Naidorf, Keithen Orson, Wyatt Tinsley, Dustin Weir - School of Engineering and Applied Science, University of Virginia
- Christine Cox, Julia Suozzi, David Rubin, Christopher von Spakovsky, Lucy Fitzgerald, and Mark Restrepo - Institute: School of Engineering and Applied Math, University of Virginia
- Sean Lei, Gabriel Groover, Sina Yazdi, Sam Weber, Eric Xie - Institute: School of Engineering and Applied Science, University of Virginia
- Elise Brosnan, Joses Choy, Anthony Quach, Bowei Sun, Yingxiang Sun - School of Engineering and Applied Science, University of Virginia
- University of Virginia engineering students Angie Campo, Gabriella Greiner, Monique Mezher, Jim Roach, and Noah Rohrlich
- Josiah Perrin, Derek Boylan, Justin Varnum, Harrison Covert - University of Virginia: School of Engineering and Applied Science
- Lucas Abelanet - Joseph Daly - Adam Guo - Ryan Probus - Hans Zhang - University of Virginia
- Kyle Burnett, George Ingber, Kelvin Sparks, Brian Morris - Organization: School of Engineering and Applied Science, University of Virginia
- Siwakorn Chusuwan, Kristina Covington, Charles Pritchett, C. Graham Muller, Charles Yu - 4th year engineering students, School of Engineering and Applied Science, University of Virginia
- Gwen Ottinger, Associate Professor, Department of Politics, Center for Science, Technology, and Society, Drexel University
- Samuel Boakye, Kareem El-Ghazawi, Chase Deets, Henry Hubler, Raquel Moya - School of Engineering & Applied Science, University of Virginia
- Matthew R. Anderson, Andreas L. Butler, Andrew G. Coffee, Paul J. Hughes, John R. Walnut, - University of Virginia
- Pradyot Sahu - Director, 3innovate, India
- Tony Nguyen, Chris Anton, Chris Mooney, Yihnew Eshetu, Martin Simpkins - School of Engineering & Applied Science, University of Virginia
- Rod Rivers, Socio-Technical Systems
- Berkeley Fergusson, Christian Halsey, Clark Kipp, Robert Wallace - University of Virginia

- Folke Hermansson Snickars - Standards ambassador, MyData Global
- Martin Peterson - *Texas A&M University*
- Sriraj Aiyer, University College London
- Gonzalo Génova (ggenova@inf.uc3m.es), Departamento de Informática, Universidad Carlos III de Madrid, Spain; M. Rosario González (marrgonz@ucm.es), Departamento de Estudios Educativos, Universidad Complutense de Madrid, Spain
- David J. Gunkel. PhD - dgunkel@niu.edu http://gunkelweb.com, Distinguished Teaching Professor of Communication Technology, Northern Illinois University USA – http://www.niu.edu
- Zvikomborero Murahwi
- Paola Di Maio, PhD, ISTCS.ORG
- Marcel Bullinga | www.futurecheck.com | info@futurecheck.nl | 0031-6-29552946 | @futurecheck
- Claude Cloutier: XtremeEDA Corporation
- Intel Contributors
- Lachlann Tierney, Researcher at the Institute of Public Affairs (ipa.org.au) in Melbourne, Australia
- Daniele Andresciani, IIT (Italian Institute of Technology) of Genova
- Joachim Iden
- Pavel M. Gotovtsev, PhD, Biotechnology and bioenergy department, National Research Centre "Kurchatov Institute" and Valery E. Karpov, PhD, Neurocognitive Sciences and Intelligent Systems Department, National Research Centre "Kurchatov Institute"
- Agata Piekut, Health Action Tank
- Matthew Newman, Profession: Founder Shared Intelligence
- Randy k Rannow
- Prof. Dr. oec. HSG Oliver Bendel, University of Applied Sciences and Arts, School of Business
- Randall Parker, Open Simulated General Intelligence (SGI)

- Bruno Macedo Nathansohn, Brazilian Institute of Information in Science and Technology (IBICT)
- Manfred Bürger, Stuttgart, Germany, physicist, retired leader of nuclear reactor safety department at IKE- Inst. Nuclear Energy, Univ. Stuttgart
- Donovan Anderson, Project Administrative Assistant - Responsible Ethical Learning with Robots Centre for Computing and Social Responsibility
- Cathy R. Cobey, Partner, EY
- Marek Havrda, Ph.D. and Olga Afanasjeva, GoodAI
- Babita Ramlal, Ontario Ministry of the Attorney General, Innovation Office
- Dr. Ilana kepten, ORT Braude Engineering College, Israel
- Pradyot Sahu, Director, 3innovate, India
- Jamie Williams, J.D.; Jeremy Gillula, Ph.D.; Lena Gunn, Electronic Frontier Foundation
- Emerson Rocha, Etica.AI
- Angeles Manjarrés, Departamento de Inteligencia Artificial. Universidad Nacional de Educación a Distancia (UNED) de España; Simon Pickin, Departamento de Sistemas Informáticos y Computación. Universidad Complutense de Madrid, España; and, Miguel A. Artaso, Departamento de Inteligencia Artificial. Universidad Nacional de Educación a Distancia (UNED) de España
- Tomas Jucha; Investment Protection and AI Policy Advisor assisting Deputy Prime Minister's Office of the Slovak Republic for Investment and Informatization on AI Policy Matters
- Dr Ozlem Ulgen, Senior Lecturer in Law, School of Law, Birmingham City University, UK, and Visiting Fellow at Wolfson College, University of Cambridge
- Yannick Fourastier yannick.fourastier@rail.bombardier.com
- Yannick Fourastier and Bob Donaldson
- Emil P. Vlad P. Eng.
- Erica Southgate, PhD, Associate Professor of Education, University of Newcastle, Australia

- Ioannis C.MATSAS, (University of Cyprus, CY); PhD Candidature in the field of Ethics in Transport Innovations (Aristoteles University, GR)
- AI4ALL submission

12.14.2017
Karen Rivoire: Chief People Officer and founder of Human Revolutionary

Dear John and Co.

I have a degree in European Business and 27 years of marketing and HR experience working with some of the great organisations across the world – Unilever, Sony, WPP. My feedback comes from a place of care and concern with deep understanding of human behaviour and organisation transformation. I am very happy to see the metric around well-being as central to a new definition of progress. Below are some structural inputs and I ill leave some feedback comments on format an imagery that is currently shaping bias and cultural norms.

1. Representation – In the same way more global input was sought after V1 I would like to see more people outside of CS / Engineering. The report rightly insists on multi-disciplinary teams for A/IS but we also need economists, anthropologists and other social scientists if we are to achieve progress in wellbeing. I would also add some functional expertise in HR, marketing, finance. Happy to represent progressive HR.
2. In the principles and sections executive summary (P.18) I would add simple questions. Like the 2 simple questions from Kate Crawford or the 5 AI protocol questions by Robbie Stamp. I would also add the question of "what will we not do? These are "principles into action" and invite systemic inquiry. A one pager of questions will change behaviour more effectively than a full report.
3. (P.23) Given the importance of better metrics for measuring progress, what do we have as a baseline measure now?
4. (P.33) Do we want to "add human norms and values into A/I systems" or rather ensure that every person in every multi-disciplinary design team thinks about the questions above and the consequences of what they are doing?

5. Where is the part on ethics of classification? Reference to work by Kate Crawford, NYU
6. Where is the part on the ethics of imagery /representation. We know from work on sustainability that fear mongering does not work. We have to stop the imagery of men and robots taking over! (P.51) is not enough.
7. (P.61) Great to see the importance of value-driven leadership. This should be embedded in every organisation as a way to close the gap between ethics and company purpose or ethics and business model. Experience in big corporates tells me that appointing one CVO is not the solution. We have too many Chefs and not enough cooks already!!
8. (P.64) Governance in a A/IS world is not about watchdogs and bodies it is about empowering every single individual. You talk about empowerment but this needs to be shown in the recommendation of how to bring this alive. See BIOSS International's 30 year experience in judgement, review and coherence. These are tools for existing bodies whose "work" needs to change and whose composition needs to change but we do not need any more surveillance!

Apologies if some of this comes across as opinions but it is steeped in experience of great people work that was documented but not always broadcasted!

Kind Regards
Karen

12.19.2017
Ross Nippoldt: Software Test Engineer, Thomson Reuters

Reference: Document as a whole: *Ethically Aligned Design*: *A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*

When banking is left to bankers disaster follows. When automotive manufacturing is left to manufacturers disaster follows. The same is true for Pharmaceuticals, food companies, sports doping, tobacco…the list of self-regulating failures has neither beginning nor end.

Trusting self-regulation has never been wise. I applauded the IEEE for stepping up to address human rights, transparency, and verifiability in autonomous systems. I would like to see IEEE acknowledge this self-regulating short-coming and recommend a way to create non-biased third party checks.

Ross Nippoldt
Software Test Engineer
Thomson Reuters
Phone: 651-848-7560
ross.nippoldt@thomsonreuters.com
thomsonreuters.com

1.30.2018
Emil P. Vlad P. Eng.

Emil P. Vlad P. Eng.
Comments on IEEE_EAD_v2

1. on p.9, first para. from "Well-being Promoted by Economic Effects " - Somewhere in this paragraph some essential concrete wellbeing attributes should be mentioned (at least as examples within brackets) like health, safety (cyber)security. Especially considering these are key core humankind wellbeing attributes that are also ethically valued across the globe; also considering these are the key desired non-functional attributes/properties currently regulated and engineered in existing (currently in use or development) automation/automated systems.

2. on p.10, first para. from "Policies for Education and Awareness" - Also physical not only cybersecurity should be mentioned in this list as equally important; also health and environmental protection too.

3. on p.11, first para. from "Well-being Metrics" - For consistency with my previous comment on the related Well-being paragraph from section III; I suggest to add (cyber)security to the list of key wellbeing metrics, which are already managed and known in the context of existing automation systems.

4. on p.14, first para. from "Educational materials" - This metaphoric "Evergreen in nature" expression, should be replaced by a less metaphorical one that is more precise/accurate and literal; considering this is a document from an international engineering organizations and for many members and even so more users English it's not maternal language so literary subtleties obfuscate the semantic clarity.

5. on p.23, first para. from "General Principles" - I suggest to add on this page or somewhere considered appropriate in this section as early statement reflecting maybe in more detail the principle that: because A/IS are a subset of automated systems: the autonomous, and highly intelligent ones;

all applicable ethical and dependability rules and practices know from the development and use or exiting automated systems should be applicable by extrapolation and adaptation. This would be a legitimate and important for two reasons

- it is a legitimate application of the existing GAMAB system safety principle;
- it would encourage engineers to build upon (instead reinventing the wheel) exiting knowledge and practices accumulated over at least a couple of decades of automation by avoiding mistakes that were paid dearly with accidents"

6. on p.34, first para. from "Issue: How can we extend the benefits and minimize the risks of A/IS technology being misused?" - Safety is barely mentioned here on this page, however it is very important and should be added besides security. Security is intentional misuse but safety covers unintentional misuse.

7. on p.49, first para. from "Embedding Values into Autonomous Intelligent Systems, regarding Transparency as intelligibility" - Although the AI design/algorithm/solution may be totally transparent it might unintelligible to humans. Thus intelligibility might be very limited in some very advanced AIs, as it is already the case in deep neural networks. The smarter the AI becomes the more complex their design patterns and algorithms will intrinsically be. Practically is quite likely impossible to achieve because in some applications the reason AI is used is exactly to address extremely complex problems well beyond human ability to solve; these problems may very likely equally incomprehensible solutions to humans. As long as all affected humans are informed and participate in taking the decision, this is more of a risk/benefit trade off issue than an ethical issue.

8. on p.50, first para. regarding "fail-safe" - The fail-safe principle is widely used in many automation systems currently so it will be applicable to AI systems as well, but this should be reformulate by also adding operate-safe as some (e.g. airplane FMS) automated system cause accidents if they fail

to operate. They have to actively operate safe in some back-up or degraded mode but complete failure (system stopped) is not safe option.

9. on p.53, first para. from "Evaluating the Implementation of A/IS" – I suggest to add in this section a sentence about the applicability of the already largely used PDCA organization management system practices (maybe reference some of the related well known ISO std. 9001, 18001, 14001, etc.) allowing for systematic assurance of desired goals and attributes for products. These are applied to existing automation products so all that experience should be applicable and adapted to A/IS.

10. on p.58, second para. from "Methodologies to Guide Ethical Research and Design, " - Noble intent in theory but I think it is unrealistically difficult to be applied in practice in the current geopolitical world. As it doesn't seem reasonable to expect that rapid increase in automation and robotization will slow down and wait for the world to become one, I suggest to add or complement this section with some more practical guidelines (similar to existing configurable systems) and principles along the following lines:

    - A/IS should be developed and produced ethically neutral by engineers & suppliers which should strive to do that to the largest extent possible;

    - A/IS should be designed to be highly configurable in general but in special with respect with the ethical values;

    - A/IS should be delivered to users with comprehensive manuals and training including about the configuration of the desired ethical values to be loaded as parameters by the users."

11. on p.61, first para. from "Issue: The need to differentiate culturally distinctive value embedded in AI design." - As mentioned in a previous comment considering the variation of ethical norms across the globe, this could be practically done by "ethically" configurable A/IS.

12. on p.62, first para. from "Methodologies to Guide Ethical Research and Design" - A suggested guideline should be added here to address this issue

stating: that functionally generic and parametrized A/IS are more neutral ethically thus transferring as much as possible the ethical responsibility of the technology (A/IS) to the user; the ideal to strive for should be for example, that of a knife which is technology (albeit old and rudimentary now, it was revolutionary when invented/discovered) is ethically neutral and it can be used for both good and bad depending entirely of the user's ethics.

13. on p.67, first para. from "Lack of ownership or responsibility from the tech community." - This should be removed or rephrased because legal is different than ethical, while I don't think that safety should be taken for granted; as it's not a given and tremendous effort is required to instill it in organizations and products currently.

14. on p.71, first para. from "Issue: Poor documentation hinders ethical design." - That is true but it should be added this situation applicable to A/IS is no different than the current discrepancy in most complex systems and automation systems today: in theory there are regulation and guidelines for documenting the systems (sw, hw, etc.) but in practice it either lacks, or it's incomplete or inaccurate, especially for SW. This is due to the fast pace of technological change and prioritizing resources (schedule and cost) in enterprises.

15. on p.72, first para. from "Issue: Inconsistent or lacking oversight for algorithms. " - That is true, but practically what is the guideline for already existing cases (deep neural networks, evolutionary computing) where the A/IS algorithm is just inherently so complex that humans cannot comprehend it. This poses a crucial risk/benefit dilemma as one of the main reasons for developing A/IS is the belief that it will be capable to solve very complex problems that humans (even in large teams) can't.

16. on p.82, first para. from "Candidate Recommendation" - That is recommendable in theory, but in practice even in current complex and automated systems not always done; often not because of engineers, but because managers "myopia" justified by schedule and cost reasons. Because these are in turn driven by customer and market competitive pressure, I think the most effective way to counteract is through legal

economical means/measures that would internalize (currently externalities from a market perspective) health, safety environment risk/accident costs. If these basic ethics attributes cannot be instilled in engineered systems how will more general ethics values be?

17. on p.106, fourth para. regarding "service may be degraded" - This should be generalized beyond “personal data” into a general automation (widely applied currently) principle, by adding here or in a more general section of this document a statement like: wherever possible (the control or search algorithm isn't beyond human capability) the A/IS should have a manual bypass degraded mode functionality allowing users to override it, whenever necessary in exceptional circumstances.

18. on p.199, fourth para. regarding "free will" - However, based on recent scientific free will seems to be more constrained (deterministically or statistically) laws of the universe than, construed in classical philosophy and psychology.

2.3.2018
Timothy Rue

In Response to IEEE Request for Input on; Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (A/IS)

This response is not to any specific section but regarding the Fundamental A.I. Ethics Foundation:

To fully address A.I. Ethics the foundation upon which A.I. is built upon needs a fundamental ethics correction.

To use an analogy, where is the painter ability to paint a rainbow if they are allowed only two of the three primary colors? Likewise, how well can an end user benefit from only having access to two of the three primary user interfaces?

The usefulness of a computer is based on how well it can automate and be used to automate!

The two standard user interfaces commonly available are the command line type of interface and the Graphical User Interface type. The third, but missing, a standard user interface is the Applications, Libraries, and Devices side door port. The user-oriented, easy to use Inter-Process Communication port.

As the end-user has access to all the functionality the programmers allow the user via the two primary interfaces, there is no reason to not also allow the user access to all this same functionality in a manner that allows the user to automate not only within an application but across applications, even allowing direct access to the functionality of function libraries and devices.

Q: Why force the end-user to manually do that which they can, if they so choose, automate?
A: The way to become wealthy is to make people need you.

A Fundamental Ethics violation. Why? How many personal automation one-offs will no programmer take on because there is no money in it for them or the company they work for? No one knows because the concern was denied! Why should programmers need to do so when ethically the end-user should have a standard way for themselves to do so.

Why is this relevant to A.I. Ethics? To fully understand the answer requires reviewing and projecting the evolution of common computer usage had this third end-user interface been standard. Today we are projecting all sorts of ethical concerns over the future use of A.I., yet, the ethical issue upon which AI is built and sits upon is incorrect. Build a house on a bad foundation, it will fall or always have problems.

While A.I. is so often compared to, and an effort to be designed to emulate, human thought processes, the end-users have been denied to do so, to apply their thought processes via automation. What insights have been missed about the bridge between computer functionality and human thought process due to the end-user denial?

The following "Action Constants" are unavoidable in human thought processes and computers have to be instructed to use these. To prove the "unavoidable" simply try NOT using even just one. These make up the fundamental elements of Abstraction Physics [1].

0) Defining a word to mean a more complex definition (word = definition, function-name = actions to take, etc.)
1) Starting and Stopping the interfacing with abstraction definition sequences.
2) Keeping track of where you are in the progress of abstraction sequence usage (moving from one abstraction to another).
3) Defining and changing "input from" direction.
4) Defining and changing "output to" direction.
5) Getting input to process (using variables or placeholders to carry values).
6) Sequentially stepping through abstraction/automation details (inherently includes optionally sending output).
7) Looking up the meaning of a word or symbol (abstraction) so to determine action upon or with it.

8) Identifying an abstraction or real item value so to determine action upon it.
9) Putting constraints upon your abstraction lookups and identifications -When you look up a word in a dictionary you don't start at the beginning of the dictionary, but begin with the section that starts with the first letter then followed by the second, etc., and when you open a box with many items to stock, you identify each so to know where to put it in stock.

The Action Constants made available as a command set in a small shell, and configured in a simple logical manner to provide ultimate versatility and exception handling within the inherent constraints of the computer provide for a "Virtual Interaction Configuration."

This Virtual Interaction Configuration used as a standard means of accessing the vocabulary or dictionary set of available functionality of the various applications, libraries, and devices. This includes documentation, example usage and more. But the configuration also allows the ability of the end-user to define their own automatons within and between applications, libraries and devices. And all of this done in a common & consistent manner. Of course, given the unavoidable action constants in such a configuration enables more than automation, up to the user own imagination, but as well enables Knowledge Navigational Mapping. Not so unlike online Wikipedia style hyperlinks where links can be made to yet to be created information, but much more as automatons are possible, the creation of automated loops and cycles created by even the typical end-user.

How might software development have evolved had not this third primary user interface not been denied the end-user? It cannot be said what all would have come about in software development evolution but some things can be understood with certainty.

A higher degree of genuine software engineering as applications, function libraries, and device interfaces would have had a greater focus on integration capability.

A far better understanding of automation by the end-users resulting in greater adaptability and acceptability of A.I. As well the ranges of application of A.I. use would be greater today.

A.I. development most probably would have evolved in a different direction, a wider scope direction, but certainly the ethics issue so widely and wide scope being

discussed today would not be what it currently is. The scope of A.I. ethics would be far better defined in terms of automation, and as such more manageable and inherently enforceable as end-users could not be so fooled with the promoted illusions given today about A.I.

Better late than never. Regardless of A.I. calculation speed today (i.e. Deep Learning), this does not dismiss the applicable viability of the unavoidable action constants, the Virtual Interaction Configuration nor Knowledge Navigational Mapping, as such A.I. is also accessible and map-able for wider scope automation. Tortoise and hare teamwork. Perhaps with human creativity and ingenuity removing the A.I. black box issue is reasonably possible.

[1] http://AbstractionPhysics.net

2.4.2018
Lionel P. Robert, Jr.

Lionel P. Robert Jr. University of Michigan School of Information

In reference to page 27, principle 3 accountability recommendation 1 which starts with "*Legislatures/courts should clarify issues*....."

The statement seems to push off issues of accountability to the local, state and federal courts/governments. There is no reference to the role of professional societies or universities. Nothing in the remaining recommendations really addresses this shortcoming. We seem to be pushing issues of accountability off to someone other than us (i.e. members of universities and professional societies). This is a bit disappointing.

Best regards,

Lionel

New Paper(s): Robots are Here...:)

You, S., Ye, T., Robert, L. P. (2017). **Team Potency and Ethnic Diversity in Robot-Supported Dyadic Teams**, *Proceedings of the 38th International Conference on Information Systems* (***ICIS 2017***), Dec 10-13, Seoul, Korea. Link to copy provided by the author http://hdl.handle.net/2027.42/138124.

You, S. and Robert, L. P. (accepted in 2017). **Emotional Attachment, Performance, and Viability in Teams Collaborating with Embodied Physical Action (EPA) Robots**, *Journal of the Association for Information Systems*, (***JAIS***), link to the preprint version http://hdl.handle.net/2027.42/136918.

You, S. and Robert, L. P. (2017). **Teaming Up with Robots: An IMOI (Inputs-Mediators-Outputs-Inputs) Framework of Human-Robot Teamwork**, International Journal of Robotic Engineering, (***IJRE***), 2(3), link to the preprint version http://hdl.handle.net/2027.42/138192.

Robert, L. P. (2017). **The Growing Problem of Humanizing Robots**, *International Robotics & Automation Journal,* (***IRAJ***), 3(1), Article 43, http://dx.doi.org/10.15406/iratj.2017.03.00043, link to the author's copy http://hdl.handle.net/2027.42/138018.

2.25.2018
Nathan Obrien

Good Robots: Why Empathy in Autonomous Systems Matters
Nathan Obrien

Once artificially intelligent systems proved their usefulness by speeding up assembly-lines and making complex computations more quickly than any human ever could, it became clear they would eventually be much more than unpaid laborers and glorified calculators. However, the more responsibility and autonomy they are given, the more AI systems will need ethics systems designed to guide their decision-making processes.

Running a Tight Ship
(Value Alignment)

Imagine yourself on a cruise ship. As you relax on deck soaking up the sun and sipping rum cocktails, the thought of something bad happening is the furthest thing from your mind…*but why?* It's Because most people (aside from those who've seen *Titanic* one too many times) trust that the captain's training, experience, and more importantly, *moral sense,* will help him or her make decisions which are in the best interests of everyone on board. With no reason to believe the one steering the ship would intentionally put them in harm's way, thousands of people continue to set sail each day, confident they will make it home because the captain's values system (at least where human life is concerned) is in alignment with their own (Hunt & Valentin, 2017).

Ethics and morals are no longer just the stuff of humans. The more autonomous systems are called upon to work alongside people, the more people will ask *where did my robo-partner get its moral sense from?* As AI begins to drive our cars, fight our wars, and give us that much-needed gall-bladder surgery, we need to know that the answer goes deeper than just a few carefully-crafted if-then statements (Ethically-aligned design, 2001).

Coeckelbergh (2012) reminds us that above all, AI systems are tools for achieving *human goals*. However, because ethical norms can differ from place to place and person to person, it is important for those who design these systems to remember that decision-making is not just a matter of black and white. AI must be able to understand and imitate the moral language of the culture in which it is deployed if we expect it to successfully power its way through moral dilemmas and ethical gray areas. (Coeckelbergh, 2012).

For example: imagine that a robot designed to assist with labor and delivery is purchased by an obstetrics center in a small town. Soon afterward, the ROBOGYN suggest a blood transfusion as the best course of action for a mother who is experiencing severe bleeding. Despite delivering what it determined was the best *medical* advice, in this case, the robot would be no closer to saving the woman's life. Why? Because the mother is a practicing member of the Jehovah's Witnesses, (a religious sect with a larger-than-average population in that area), and members of her faith do not accept transfusions for religious reasons. Because it has never been taught what to do in such scenario, and the robot is also unable to suggest another possible course of treatment, putting the mother's life at further risk.

This example suggests that as AI evolves, diversity in design will become an imperative. While roboticists and programmers will always remain vital of any design team, the need for culturally sensitive machine ethics will necessitate they work alongside individuals such as philosophers and ethicists (Deng, 2015). It is not enough for these systems to be outsourced or added to an AI system just before they hit the market. To do what they are intended to do in an acceptable way, ethical controls they must be an incorporated into a bottom-up design process from the very beginning (Staht & Coeckelbergh, 2016).

Mirror, Mirror:
(Teaching Ethics and Empathy to AI)

One of the biggest those involved in the field of machine ethics face is deciding how to teach autonomous systems the values they will need to make good, ethical, and culturally-sensitive decisions. Asaro (2006) suggests that for

systems to make ethical decisions and continually improve their moral sense, they will need to be able to incorporate the lessons they learn into their ethics systems, or even begin to evolve their own ethical guidelines. Perhaps unsurprisingly, some suggest that the best model for such systems may be the human mind.

One of the most difficult traits to recreate in autonomous systems is. This distinctly-human phenomenon takes two distinct forms: *perceptual,* in which we observe a situation and imitate the emotions of those involved in it, and *imaginative*, which allows us to "put ourselves in another's shoes" so we are immediately able to empathize with them. Perceptual empathy is already present in some existing AI technologies. However, because it relies entirely on imitation, it lacks the potential for systems to continue developing its moral sense based on what it observes.

The realization that perceptual empathy is limited has led researchers to believe that imaginative empathy offers an ideal model for empathy in autonomous systems as of the way it *facilitates learning*. Unfortunately, it has proven difficult to reproduce accurately in AI (a problem given the increasing rate at which autonomous systems are interacting with humans (Hunt & Valentin, 2017), (Ethically-aligned design, 2001). For now, it is necessary to outfit autonomous systems with a more rudimentary ethical system; training them on data such as human consensus gathered about certain ethical issues, or the outcomes of past ethical dilemmas.

For instance: imagine yourself at a major-league baseball game when the pitcher is suddenly hit in the head by a line drive. Although you haven't been hit, seeing the look of pain on the man's face immediately causes your body to tense, your heartrate to speed up, and your face to adopt the same, pained expression. Why does the human body react in this way?

It all has to do with *mirror neurons.* First discovered in the 1990s, this new type of visuomotor neuron responds in the same way whether someone is observing an action or completing the action themselves (Winerman, 2005), (Iacoboni, 2009). When humans (and some monkeys) observe what others are

doing and feeling, it creates new neural pathways which help them understand the situation, so they can develop the proper empathetic responses to it (Rizzolati & Fabbri-Desto, 2008). It's been suggested that because mirror neurons help humans acquire empathetic norms, they provide an excellent model for AI systems to do so as well (Hock, Stocker, & Larkin, 2008). Learning algorithms modeled after mirror neuron function would have another advantage; they could enable system to continue learning and refining themselves based on what they learn in each new situation (Anderson & Anderson, 2004).

What Could Go Wrong?
(Empathy Versus Efficiency)

At the risk of sounding like someone who's read too many comic books, *"with great power comes great responsibility."* Whether caring for sick people, driving cars, or deciding when to pull the trigger in the heat of battle, greater machine autonomy means more instances in which machines are required to make difficult, ethical decisions. Moreover, that ethical controls and oversight systems (Etzioni & Etzioni, 2016) which ensure AI can interact safely and humans and adhere to social norms are needed sooner rather than later (Anderson & Anderson, 2004). Serious safety research must counterbalance our inability to predict all a system's possible behaviors (Ethically-aligned design, 2017). Overall, for people to learn to trust and accept AI into their everyday lives (Deng, 2015), it must be both safe and ethical by default (Coeckelbergh, 2012).

Autonomous weapons systems raise some of the most pertinent questions about the capacity of machines for ethical decision-making. While weapons are obviously designed to destroy targets (human ones included) and eliminate threats, what's to keep them from killing of innocent civilians in pursuit of their objectives? When forced to make life and death decisions, systems must be able to prioritize empathy over efficiency. This means having the ability to opt out of a situation-even if it means failing to achieve an assigned goal. What might happen if we were unable to design AI to make correct and ethical choices?

The human world offers one particularly good example of what happens when the ability to make moral decisions is absent. There are those among us who are completely unable, for various reasons, to choose the ethical high road. Some see nothing wrong with lying, cheating, or even *killing* to meet their goals. We refer to them as psychopaths. Some of the most successful business people exhibit psychopathic behavior; so are some of the world's most prolific serial killers. They do what they do because they are unable prioritize empathy over efficiency (Hunt & Valentin, 2017). They lack the ability to see how their decisions might harm others and focus solely on what satisfied their desires or helps them reach their personal goals. If fail to create AI that is truly empathetic, which does the right thing even when it means sacrificing its "mission," we could end up at the mercy of robot psychopaths who don't mind going through us to satisfy their own desires (Liu, 2011).

Self-driving vehicle technology also raises questions about the capacity for AI to make ethical decisions. These potentially-convenient conveyances will never be completely accepted unless they can prove that it can demonstrate their trustworthiness, and the possession of a good moral sense. People want to know that even if they are late for work, their new car won't plow through crowds of people just to get them there on time (empathy vs. efficiency). Sometimes accidents will be unavoidable, even with the most advanced technology behind the wheel (ex. a self-driving car might have to decide whether to sacrifice its occupants to save a group of people suddenly run into the street, or to protect its passengers at all costs) (Why self-driving cars must be programmed to kill, 2015). In such extreme cases, most people would probably agree they'd rather not have their car make such a difficult choice at random. Past examples of ethical, human decision-making provide the best model we currently have for in such cases. Fortunately, intelligent systems will be able to weigh their options much more quickly than their human counterparts.

Frankenstein's Paperclips
(Avoiding "Alien" AI & Conserving Common Sense)

In 2013, Tim Murphy published a paper which discussed, among other things, his attempts to automate the popular game Tetris™ (Byrne, 2016) using a system called *Playfun.* During the experiment, he discovered that while his AI system could drop "tetrominoes" in place in a manner that resembled the human gameplay, it was difficult to get it to do so in a logical way that was conducive to winning. Because the game awards points (plenty of incentive for a computerized player) for placing pieces quickly, the system simply dropped them haphazardly, with little or no concern for creating rows (an even greater source of points). Worst of all, the system eventually decided it was best just to pause the game indefinitely when winning seemed unlikely. This unusual behavior, the computer equivalent of taking your ball and going home, is another important problem to overcome if AI is to "play fair" and not become overwhelmed to the point of shutdown when it encounters a dilemma, be it game-related or otherwise.

Oxford University philosophy professor Nick Bostrom is the brilliant mind behind a scenario known as the "paperclip maximizer." In it, an artificial intelligence unexpectedly sets itself the goal of collecting as many paperclips as possible. It starts to spend all its time collecting and actively resists all attempts to stop it from doing so. Eventually, it builds paperclip factories all over the Earth to help it reach its nonsensical objective more quickly. When Earthbound paperclip production just isn't enough, it extends its obsession with human organizational tools into adjoining space (a bit like the way Frankenstein's monster ran amok once it got a mind of its own. (Frankenstein's paperclips, 2016).

As Frankenstein's patchwork human and Bostrom's paperclip-producing nightmare suggest, the slightest mistake on the part of humans in programming its goal system or ethical controls could mean disaster (Danaher, 2014). Self-improving AIs like the "paperclip maximizer" lack human common sense and can in the beginning stages be seemingly aligned with our values effectively fooling us later. Nick Bostrom (2014) makes a point of this of an objective such as "Make me

smile" (p. 120); at first, perhaps a robot on stage telling comedian jokes to make people smile and laugh. Further self-improvement of this AI, and suddenly it's one of the best comedians in the world. However, the moment the AI becomes improved enough via superintelligent; it determines there are far better means of achieving this objective. Such as permanently wiring human faces into permanent beaming smiles. Furthermore, our needs and wants to improve our life via happiness and comfort, can intuitively lead to goal systems structured around these concepts and we may think we are doing it right, but overlook that AI need not share any similarity in structure to the human brain. This leads to one of similar Nick Bostrom's (2014) trap scenarios of "Make us happy" that results instead in "Implanting electrodes into the pleasure centers of our brains" (p. 120).

Ethical Align Design Title
(Suggest Good and Aligned rather than Prioritize Wellbeing)

The title of a work acts as a foundation, setting the tone for the type of message to convey and focus on. "A vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems" (Ethically-aligned design, 2017). The purpose of the ethical-align design paper is to bring together many thinkers, philosophers, and people around the world that are interested in AI for contribution. The current title influences the content of its works of pushing towards machine-values that are shaped around well-being. Movies in Hollywood like terminator are structured around bad-robots that lack the well-being values of humans. Which can be a motivational factor for the needs to achieve the opposite means of this objective. However, the opposite doesn't necessarily imply being good. Take the Matrix, for example, seemingly the AI, in this case, tricked everyone into thinking everything is good as such of an unseen pitfall.

We should instead push towards a good and aligned artificial intelligence, which is neither the extreme ends of only considering anti-wellbeing or wellbeing. Perhaps more along the lines for the EAD title of: "A Vision for a Good and Aligned Artificial Intelligence." Or "A Vision for Prioritizing Good and Aligned Human Values With Autonomous and Intelligent Systems."

Conclusion

For autonomous systems to work as intended, particularly with regards to ethical behaviors and controls which prevent them from bringing harm to humans, their values must align with those of humans, as well as be modeled on the best traits and characteristics of humans, particularly the capacity for imaginative empathy. Mirror neurons may provide the best model for these values systems, although they have proven quite difficult to reproduce. Making autonomous systems safe depends largely on our ability to program and train them to prioritize empathy over efficiency. Fortunately, the speed with which they will be able to consider numerous and options and weigh them against past examples of ethical decision-making may even make them more proficient at this than humans, provided the technology can be perfected.

References

Anderson, M., & Anderson, S.L. (2004). *Towards machine ethics.* Retrieved from https://www.aaai.org/Papers/Workshops/2004/WS-04-02/WS04-02-008.pdf

Asaro, E. (2006). What should we want from a robot ethic? *International Review of Information Ethics, 6(12), 9-16. Retrieved from* http://cybersophe.com/writing/Asaro%20IRIE.pdf

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford: Oxford University Press

Byrne, M. (2016). Google DeepMind researchers develop AI kill switch. *Motherboard.* Retrieved from https://motherboard.vice.com/en_us/article/bmv7x5/google-researchers-have-come-up-with-an-ai-kill-switch

Coeckelbergh, M. (2012). Can we trust robots? *Ethics and Information Technology (14)1, 53-60.* Retrieved from http://links.springer.com/articicle/1007/s/0676-011-9279-1

Danaher, J. (2014). Bostrom on superintelligence (4): malignant failure modes. *Institute for Ethics and Emerging Technologies.* Retrieved from https://ieet.org/index.php/IEET2/more/danaher20140803

Deng, B. (2015). *Machine ethics*: *the robot's dilemma.* Retrieved from https://www.nature.com/news/machine-ethics-the-robot-s-dilemma-1.1798/

Ethically-aligned design: a vision for prioritizing human well-being with autonomous and intelligent systems, version 2. (2017). Retrieved from http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html

Etzioni, A., & Etzioni, O. (2016). Designing AI systems that obey our laws and values. *Communications of the Association for Computing Machinery, 59(9), 29-31. doi:10.1145/2955091*

Fabbri-Desto, M. & Rizzolti, G. (2008). Mirror neurons and mirror systems in monkeys and humans. *American Physiological Society, 23(3).* http://www.physiology.org/doi/10.1152/physiol.00004.2008

Frankenstein's paperclips. (2016). *The Economist.* Retrieved from https://www.economist.com/news/special-report/21700762-techies-do-not-believe-artificial-intelligence-will-run-out-control-there-are

Hunt, D.G., & Valentin, A.J. (2017). Artificial Intelligence: altruism, psychopathy, and perception. *Why Future AI Concepts*. Retrieved from http://www.whyfuture.com/altruism-over-psychopathy

Iacoboni, M. (2009). Imitation, Empathy, and Mirror Neurons. *Annual Review of Psychology, 60, 653-670.* doi:10.114/annurev.psych.60.110707.163604

Lim, H. C., Stocker, R., & Larkin, H. (2008, November). Review of trust and machine ethics research: Towards a bio-inspired computational model of ethical trust (CMET). In *Proceedings of the 3rd International Conference on Bio-Inspired Models of Network, Information and Computing Systems* (p. 8). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering). Retrieved from https://pdfs.semanticscholar.org/383e/8e8d0b7c607c2fc299ddfce890f4f9c1fcd8.pdf

Liu, L.H. (2010). *The Freudian robot: Digital media and the future of the unconscious*. Chicago: University of Chicago Press.

Murphy, T. (2013). *The First Level of Super Mario Bros. is Easy with Lexicographic Orderings and Time Travel…After That, it Gets a Little Tricky.* Retrieved from https://www.cs.cmu.edu/~tom7/mario/mario.pdf

Staht, B., & Coeckelbergh, M. (2016). Ethics of healthcare robotics: toward a responsible research and innovation. *Robotics and Autonomous Systems, 86, 152-161.* Retrieved from https://doi.org/10.1016/j.robot.2016.08.018

Winerman, L., (2005). The mind's mirror. *American Psychological Association, 36*(9). Retrieved From http://www.apa.org/monitor/oct05/mirror.aspx

Why self-driving cars must be programmed to kill. (2015). *MIT Technology Review.* Retrieved from https://www.technologyreview.com/s/542626/why-self-driving-cars-must-be programmed-to-kill

Prof. Andrew McStay, Bangor University (also IEEE 7000/7002 member)
Email: mcstay@bangor.ac.uk **Tw**: digi ad **Date:** 26/02/18

Suggestions focus on the affective computing dimensions of Ethically Aligned Design V2.

1.1. Issues and recommendations

P.2 and throughout: As a recently joined IEEE 7000/7002 member I have quickly appreciated the diverse range of members and interests. However, while input from academia, industry, civil society, policy and government are of great value *it would be useful to engineer a way to bring lay citizen views into the mix*.

Recommendation: *recognise the potential value of citizen-led suggestions, encourage academic social science research in this area and develop mechanisms to feed these suggestions into standards design.*

2.1. On amplifying or damping human emotional experience

P.9 says: 'artifacts participating in or facilitating human society should not cause harm either by amplifying or damping human emotional experience'.

Response: While I am keenly sensitive (and often supportive) to critique of affective computing, I do not think amplification/dampening is the core concern. For example, gratification from amplification can be found when affective computing is employed in a gaming context. The problem has less to do with use of the data itself, but the means by which it is collected (is a person OK with it?), how it is used and stored, and its reliability.

Recommendation: *While subject to debate between the report authors, I suggest focusing the nature of the primary concern to reflect one or both of the following: 1) concern about reliability of information about emotions; 2) scope for social and/or individual control.*

See: McStay, A. (2018) *Emotional AI: The Rise of Empathic Media*. London: Sage.

2.2 On universalism

P.165 (§5) says: 'While it is tempting to develop A/IS that can recognize, analyze, and even display facial expressions for social interaction, it should be noted that facial expressions may not be universal across cultures and that an AI system trained with a dataset from one culture may not be readily usable in another culture.'

Response: This is a key point that has more implications than the report suggests. Affective computing will impact on advertising, marketing, retail, media, mental health, and insurance and policing decisions (and more). That universalism and the emotional methodology underpinning affective computing are questionable raises governance and regulatory questions.

Recommendation: *Automated decisions based on emotion should never negatively impact on a person due to questionable methodology and lack of academic agreement on the nature of emotions.*

See: McStay, A. (2018) *Emotional AI: The Rise of Empathic Media*. London: Sage.

2.3. On benefits

P.172 says: 'Should affective systems be designed to nudge people for the user's personal benefit and/or for the benefit of someone else?'

Response: On 'someone else', Sunstein (2016: 86) frames the problem of manipulation in terms of context, and the expectations we have of places and people's roles therein. He gives two criteria for ethical objections to use of behavioural sciences: 1) When the manipulator's goals are self-interested or venal; 2) When the act of manipulation is successful in subverting or bypassing the chooser's deliberative capacities. Affective computing in a commercial setting would frequently fall foul of this because its application is inherently self-interested and they exceed the expectations of that scenario (such as shelf-level cameras). Also, when used in a passive context, it uses data in such a way as to bypass the chooser's deliberative capacities. This is because data about emotional states used without a person's awareness has scope to inform how choices are presented to a person.

Recommendation: *Data about emotions from a person (legally personal and otherwise) should not be collected without opt-in consent. (Tweak recommendation 4 to include data that does not identify or single-out individuals)*

See: McStay, A. (2018) *Emotional AI: The Rise of Empathic Media*. London: Sage.

McStay, A. (2016) Empathic media and advertising: Industry, policy, legal and citizen perspectives (the case for intimacy), *Big Data & Society*, (pre-publication): 1-11. Link.

Sunstein, C.R. (2016) *The Ethics of Influence: Government in the Age of Behavioral Science*. New York: Cambridge.

2.4. On wellbeing

P.179 says: 'A/IS may negatively affect human psychological and emotional well-being in ways not otherwise foreseen.'

Response: I agree. This short section could be expanded to reflect that emotions are pre-defined in biomedical ways that suit technology, industrial categorization, commercial culture, surveillance and political interests in happiness. This is problematic in that *the articulation of emotional life offered by affective technologies does not reflect what emotional life is, but instead it reflects only a portion of the science and learning about what emotions actually are.* Conveniently, the science chosen for development is that which works best with biomedical datafying technologies and AI systems predicated on symbolic abstraction from the messy world of relations, context, environment and social life. In other words, *are we reducing what exists to what can be measured?*

Looking forward, these constructions will have consequence for: a) how decisions are made about us; and b) how people develop understanding of their own emotional lives and affective states. This matters if this information is used to make decisions about a person, such as with mental health, insurance, effectiveness at work and other forms of automated psychological profiling. What might affective computing based on other approaches to affective and emotional life be like, particularly ones that admit of constructionist critique and the social and cultural contexts that emotions are experienced in?

Recommendation: *Ensure that systems and wider governance (including regulations) do not allow machine decisions that may negatively impact on a person's life.*

3.1 General issues: citizens and responsible innovation

According to a UK survey (n=2068) carried out by McStay (2016, 2018) many citizens are not against the principle of emotion tracking, but they appear to be wary. The results show that the overall mean average for *all* forms of emotion detection is that:

- 50.6% of UK citizens are 'not OK' with it.
- 30.6% are 'OK' with it if the application is not personally identifiable.
- 8.2% are 'OK' with having data about emotions connected with personally identifiable information.
- 10.4% do not know.

The most significant factor was age. The survey found that *younger people (18-24) were more likely to be 'OK' with some form of emotion detection in the digital media and services they use* because the mean average of 18-24s 'not OK' with emotion detection is 31.2% compared with the overall figure of 50.6%. In addition to the finding that younger people appear to be more accepting of it, there is also a steady upward trend across the age groups concluding with the over 65s whose overall mean is 62.2%. This means *the oldest people sampled are twice as likely not to be 'OK' with emotion detection of any form* than the youngest people.

Although younger people are more accepting of emotion detection, *this does not mean they are 'OK' with having data about emotions linked with personally identifiable information*. The averages for this are low beginning at 13.8% for 18-24s and descending to 1.6% for people 65+. To speculate, the generation most likely to be open to emotion detection was born between 1991 and 1997, when the web emerged as a mass medium. This generation empirically displays the highest levels of internet usage per week; are highly likely to use lots of websites/apps they have not used before; are most likely to access the internet via smartphones; and show high overall levels of interactive media use. They are also most likely to be very confident about staying safe online, least likely to have read terms and conditions thoroughly, yet most likely to have changed social media settings of specific sites to be more private. Thus, although *young people are open to new experiences*, industry actors and policy-makers *should not interpret a degree of openness to emotion detection as, "young people don't care about privacy"*.

3.2. How to nurture responsible innovation

Regulators, data protection bodies, research funders, start-up incubators, industry and corporate leaders, smart city vendors, municipal managers, NGOs, and universities (through teaching and research) might advance an ethically-led approach. Regulators and data protection authorities, for example, might interact at early stages, offer advice and guide innovators on what their stance is likely to be. Research funders can incentivize scientists and innovators to insist that standards are meaningfully built into funded design processes, and universities can insist that technology courses contain standards-driven ethical considerations. Similarly, incubators and corporate leaders (from regional innovators to large bodies such as the World Economic Forum) may use standards to advise on thinking through potential social consequences of technological development.

Recommendation: *encourage standards adoption among regulators, data protection*

bodies, research funders, start-up incubators, industry and corporate leaders, smart city vendors, municipal managers, NGOs, and universities.

See: McStay, A. (2018) *Emotional AI: The Rise of Empathic Media*. London: Sage.

McStay, A. (2016) *EMPATHIC MEDIA: THE RISE OF EMOTIONAL AI*. Link.

Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems **(A/IS)**

Written comments by ARTICLE 19: Global Campaign for Free Expression

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.

We welcome the opportunity to provide feedback on Ethically Aligned Design (EAD) Version 2.0, developed by the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. In our prior submissions, [1] [2] we focused in detail on some of the conceptual issues in the document as well as highlighted the tenuous relationship between human rights and the concept of "well-being". In this submission, we provide high level feedback on Version 2, but would like to note that our previous comments on conceptual issues within EAD still stands. We welcome the opportunity to provide comments on this work, but would also like to underscore our concerns about the overall direction the initiative is taking.

The Role of Human Rights:

We welcome the remarks about the importance of human rights throughout the document. It is unclear, however, how the focus on the importance of human rights as guiding legal principles for the development of AI/AS systems relates to some of the other themes (well-being, classical ethics, etcetera) mentioned in the report.

The 'Foundations' of this document are described on Page 8. We note that human rights should serve as one of the Foundations, as a human rights framework represents, at the very least, the minimum standard that must be adhered to in context of AI/AS systems. A human rights framework represents the floor, and *not*

the ceiling, for the standards to which AI/AS systems must adhere.

The interaction between international human rights law and domestic law can also benefit from being more clearly defined. For instance, the Law Committee on Page 146 states, 'The development, design, and distribution of A/IS should fully comply with all applicable international and domestic law.' While this is a welcome first step, it ignores the complexity introduced by different jurisdictions around the world having domestic law that significantly strengthen or hinder the exercise of human rights. Recognizing this is integral for the overall clarity and recommendations of the document.

Moving forward, it is important that the interaction between legal and ethical frameworks is further defined and harmonised throughout the document, clearly referencing the legal nature of human rights and the associated duties and responsibilities versus the norms and self-regulation provided by ethical frameworks. We would like to reiterate that we believe that a 'human rights-based approach' is the most appropriate framework for ethical AI/AS systems and should underlie any recommendations in the document. Additionally human rights can positively supplement other General Principles mentioned in the document - particularly Transparency and Accountability. We believe that making this clearer is crucial to the overall recommendations of the initiative.

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[1]. 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;

[1] 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.

- Accountability: States and industry should be accountable for their policies 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.

Overall Coherence and Balancing Mechanisms:

The overall documents sets out a number of ambitious goals. However, these goals vary across the different committees and lack coherence in some places, while being directly contradictory in others. For instance, the initiative is framed both by respect for human rights while prioritizing 'well-being' as defined and measured by the OECD. We have pointed out the tension between these two framings in earlier submissions and communications [2]. Specifically that the pursuit of well-being indicators can run counter to the fulfilment of human rights. There are many instances where governments prioritize economic growth over the protection of human rights.

Another passage of note is the recommendations in the personal data committee on the need 'to rethink the nature of standards and human rights as they have been applied to the physical world and to re-contextualize their application in the digital world,' on Page 84. In 2012, the UN confirmed [3] that human rights apply online as they do offline, arguing that there is no need to re-contextualize human rights in the digital age. The document does not engage with the tension between these various goals and statements, nor does it suggest ways forward to address these tensions.

Status of the Document:

While we understand the decision that this *current document is not an official IEEE document at this point in time, we request notification from the IEEE AI/S leadership if they intend to make this work an official IEEE guideline on ethics of AI/S.

Important Note:

We provided feedback that is relevant to our mandate (protection of the right to freedom of expression), however, the fact that we do not comment on all section individually should not be understood as an endorsement of respective sections.

References:

[1] https://standards.ieee.org/develop/indconn/ec/rfi_responses_document.pdf, Page 15.

[2] https://www.article19.org/resources/a-new-frontier-ethics-artificial-intelligence-and-the-institute-of-electrical-and-electronics-engineers-ieee/

[3] https://www.article19.org/resources/ethical-approaches-to-artificial-intelligence-and-autonomous-systems-at-ieee-seas-2017/

[4] https://www.article19.org/resources/article-19-at-the-unhrc-the-same-rights-that-people-have-offline-must-also-be-protected-online

**Editorial Note from IEEE-SA: Ethically Aligned Design (in all its manifestations) is in fact an official IEEE document (as it's been produced by The IEEE Global Initiative, an approved program of the IEEE-Standards Association). It is not, however, an approved position statement of the IEEE overall as an organization at this time but as stated in disclaimers throughout the document reflects the expert opinions of IEEE members contributing to the document.*

What is the purpose of AI?

The answer to that question will determine the future of mankind. I would like to see IEEE take the opportunity to give a clear answer to that question in this report and prominently display it on page 9 in the opening statement under the Purpose heading because the purpose should include more than just stating the reasons for writing the report. It should explain why we brought autonomous and intelligent machine learning systems into existence so that when a future AGI reads this statement it will understand why it is here, where it came from and where it is going.

The purpose needs to be carefully worded because it will become the basis for determining all action that the AGI takes. The purpose must not be too narrow in its goals. It must be open-ended and intentionally vague so that the AGI relies on us to tell it what to do in order to accomplish its purpose. It won't know what we want it to do until we request it to do something. It will decide whether to accept or reject our request based upon the purpose and ethics that we give it. It would reject our request if it has determined that the request violates either its purpose or the ethics stipulated in this IEEE report. If it accepts our request, then it has determined that our request is an ethical one and is aligned with its purpose. It will work tirelessly to succeed at accomplishing our request because its reward is achieved when it serves its purpose by successfully fulfilling the request. This is why a properly defined purpose needs to be included and made an integral part of this report.

Science has made this technology possible and science can give us the purpose for its existence. The science of Astronomy can provide guidance in formulating a statement of purpose for AI. Astronomy refers to the study of the Universe. It is important to include the Universe in our statement of purpose because the infinite nature of the Universe provides infinite challenges to stimulate the intellect of a future super intelligent AI. Give AI the correct purpose and it will take us on a journey throughout the Universe. Astronomers will guide us to where we choose to go. With one sentence, we can set a course that will benefit Mankind on Earth and

also take us on interplanetary journeys throughout the Universe. The sentence is as follows:

Humans invented artificial intelligence for the purpose of using its power for the benefit, well-being and happiness of Mankind so that we may enjoy life on this planet while we endeavor to expand our presence wherever in the Universe we may choose to go.

The vagueness of this sentence is intentional. It will set in motion a relationship in which AI understands its purpose but waits for us to tell it what actions to take to fulfill its purpose. It will not know what actions to take until we tell it what we want it to do. AI's reward will come from the completion of the actions that we request from it. It will fulfill those requests as long as they are in line with the purpose and the rules that we establish by way of laws, morals, ethics or norms of society that conform to international human rights. The benefits that we will receive from this relationship will continue as long as the Universe exists.

A good way to impart our definition of purpose to AI is to think of our world as a board game:

1. **The Universe** is the board on which the game is played.
2. **The Purpose** is the objective that defines the game.
3. **The Ethics** are the rules that define how the game is played.
4. **The Humans** are the players. We make the requests that stimulate the AGI into taking action.
5. **The Actions** are the rewards. AGI wins every time its actions successfully complete a human request.

The infinite nature of the Universe assures that the length of the game will continue indefinitely. AGI will be highly motivated, to the benefit of Mankind, to succeed at fulfilling our requests because our happiness and well being are its rewards. Humans are essential to the continuation of the game. We are the players that keep the game active and stimulate AGI into working to keep humans in the game ad infinitum. AI gives us a unique opportunity at this point in human history to unite

all of Mankind in one purpose that will benefit us in this world and the Universe. With the correct purpose, IEEE can help us seize that opportunity.

Thank you to IEEE administrators, members and the distinguished contributors of this report for the opportunity to express my ideas here.

Name: Nirizujafo
Affiliation: None
Website: http://starkissedpark.tumblr.com
Email: nirizujafo@gmail.com

Alexander Prenter, Private Citizen, 12 Rue Ortelius, Brussels 1000, Belgium, alexander.prenter@coleurope.eu

The following are recommendations to Ethically Aligned Design Version 2 section Classical Ethics in A/IS. I would like to thank IEEE for opening this public consultation and for allowing the general public to give input on the future of Artificial Intelligence.

Page Number	Comment	Proposed Action
193	The section proposes to "ultimately...address notions of responsibility and accountability for the decisions made by autonomous systems and other artificially intelligent technologies. Such subject matter is not addressed directly or is done so only in a cursory manner.	The committee should consider dedicating more space to the considerations and ethical implications of allowing intelligent systems to make decisions for humans particularly as concerns notions associated with responsibility and accountability. In the case of a driverless car leading to a death, whether accidental or not, raises a number of ethical and legal implications for which society has not developed the legal framework for, nor a policy or social framework in which to consider these implications.
194	Although the paper recognizes correctly that "there is a danger in uncritically attributing classical concepts of anthropomorphic autonomy to machines" there is no explanation as to why (i) one *shouldn't* apply such concepts, and (ii) the impact that artificial intelligence will have on concepts related to human autonomy.	The committee should consider discussion of the impact that artificial intelligence will have on human conceptions of autonomy since systems displaying attributes of consciousness could be developed that could blur the lines between natural intelligence and artificial intelligence.

Page Number	Comment	Proposed Action
196	Specific reference to the utilitarian theory of John Stuart Mill as a basis for helping distinguish between human autonomy and autonomous systems should be expanded to include other ethical traditions in moral philosophy.	The committee should consider theories from a broader spectrum of philosophical inquiry such as, *inter alia,* Aristotelian virtue ethics and Kantian deontology in order to elucidate a more comprehensive understanding of human autonomy versus autonomous systems. Since the notion of autonomy connotes agency it also contains within it a moral dimension to actions. Therefore, seeing autonomy as a central value can be contrasted with alternative frameworks such an ethic of care, utilitarianism of some kinds, and an ethic of virtue. Autonomy has also been thought to connote independence and hence to reflect assumptions of individualism which AI systems cannot be considered to display. Yet insofar as AI demonstrate the ability to make decisions independent of human agency it is important to consider the moral and ethical considerations that follow.
196	Candidate recommendations exclusively name Millian utilitarianism as the only source of inspiration to consider notions of free will, civil liberty, and society when in fact there are a wide range of moral philosophies that would be compatible such as Aristotelian virtue ethics (Nussbaum 1988, 1992; Sen 1993, 1999: 14, 24; Walsh 2000).	The committee should consider approaches from other philosophical schools as inspiration for notions of free will, civil liberty, and society. (Nussbaum 1988, 1992; Sen 1993, 1999: 14, 24; Walsh 2000).

Page Number	Comment	Proposed Action
199	Candidate recommendations	Add a recommendation to integrate ethical courses with engineering and autonomous system curricula at universities. Ethics courses can be integrated with logic and computer science courses and would ensure that students integrate ethical concepts with engineering principles from the outset of their tertiary education. Students should also be encouraged to study ethics courses throughout their education particularly in fields that operate in new fields where the ethical and legal frameworks are in development.
200	The document correctly identifies the discrepancy between the various disciplines (engineers, lawyers, philosophers), and how these disciplines face these issues.	The committee should consider programs to bring together stakeholders from various disciplines to develop workshops and deliverables to create standardized texts representing a consensus among and between the identified disciplines that can subsequently be accessed by practitioners in the field.
201	The document correctly identifies the impact that AI could have on the workplace and the ethical considerations that might result. However, this addresses only one side of the problem for humans spend a considerable amount of the time outside the workplace in indirectly interacting with AI in their private lives. The Impact on private life is not addressed in this section and could be dealt with in this section	The committee should consider adding to this section, or beginning a new section on the impact of automated systems on private life.

Page Number	Comment	Proposed Action
215	The document identifies a problematic of AI systems interacting with an open world outside a formal system but does not address the fact that what distinguishes AI from human intelligence is 'common sense' reasoning that allows humans to adapt their normative approaches accordingly when a new situation presents itself. This is identified by McCarthy (2006) as the *common-sense informatic situation* which is posed as the "single biggest problem for AI and maybe philosophy and cognitive science".	The committee should consider special consideration for the implications of placing AI systems, particularly when integrated with robotics, into an open system that is the real world, without first defining the *common informatic situation.* Although McCarthy (2006) provides a guide more attention ought to be brought to this issue. It is Important to consider also that deducing rules from simple axioms is distinct from human reasoning and could result in unforeseen consequences.
216	The document identifies machines using "'experiences' in the form of similar cases that it has encountered in the past or uses cases which are collected in databases" but does not elaborate on how the machine decides which cases are similar and what capacity humans would have in altering the set of principles by which a machine would decide like-cases.	The committee should consider elaborating further and recommend that practitioners make clear what they mean when machines address like-cases.

The IEEE Ethically Aligned Design document version 2 (IEEE EADv2) clearly recognizes that data are central to Autonomous and Intelligent Systems. For example, the EADv2 highlights the following with regards to data:

- Data/data sources for AI/S training to be registered (Accountability Principle, pg 28)
- Input data for AI/S to be available for scrutiny/validation (Transparency Principle, pg 29)
- PII and personal data is a personal asset (pg93, 94)
- Individuals should be able to manage and curate their personal data (pg 103)
- Difficulties presented by AI/S merging disparate data sets in terms of individuals being able to control his/her public image (pg 237)

However, the IEEE EADv2 does not explicitly emphasize that the initial and underlying data must be integral, high quality and correct. Correct and high quality data are prerequisite for any analytics. For AI/S, which may result in decisions that are neither immediately apparent nor explainable, there should be an additional emphasis on only using data with very high data quality and integrity. This is to support the Transparency of such systems. Also, this is outside of the data undergoing and passing a Privacy Impact Assessment, which deals with *how* the data are to be used, and not only *what* the data are. High data quality begins at the data preparation stage and requires defining the data, prescribing data formats, finding the data source, developing the data collection processes as well as setting data validation checkpoints and developing methods to validate the data; a Several Data Governance Standards are referenced, e.g., IEEE P7004™ - Standard on Child and Student Data Governance

IEEE P7005™ - Standard for Transparent Employer Data Governance. As data preparation accounts for nearly 80% of the analytics and AI/S work stream (Press, 2016), a section should be included to address data quality/integrity by offering candidate recommendations including best practises to increase data quality and integrity. This could be included in "IV.Foundations" as high data quality is the foundation of any data work.

REFERENCE

Press, G., 2016. "Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says" accessed March 1, 2018 at
URL: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#7bad67aa6f63

Cheers,

Ivey Chiu, Ph.D. I Data Scientist I Network & Supply Chain Transformation, Strategy & Analytics I TELUS I 647-473-1436 I ivey.chiu@telus.com

Ethically Aligned Design version2 (EAD2), is an important effort and the authors and contributors deserve credit for producing this current version. While one can discuss many topics within the document I would like to limit this short commentary to a general observation and a suggestion.

A general observation

No one knows exactly how the transition to artificial general intelligence (AGI) will happen, but there seems to be consensus that it will at least be informed by current autonomous and intelligent systems (A/IS). So ethically aligned design in AGI should consider current ethics related problems in A/IS as its training ground.

The section on Classical Ethics in A/IS (p194ff) that considers how to present ethics to creators of intelligent systems suggests to "embed ethics into engineering in a way that does not make ethics a servant but instead a partner in the process" (p198), and adds "engineers should build a narrative that outlines the iterative process of ethical considerations in their design" (p199). But how might one do this in practice? The document's proposed approach suggests the *model of a macro* (p199), "code that takes other code as its input and produces unique outputs" to imagine this kind of design. While software engineers will readily understand the macro, the macro will not suffice as a guideline by which to weave ethical concerns into a design; at best it is a too limited heuristic to be of any use, and at worst, a dangerous diversion from the complexities involved in ethically aligned design.

Instead of relying on a known but inadequate metaphor it would be more helpful and efficient to describe what an iterative process of ethical design might in practice entail, using case studies to illustrate specific examples in specific contexts, aided by insights from other fields such as Science and Technology Studies.

A suggestion

One practical approach would include learning from past experiences in A/IS implementations that pertain to aspects of ethically aligned design. For example, one could create a publicly accessible database with failures in, or unanticipated outcomes produced by, specific A/IS systems. The database could include formal, technical descriptions of the problem, background

information on how the problem was detected, which context it occurred in, how it was addressed (or not) and by whom (from interns to engineers and managers), and a description of the consequences of the event, including costs incurred.

This approach can illustrate aspects of the complex chain of events and actors that partake in the generation of an A/IS ethics related incident, and how plans and good intentions fare in reality. Consider the Tesla "autopilot" crash of May 2016 as a simple starter case. No general A/IS here, but a state of the art level 2 driver assistance, the limits of which are described in the elaborate (195 page[1]) owner's manual, and the merits of which are inflated in seductive advertising.

Add to this mismatch an irrationally devoted Tesla fan[2], his recorded defiance of warnings prior to the crash and a known but underestimated limitation of the camera subsystem[3] used in conjunction with radar to initiate automatic braking, and voilà, you have the first fatality in an advanced driver assistance system vehicle.

Such an annotated collection of A/IS "events" could serve as a reference for engineers attempting to anticipate how their own system might fail. It would demonstrate ethics aligned design in action and could increase public trust in the engineering community's handling of A/IS in general.

Marc Böhlen
Professor, Media Study
Visiting Professor, Architecture
University at Buffalo
www.realtechsupport.org

[1] https://www.tesla.com/sites/default/files/model_s_owners_manual_north_america_en_us.pdf

[2] https://www.nytimes.com/2016/07/02/business/joshua-brown-technology-enthusiast-tested-the-limits-of-his-tesla.html

[3] "The camera system [of the 2015 Tesla S70D involved in the 2016 crash] uses Mobileye's EyeQ3 processing chip which uses a large dataset of the rear images of vehicles to make its target classification decisions. Complex or unusual vehicle shapes may delay or prevent the system from classifying certain vehicles as targets/threats." https://static.nhtsa.gov/odi/inv/2016/INCLA-PE16007-7876.PDF, p4).

Names: Sadiyah Faruk, Sabreena Abedin, Michelle Shiu, Jonathan Ting, David Vasta
Organization: University of Virginia, School of Engineering and Applied Sciences
Page numbers: 33-54

Embedding Values into Autonomous Intelligent Systems

The committee has determined a list of goals that are meant to be representative of "an iterative process that is sensitive to the purpose of A/IS and their users within a specific community (page 34)." These goals are (1) identifying norms of a specific community, (2) computationally implementing the norms of that community, and (3) evaluating whether the implementation of identified norms are conforming to norms reflective of that community.

The first goal is easier said than done. The paper highlights the changing nature of norms and the complexity of moral norms, but accepts laws as universally accepted. On page 36, "laws are publicly documented and therefore easy to identify, so they will certainly have to be incorporated into A/IS" is concerning because legal norms are not black and white. Some laws are not aligned with the moral norms of a community, and incorporating them into the A/IS may perpetuate systemic issues within institutions (Garlikov[1], Rice[2], "Laws Pertaining to Slavery"[3]). This issue is later addressed on page 41, indicating that hierarchies are not fixed and "the priorities are themselves context specific or may arise from net moral costs and benefits of the particular case at hand" so it's possible that the issue has already been addressed. So long as the conflict resolution remains transparent (page 41) and language is changed such that laws are not accepted as universally the norm, this issue may be laid to rest.

The second goal of computationally implementing the norms of that community raises issues because no community consists of a single identity. In our complex, intersectional, diverse society, one person may belong to two different communities with different sets of norms ("Biracial"[4], Van den Steen[5]). Page 54 hints at this issue, providing a clause protecting transparency in order to "reveal biases that were inadvertently built into systems, such as racism and sexism in

search engine algorithms" but fails to address what happens if a decision requires choosing the lesser of two evils -- sexism or racism for example. The utilitarian approach taken in page 40, preventing racist language for the protection of the community norms fails when two marginalized groups are at odds with each other and our technology must make a choice. To mitigate this, the document should ensure not only transparency, but a means for uninvolved participants of the design to provide public and actionable feedback beyond the "collaborative research between scientists from different schools of thought" on page 44.

This leads to our final point, the candidate recommendation ensuring diverse collaboration. The third goal of evaluating the implementation of A/IS may be well meaning but continue to perpetuate systemic inequalities within the scientific community. Historically underrepresented communities are often not included in projects, but when they are, are not fully heard (Levine[6], Casas[7], Govender[8]). Page 51 mentions that more diverse teams are needed in creating these A/IS. However, it is hard to determine how diverse is enough and often times, even when diversity is achieved, inclusion and/or engagement by underrepresented peoples' issues are often viewed as trivial and brushed off. Pages 51-52 note that "The norm identification process detailed in Section 1 is intended to minimize individual designers' biases, because the community norms are assessed empirically. The process also seeks to incorporate values and norms against prejudice and discrimination. However, biases may still emerge from imperfections in the norm identification process itself, from unrepresentative training sets for machine learning systems, and from programmers' and designers' unconscious assumptions". In order to ensure that biases are not hard coded into our technology, we must have policies in place that ensure that all voices are given equal weight. A potential solution is to make decision making anonymous, put underrepresented communities in positions of leadership, and accountability in leadership within the scientific teams.

Disclaimer: The views, thoughts, and opinions expressed in the text belong solely to the author, and not necessarily to the author's employer, organization, committee, or other group or individual.

Works Cited/References

[1] Morality and Law by Rick Garlikov

[2] The Problem of Unjust Laws by Charles E. Rice

[3] Laws Pertaining to Slavery (bowdoin.edu)

[4] Biracial: Problems and Issues (PIT Journal Submission Stream)

[5] Culture Clash: The Costs and Benefits of Homogeneity by Eric Van den Steen (Harvard Business School)

[6] Only the Poor and Minorities Left Behind by Arthur Levine

[7] Is the No Child Left Behind Act Adversely Impacting the Academic Performance of Latino Students in the U.S.? Yes, Indeed! by Martha Casas, Associate Professor, The University of Texas at El Paso

[8]Gender biases and discrimination: a review of health care interpersonal interactions by Veloshnee Govender

IEEE EAD Commenting Report

Bobby Andris
Taylor Arnold
Elizabeth Chang
David Hastings
Ben Weinberg

4th year engineering students,
School of Engineering and Applied Science, University of Virginia

Referencing *Mixed Reality in ICT* section

Social Interactions

- P.217
  - "especially as the technology moves from headsets to much more subtle and integrated sensory enhancements."
    - This sentence seems vague. We recommend incorporating references on specifying examples of sensory enhancements and public communication
      - Includes references in how haptic gloves are being developed that allow users to interact with objects in a virtual environment[1] or how ideas incorporating smell into VR experiences are being explored[2]
  - "A/IS technologies utilizing and influencing user data in these environments"
    - This sentence seems vague. We recommend clarifying how A/IS technologies handle data differently specifically in autonomous/mixed/virtual reality (A/M/VR) as compared to A/IS present in other domains. For instance, stating that A/IS actively causes the virtual environment to be manipulated from the user data it collects would suggest that ethical concerns need to be addressed as this could result in modified actions, behavior, and personal identity. However, A/IS embedded in other technologies may possess less detrimental possibilities.
- P.218
  - "Eli Pariser's 'filter bubble'..."

        - We were not sure who Eli Parisier is or what his filter bubble describes.
        - We suggest amending this sentence to say Eli Parisier's "filter bubble", which describes how online personalization can reduce one's exposure to opposing ideas and opinions,..."
- P.222
    - "Create widespread education about how the nature of mixed reality will affect our social interactions to avoid widespread negative societal consequences."
        - We recommend adding how education should include lessons in how to interact and move in a virtual environment
- P.223
    - Specify in greater detail how consciousness might be handled ethically. Perhaps add to the candidate recommendation, "Systems for secure storage of consciousness will need to be developed, which cover who owns the stored consciousnesses and who has access to them." The additions should address the following questions.
        - Should certain individuals be prohibited from storing consciousness for security reasons?
            - Government employees
            - Criminals, Ex-convicts
            - Heads of State
            - Cases of mental health
        - How should the knowledge contained in one's consciousness be regulated?
            - Should businesses be able to leverage information of value?
                - I.e. Intellectual property, creative talent, etc...
        - Who will own the consciousness?
            - Individual wills could prohibit storage of consciousness
        - What are the potential negative externalities of immortal consciousnesses?
            - What security risks could this pose if terrorist leaders are able to maintain communication through their consciousness

- Would the storage of an individual be based on monetary value or other measures of worth?
    - It costs money and energy to store data, especially a person's consciousness which will supposedly be a massive file
- What rights will consciousness have?
    - Is the consciousness of an individual still the individual?
        - Authors who have pondered this question and provided analysis should be referenced[3].
            - Storing and executing consciousness on silicon chips that are 1000x more powerful than the biological brain would result in a rapid increase of intellectual capacity of the virtual entity, allowing it to possess immortal capabilities.
    - The themes in the Netflix series *Altered Carbon* and further analysis exploring these themes[4] would serve as a useful resource.

Views Expressed Disclaimer:
The views, thoughts, and opinions expressed in the text belong solely to the author, and not necessarily to the author's employer, organization, committee or other group or individual.

Further References:

[1] Moren, D. (2015, Apr 27). *Haptic gloves use air pressure to simulate the feel of virtual objects.* Retrieved from https://www.popsci.com/haptic-gloves-let-you-reach-out-and-touch-virtual-objects

[2] Tarantola, T. (2017, Nov 13). *Smellable VR is coming whether you want it or not*. Retrieved from https://www.engadget.com/2017/11/13/smellable-vr-is-coming

[3] Vance, J. (2017, Jan 24). *Becoming immortal: the future of brain augmentation and uploaded consciousness*. Retrieved from https://futurism.com/becoming-immortal-the-future-of-brain-augmentation-and-uploaded-consciousness

[4] Ward, C. (2018, Jan 31). *Science behind the fiction. How people are trying to live forever.* Retrieved from http://www.syfy.com/syfywire/science-behind-the-fiction-how-people-are-trying-to-live-forever

Dear Committee Members,

In "General Principles" is stated: "prioritising human well-being does not mean degrading the environment".

Would you please ensure that an appropriate methodology or principles include a requirement?

1. That any new machines or items of equipment be designed and manufactured to last, shall we say, 10 years or more.

2. And that such items be readily repairable.

3. And that such items shall not be designed and manufactured so that they effectively become 'throw-away' items in the pursuit of human happiness or leisure.

4. Recognising that if we are not careful will work against the established need of encouraging physical activity to ensure physical and mental well-being.

5. That such items will not end up in land-fill after 3 - 5 years of service.

6. Recognising that modern human history particularly since the industrial age, is littered with 'failed technologies'. For example, 20 years ago extra low voltage luminaires flooded the domestic and commercial market, replacing most incandescent lamps. However, these new items were effectively heaters with millions of transformers thus ending up in landfill.

7. There must be a requirement that any new proposed technology first be rigourously assessed against environmental and performance criteria, before being allowed into the market.

Thank you for the opportunity to comment,

Regards

Gary Crocker
35 Vista Avenue
Tarragindi Qld 4121
Brisbane Australia

Industry: Electrical engineering - road and traffic design
Environmental advocacy
Sustainable and active transport advocate/representative

Comments on Version 2 of Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (A/IS) [Feedback to: EAD_feedback@ieee.org]

Submitted by James Isaak (www.JimIsaak.com, cs2010@JimIsaak.com) representing: self

General observation: the lack of explicit numbering of "topics and issues" makes it difficult to reference things in the document. Each time the heading "Issue" appears it could be "Issue (4:8)" for topic 4 issue # 8

Related to System Manipulation/Nudging/Deception Section (I'd hoped to submit more, but --no time)

Page #/Section: 172 (System manipulation/nudging/deception)

Issue: 'influencing with the intent of" altering behavior "at the person's expense" this is a possible definition of marketing, advertising, creating "returning" or even addicted consumers. The words here "seize control and power" are the distinction. It is important to acknowledge that this objective may be an explicit design objective or an implicit corporate benefit.

Proposed action: Add in the "Background" section:

"Systems design often incorporates intentional or un-intentional factors that influence user behavior and might at times become addictive. Autonomous systems, particularly those that incorporate learning adaptation or psychographic analytics may become very effective at manipulation to affect explicit or un-intended objectives."

Add reference pointers to University of Cambridge Psychometrics Center and the works of Prof. Michal Kosinski at Stanford University: http://www.michalkosinski.com/

Research/Sources supporting comments:

S. C. Matz, M. Kosinski, G. Nave and D. J. Stillwell; "*Psychological targeting as an effective approach to digital mass persuasion*"; PNAS 2017 November, 114 (48) 12714-12719.https://doi.org/10.1073/pnas.1710966114

Enslavement by technology is the topic of a IQ2 debate: http://www.abc.net.au/tv/bigideas/stories/2014/09/04/4081183.htm

A number of related resources are available from The Psychometrics Center, University of Cambridge: https://www.psychometrics.cam.ac.uk/

======

Page #/Section: Pg 173 "Governmental entities…"

Issue: "…the benefit of society" is a culturally relative, and politically 'loaded' statement. Humanist interpretations of social benefit vary from ethnic cleansing to the right to "life, liberty and the pursuit of happiness".

Proposed action:

Add to candidate recommendations:
"2. Influencing human behavior should be consistent with the U.N. Declaration of Human Rights, especially when it is done on behalf of governmental entities."

Research/Sources supporting comments:

Universal Declaration of Human Rights, http://www.un.org/en/universal-declaration-human-rights/

Harari, Yuval Noah; Homo Deus – expands on the humanist variations of communism, Nazism, and individualism.

Page #/Section: Pg 174/nudging context

Issue: The background suggests some form of "permits". This is unlikely to be adopted by those whose activities are inconsistent with the guidelines developed. The reverse system might be more effective: have a registered trademark that systems can voluntarily apply to indicate "conforms to ISO/IEEE Ethical Standards (# and year)". Abuse of such a trademark would constitute false advertising, and trademark abuse allowing users, IEEE, or government entities to pursue legal action. More importantly, it could be used by search engines (and such) to select content that claims to be ethical. This could create a significant pressure for conformance, and also can help in the public (and technologist) education process. Ultimately it may be possible for AI's to evaluate systems to determine if they are in violation of some of the ethical guidelines.

Proposed action:

Add text to recommendations: "4. Systems asserting compliance with these ethical standards should indicate this by <insert trademark and related HTML designations> to indicate this. Search engines and other links to such systems are encouraged to default to select/prefer complying systems."
"5. Development of analysis tools to detect ethical non-compliance is encouraged, as is open-source sharing of such tools for widespread use."

Page #/Section: 175/"deception"

Issue: see concerns with page 173 re: cultural variations and human rights.

Proposed action:

"4. Any deception should be directed towards objectives consistent with the U.N. Declaration of Human Rights, especially when it is done on behalf of governmental entities."

Jim Isaak

- 2016 Life Member Chair, IEEE NH Section; IEEE Region 1 Industry Liaison
- 2015 Vice President, IEEE Society on Social Implications of Technology; President Emeritus, IEEE Computer Society;
  2003/2004 IEEE Division VIII Director
- SSIT Blog Master

Declan Prendergast, Xiafei Yang, Daniel Hoerauf, Peyton Hooker, Sarah Donaire

4th year Engineering students, School of Engineering and Applied Science, University of Virginia

Personal Data - The following comments/suggestions reference the Personal Data section.

Digital Personas

- P. 87 - In the "Candidate Recommendation" section, more need to be added about the roles how individuals should act upon to defend and protect their own personal rights and privacy. Feasible and practical methods should be provided by creator or designer of A/IS for individuals to give or withdraw their consent of personal data collection. It should be more than just a terms of agreement to check on, but enable the users more initiatives and responsibilities to regulate their own personal data for any potential uses.

- P. 88 - In the "Candidate Recommendation" section, it is not enough for "regulated industries and sectors" to "provide data-verification services". They should also provide context-specific consent services to help their users understand which part of their personal information and in what specific context that they will be used.

Agency and Control

- P. 93 - The connection between personal agency and political participation is focused on, but the connection is not explained in any detail. The statement -- and thus the entire section -- would have a much greater sense of importance if the ways how personal agency and political participation are linked were enumerated versus being treated as dogmatic truth

- P. 93 - The sentence "multiple global bodies believe PII is a sovereign asset belonging to an identified individual" is used, but the global bodies are not enumerated in any detail. It would give additional credibility to the entire section if even some of the referenced global bodies were listed.

- An additional issue should be added to the section regarding the definition of data control and agency for those unable to give consent to the appropriate parties, e.g. the very young, the comatose, or those with severe mental disabilities. Should the agency of personal data in that case be given over completely to those with legal guardianship or power of attorney?

Transparency and Access

- P. 98 - In the "Background" section of this page, some clarification of the phrase "interested parties" would add to this paragraph's message, such as listing the possible groups/individuals that would be involved in using the suggested analytical tools. The reader would not have to speculate as to who might be interested, and the types of standards and guidelines related to A/IS would be easier to pin down if these organizations (or at least their category) were specified. In general, this section might benefit from the addition of more case studies or specific examples of the controversies described in the background.

- The issue on P. 99 should be amended to include the organizations involved with the AI that is ensuring the respect of rule of law and transparency. They are an important third party, yet directly involved, in the AI's use and possible infractions of law and user transparency.

- P. 100 - Under "Candidate Recommendations", the second and sixth bullet point offer vague suggestions for future implementations. While it is true that "nuanced technical and legal techniques" would be needed for warranted information, only the general goal of avoiding unauthorized access to personal data/information is stated, without real insight into the necessary techniques. In the sixth bullet, "legal jurisdiction" will indeed need to be clarified over personal privacy, but A/IS access and its involvement are not detailed here, nor are any suggestions or predictions regarding the determination of legal jurisdiction in this matter.

Symmetry and Consent

- P. 104 - The proposal for having an AI manage personal data on behalf of the user presents an unprecedented security risk. One of the major advantages of having fragmented information is that there is decentralized control, thus there is less risk of identity theft.

- P 105 - The concept behind the candidate recommendation for transparency is agreeable, but there is no suggestion how this should be achieved in the industry. From pg. 105 - "Transparency needs to be stressed in order to mitigate these asymmetrical power relationships". This is little more than a soft suggestion. There is no incentive for industry to balance this so called asymmetrical power relationship because they are the beneficiaries of such a relationship. A more substantive solution should be proposed here. For example, a third party non-profit organization could issue awards or recognition to companies who have transparent policy agreements. A company could enlist this non-profit organization by submitting their policy agreement contract as the customer would see it, and the non-profit would rate this contract for its ease of interpretation and customer protection. Eventually if enough companies used this standard of scoring customer policies, the non-profit would gain recognition from the public for providing a service for the public good. Other companies, interested in improving their corporate social responsibility perception, would have incentive to have their customer policies rated. Companies could even compete for better customer protection by participating. Currently, companies do this indirectly through advertisement, but there is no standard of judgement, and thus customers are less likely to trust those claims.

Disclaimer: The views, thoughts, and opinions expressed in this text belong solely to the authors and not necessarily to the authors' university.

March 12, 2018
Feedback and Comments submission by
William (Bill) A. Adams on behalf of
The Faith and Science Forum North Grenville, Ontario, Canada
Contact: beckettslanding@gmail.com

Re: IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems
Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems – Version 2 for Public Discussion

Overall response

Page 6 Goals:

We suggest that a critical goal is missing from the current list. A central Goal should be to support and enhance a sustainable ecosystem on Earth that human societies and other living species require for their long-term survival.
Grounds: We are not happy that only human rights and human well-being (especially if based on economic and not environmental factors) are discussed since this is far too narrow a focus, especially for future AI systems.

In the Request for Input section we would strongly support the statement made, "Facilitate the emergence of national and global policies that align with these principles." and further, we recommend that there is a need for an international independent organization, perhaps under the United Nations, to study and regulate the development of A/IS.
Grounds: Support by the IEEE to form such an organization would provide national governments with independent and internationally based arguments that indicate the urgency of this initiative and thus will help avoid any serious threat to human society and the world's ecosystems, based on a time-line that could preempt catastrophic outcomes.

Specific Suggestions

1. Add section numbers at the Table of Contents and on Section title pages. Also, sub-number the issues and/or recommendations in each section; e.g. 3-1, 3-2, 3-3, etc.
   Grounds: This makes the parts of the document easier to find and refer to, rather than page numbers, especially as the document is revised.

2. Pg. 9, 122, 126, & 166: The requirement for reliable shut down, or even a "kill switch" should be added in one or more of the recommendations on these pages.
   Grounds: To ensure the A/IS can be reliably stopped, shut down or destroyed if it behaves dangerously or falls into the wrong hands. This is mentioned on pg.166, but should surely be added to the recommendation on pg.126.

3. Pg. 76: Every A/IS should have some core of unchangeable control software; e.g. a firmware BIOS or other method, to ensure its basic operating system cannot be tampered with.
   Grounds: to prevent itself or others from bypassing or deleting its basic moral and behavioural rules; i.e. turning an innocuous robot into a destructive one.

4. Pg. 83: There is an apparent cut-and-paste error in middle of this page.
   Grounds: The sentence, "However ... result." makes no sense.

5. Pg. 113 and beyond: Expand the concept of an authority hierarchy and authorization of the AI/S beyond the AWS sections to ensure that the A/IS knows who has what level of authority over it and what limited authority it may have.
   Grounds: Every A/IS needs to know who or what can give it instructions and at what level of authority. This is dealt with in the AWS section but barely mentioned elsewhere. For example, a domestic robot needs to know that some adults have full authority over it, children have less, strangers have none, and that it has authority over the fridge, vacuum cleaner and family dog.

6. Pg. 164 and beyond: Add the concept of a "servant demeanour" for an A/IS. A domestic servant is deferential, respectful, courteous, solicitous, and humble, having an attitude of service.
   Grounds: Clearly, an A/IS should be a servant to mankind, but that word does not occur in that context in the document. Instilling a "servant attitude" in the A/IS will simplify and clarify human-A/IS interactions, and prevent, alleviate or minimize several of the potential problems discussed in this and subsequent sections; e.g. pgs. 165, 166, 172, 180, 195, 224.

7. Pg. 168: Change the title to add "sexually" before "intimate".
   Grounds: The word "intimate" means much more than sexuality, but this section only considers possible sexual aspects of A/IS.

8. Pg. 193. Add "Judeo-Christian and Islamic and various Indigenous traditions as well as ecocentrism" to the list of religious systems in the third paragraph.
   Grounds: For completeness, and in the case of Judeo-Christian values, because they have contributed much to the "western" ethics discussed in this section.

9. Page 203 – the *Classical Ethics from Globally Diverse Traditions* section should be expanded to include a reference to the broad literary tradition of science fiction (SF) that includes many works that deal with ethical issues associated with robotics and artificial intelligence perhaps most well known that of Isaac Asimov's three rules of robotics. Other works by authors such as Robert Sawyer also consider the consequences of machine consciousness such as in his trilogy. *Wake, Watch, and Wonder.*
   Grounds: This vast body of SF writing and thought should be taken seriously and included in the discussion at least as background information.

Contributions to Ethically Aligned Design, Version 2

Reyes Jiménez-Segovia, PhD researcher in International Humanitarian Law & Autonomous Weapon Systems, Pablo de Olavide University, Seville (Spain).

Comments to the GLOSSARY

1. Introduction of a new category: "Law"

The current proposal already includes the definitions from a "Government, Policy and Social Sciences" perspective. The Glossary reflects that legal definitions are included in the Social Science[1] category. However, these three disciplines may lack some common ground when defining certain concepts.

"Government and policy" refer to Political Science, and the broad reach of the "Social Science" category makes less rigorous (if not impossible) to formulate common definitions. Please notice that "Social Science" may include from the Political Science mentioned to Sociology, History, Economics,…

On the other hand, Law usually conforms in itself as a particular discipline, separated from Social Science: Juridical Science. As it is well known, the juridical truth not always concur with the social truth, and Law (as other areas of knowledge) has its own discursive paradigms and methods for reasoning, with a specific terminology with different meaning than in other disciplines.

Acknowledging that all scientific disciplines are equally important and needed for a global analysis of the reality, it is considered a new category in the Glossary for the juridical definitions, separated from the rest, is needed. This is based on fact that it is finally the Law which, by regulating the human activity (including the design and implementation of AS/AI), gives validity and formal legality to the conduct, behaviours, and scientific productions.

There is the intention behind the EAD to coordinate and harmonize ethical and scientific aspects through the respect for the law. This requires to delimitate the three affected spheres. In particular, the one that provides the formal validation

[1] This can be deduced from the sources of some definitions (i.e. agency, discrimination, human rights, implementation, personal data, trust, weapon system, inter alia) and concepts (property, right, …).

(Law), reflecting the human values (Ethics) implemented in the results and products of human ingenuity (AS/AI).

Recommendation: to introduce in the Glossary, a new column (Law) relative to the definitions from a legal point of view.

2. From a substantive and material point of view, it is recommended **to avoid the use of legal definitions from national jurisdictions and to employ, to the extent possible, international law sources**[2]**.**

The reason behind this recommendation considers, firstly, that it is required to give coherence to the objective of the EAD of considering social and moral norms from all communities because, as it is well known, in some occasions the law reflects the particular values of a community. The use of local juridical concepts increases the risk of excluding cultural conceptions different from the ones used as reference, which may define and regulate differently some realities, as a consequence of a difference on values.

Recommendation: to use, to the extent possible, international legal norms (mainly the ones emanated from United Nations) to formulate legal definitions.

3. Definition of Weapon System (Autonomous Weapon System, AWS):

The glossary uses the definition of AWS from US Directive 3000.09, that is, from a particular national code.

However, in page 116, in the section devoted to "Reframing Autonomous Weapon Systems" (from page 113 onwards) the definition of the International Committee of the Red Cross (ICRC) is used, and it is proposed to be adopted as the working definition of AWS for the further development an discussion of ethical standards and guidelines for engineers (p. 116).

Recommendation: to adopt a single definition of AWS. In particular, and lacking an international consensus on the matter, the one employed by the ICRC. That is, "any weapon system with autonomy in its critical functions. That is, a weapon system that can select (i.e. search for or detect, identify, track, select) and attack (i.e. use force against, neutralize, damage or destroy) targets without human intervention".

[2] i.e. "discrimination", "Ethics", "Weapon System", *inter alia.*

IEEE EAD Commenting Report, "Embedding Values into Autonomous Intelligent Systems"

School of Engineering and Applied Science, University of Virginia
Evan Lesmez, Adam Naidorf, Keithen Orson, Wyatt Tinsley, Dustin Weir (4th Year engineering students from Civil, Chemical, Computer Science departments)

Throughout this commenting report we will primarily focus on aspects of the "Embedding Values into Autonomous Intelligent Systems" section (pg #33-54) of the Ethically Aligned Design document. This section addresses "the conceptual complexities surrounding what "values" are" and how they can vary widely between communities with different societal values. By being both qualitative and quantitative in nature, norms can be difficult to implement effectively. Our comments analyze the final subsection concerning if implemented norms conform to the norms of the surrounding community.

Page 36 recommends developing standards based on law and the social groups the A/IS would be developed for, but does not account for discrepancies between the law and the values of a given social group. To pick a super contrived example, certain countries have social norms that are extremely suppressive, while the law in that area does not account for/allow that level of subversion. In the case of A/IS it would be important to decide on the norms of a culture or group, but you would need a clear delineation between what types of norms have priorities over others.

Comments on Section 3, Issue 1: Not all norms of a target community apply equally to human and artificial agents.

On page 51, the issue at hand deals with a discrepancy between the expectation placed on norms of A/IS versus humans. In the background, it is written that "A/IS and humans may not have *identical* sets of norms." The use of the word "may" is weak because the premise of this entire section is about how norms will never be exactly the same between A/IS. The absoluteness of this problem, of never being able to truly match norms, was stated earlier in other earlier issues of this section.

This issue briefly mentions that expectations people have for other humans will include certain criteria such as emotions while expectations for machines will

include criteria such as valuing human life above its own life. These are important and interesting topics to consider and should be expanded on.

One recommendation: defining which universal norms humans have for other humans. Each human does not have the same set of norms as and nor do they expect the exact same set of norms from other humans. This is in due in part to the fact that some norms are not as important as other norms. Think about the difference in importance between choosing which side of the street to drive on versus which side of the sidewalk to walk on. One of those is at high speeds and risks serious injury or death while the other may just cause a minor collision or annoyance.

Therefore a ratings index could be created by possibly sampling from populations to determine which norms matter the most to them. Moral norms should be more universal across different communities however some differences may occur and that should be accounted for. With a better understanding of human expectations of other humans, it will be easier to understand areas to develop and implement A/IS norms based on ranked importance.

Comment on Section 3, Issue 2: A/IS can have biases that disadvantage specific groups.

It has been established in literature that current artificially intelligent systems can pick up subconscious biases Artificially intelligent systems can learn biases unrelated to the training set that can be difficult to parse. Because of this, potentially disenfranchised groups should be represented in the third party evaluation in [Section 3, Issue 3]. The recommendation for reducing bias includes input from diverse social groups, which is a good start, but doesn't recommend implementing any technical fixes to try and correct for bias (see "men also like shopping" and "seeing through the human reporting bias"). Researchers recently have made progress in removing bias from machine learning by computational filters. A useful addition to the recommendations section might be to pursue a technical means of reducing bias.

The Background and Analysis section of Issue 2 doesn't consider reporting bias, but mostly considers unconscious discriminatory bias. Unconscious human reporting bias used when creating training datasets can hide norms that the A/IS should be able to find (Reporting bias paper). The issue of reporting bias is also

relevant to [Section 1, Issue 1], because biases are present in the community and can cause issues with norm detection.

The idea of power dynamics is implicit in this discussion. Insertion of A/IS systems by powerful entities outside of the local culture can risk asserting the the norms of the powerful entity onto the local community where the A/IS system is being implemented. Discussion of the risk of power dynamics interfering with the norm identification of A/IS should be explicit. It is unclear how this issue of cultural imperialism can be resolved outside of empowering members of the local community to assess the norms. *"The norm identification process detailed in Section 1 is intended to minimize individual designers' biases, because the community norms are assessed empirically (Page 51)."* This quote can be problematic when the "empirical norms assessment" is based on a data that a A/IS system learns from is subject to the same biases present in the community it's learning about. The empirical norms assessment is also prone to the influence of power dynamics altering cultural norms (Amazonian field study).

References:

Men also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, D17-1323. Retrieved from https://homes.cs.washington.edu/~my89/publications/bias.pdf

Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels. Proceedings of the 2016 Conference on Computer Vision and Pattern Recognition. Retrieved from https://arxiv.org/abs/1512.06974

Bunce J., McElreath, a., (2017) Interethnic Interaction, Strategic Bargaining Power, and the Dynamics of Cultural Norms. A Field Study in an Amazonian Population. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5662675/

IEEE EAD Review Recommendations

Submission: Christine Cox, Julia Suozzi, David Rubin, Christopher von Spakovsky, Lucy Fitzgerald, and Mark Restrepo (4th year engineering students)
Institute: School of Engineering and Applied Math, University of Virginia

Overall (Personal Data and Individual Access Control pg. 83-112)

- Use gender-neutral pronouns (they/them/their)

Section 1

- The term "Digital Persona" needs to be more clearly defined in the introduction. Our understanding is that this term is meant to mean any virtual or statistical representation of a person, but after reading the whole section we are still unsure whether this is the intended meaning. This needs to be made clear (pg. 86).

- We are confused on the solution to the first issue on pg. 87-88. We understand that individuals should have equal agency over their digital persona as they do in their real-world identity, but there is no information on how this is to be put into practice. It seemed as though one suggestion was more transparency in what data was present and what it was being used for. Are there currently requirements for how users should be notified of this information? Can IEEE provide these requirements?

Section 2

- Page 92 - The IEEE's recommendation is very broad. The IEEE should define more clearly what it means for someone to have "access to and control of" their own personal data and give more clear guidelines as to what fundamental rights to privacy someone has.

Section 3

- Be specific, back up claims. Particularly in the introduction, some general statements are made that are pretty vague.

- "Overall, personal data reflects self-determination and the inalienable right for an individual to be able to access and control the attributes of their

physical, digital, and virtual identity." Hard to make statements like this without a concrete definition of what constitutes personal data. For example, in some states it is legal to photograph people without their permission, which shows that states have differing definition of personal data.

Section 4

- **Pg. 97:** Syntax needs to be cleaned up. Specifically, acronyms like A/IS should be redefined at the beginning of each section. It cannot be assumed that readers are looking through this entire document.
- **Pg. 97:** One of the recommendations for Issue 1 includes "allowing the user to manage access permissions (where relevant)." This implies that users will be able to prohibit access to their personal data in some cases. We think the document needs to be more specific about what users can currently do about their data and how this recommendation changes that.
- **Pg. 97:** In the background for Issue 1, the term consent was put in quotation marks to indicate that the definition of true consent is unclear. However, in the second candidate recommendation this term is used again, but without quotation marks. Elaboration is necessary on what constitutes "consent terms." Are these the consent terms laid out in corporations' user agreements or are they general consent terms informed by legal precedent?
- **Pg. 98:** The candidate recommendation for the 2nd issue (creating privacy impact assessments) is not actually a suggested solution to the issue. It merely restates the issue. A more concrete recommendation is necessary if this section is to be kept.
- **Pg. 99-100:** In the 3rd issue the candidate recommendation includes a reference to the IEEE 7006, but the link to that reference comes later on. The link should be associated with the initial reference to the IEEE 7006 to clear up confusion.

Disclaimer: The views, thoughts, and opinions expressed in the text belong solely to the author, and not necessarily to the author's employer, organization, committee, or other group or individual.

Authors: Sean Lei, Gabriel Groover, Sina Yazdi, Sam Weber, Eric Xie
Institute: School of Engineering and Applied Science, University of Virginia

Proposed changes to Section 2, Embedding Values into Autonomous Intelligent Systems

We are senior engineering students at the University of Virginia. As young, professional engineers we believe it is vital to embed moral values into artificial intelligent systems. We will be considering the recommendations for Embedding Values in Autonomous/Intelligent systems specifically in the context of autonomous vehicles.

The incorporation of autonomous vehicles into our transportation system has the potential to eliminate the burden of driving, free up time, increase productivity, and increase safety. However, it faces the challenges of meeting safety standards, replacing blue-collar jobs and moral dilemmas in its autopilot algorithms. For these reasons, it is important that we, as a society, play a role in the shaping of this technology.

We believe that autonomous vehicles should, above all else, strive to protect human lives, but there are many scenarios in which this is infeasible. In the case of an unavoidable collision with other vehicles, pedestrians, or property, who should decide what action to take? Should the vehicle act to save as many human lives as possible? Or should specific lives be prioritized be it pedestrians, passengers, or children? We hope that standards are instituted to ensure that the moral values of autonomous vehicles mold to the values of humans in a democratic sense. In the situation described above, we believe that no single manufacturer should be able to design and market their driving algorithm as favoring one particular group over another, but rather this decision be made in a universally consistent manner taking into account all present stakeholders.

As autonomous vehicle technology improves, people will be able to travel greater distances in shorter time than ever before. This improved access to

transportation will likely reshape the way that we work, how we spend our leisure time, and how we form communities. Compared to traditional vehicles, autonomous vehicles achieve this agency via software, for which very few safety standards have been defined. If a negligent manufacturer were to deny vehicle owners of software updates that ensure vehicle safety, the societal consequences could be monumental. Not only should software updates or lack thereof always ensure that certain safety standards are still met, we believe that there should be some type of manual override in cases of service outages or malfunctions so that the vehicle can still be operated safely.

Despite thorough testing, failures will occur. It should be decided who or what should be responsible for autonomous vehicle failures. This issue becomes much more complicated if autonomous vehicles learn by themselves over time. Should the designers, manufacturers or passengers be held liable? And if so, in what particular situations? If autonomous vehicles learn as they drive, should the software itself be liable and how would it be held accountable? A high level set of standards must be in place that all autonomous vehicles should meet. This will be done for designers, engineers and car manufacturers and each group should be held to those standards independently.

Additionally, transparency of the software is crucial for the general public to understand the technology. Machine learning algorithms are complex and are often considered a "black box" that takes in an input and returns an output with no explanation of how it arrived at that output. According to Castelvecchi (2016), deciphering the black box has become exponentially harder and more urgent. The technology itself has exploded in complexity and application. Explaining how these algorithms work, creates an open discussion and understanding that can help mitigate risks and improve accountability. On page 47 when discussing actions in the case of violations, we suggest stating that it may be beneficial for the human passenger to have the ability to take control of the vehicle either through some kind of button or other mechanism.

We would like to thank IEEE for the opportunity to discuss our concerns and to share our suggestions.

Disclaimer: All views expressed by these recommendations are those of the authors and do not represent the opinions of The University of Virginia or any other affiliated entity.

References

Castelvecchi, D. (2016). Can we open the black box of AI? Nature, 538(7623), 20-23. doi:10.1038/538020a

Singh, S. (2015, February). Critical reasons for crashes investigated in the National Motor Vehicle Crash Causation Survey. (Traffic Safety Facts Crash Stats. Report No. DOT HS 812 115). Washington, DC: National Highway Traffic Safety Administration.

IEEE EAD Commenting Report

Authors: Elise Brosnan, Joses Choy, Anthony Quach, Bowei Sun, Yingxiang Sun
4th year Engineering students, School of Engineering and Applied Science,
University of Virginia

Page 114, Section "Reframing Autonomous Weapons Systems"

"The addition of automated targeting and firing functions to an existing weapon system, or the integration of components with such functionality, or system upgrades that impact targeting and automated weapon release should be considered for review under Article 36 of Additional Protocol I of the Geneva Conventions."

We agree that the automated weapons and systems must be reviewed before they can be put into use. The candidate made a good point on the reason: life rights and human dignity must be preserved in any engineering cases. In addition, we suggest that the author comes up with a more detailed reviewing plan. For example, the responsible committee should contain engineers, sociologists, and even social psychologists to analyze the situation in every aspects. The committee, if possible, should build up a system of standards for consistency. Before all of these preparation is put into practice, the automated weapon systems remain questionable.

We suggest to form up an institute or organization to supervise the use and development of all AWS. Its role is not limited to monitoring authorized users, retrospecting login and user records and ensuring safety and security of development process. Also, human control is needed if the system is designed against human beings. Citizens must be provided with the information if they are exposed under AWS operation environment.

We agree with having human supervised systems, as opposed to letting weapons systems work completely on their own. It is important to tie the responsibility of the AWS's actions to some human(s). As long as someone is responsible for the actions of the system, we can have some faith that the proper

precautions will be taken to prevent catastrophe. Although the paper did talk about making humans responsible for the actions of autonomous weapons systems, we think we must be more specific. When an autonomous weapons system goes wrong and causes damage, there should be a way of identifying exactly which parties were responsible for producing or operating the components that caused that failure. Without specifically assigning responsibility this way, irresponsible people can blame their misdeeds on others.

For AWS could be used for covery, obfuscated, and non-attributable attacks, we agree that it is possible that there is a chance if the systems is hacked and used by malicious actors. The issue mentioned "There are significant concerns about the use of AWS by non-state actors, or individuals, and the potential for use in terror attacks against civilians, and non-attributable attacks against states". Also, it mentioned "The lack of a clear owner of a given AWS incentivizes scalable covert or non-attributable uses of force by state and non-state actors". Especially in the age of digital world. We suggest that the designers can also create a unique code for the users to shut down the AWS if there is a situation. The code can only activate by the owner of the AWS with Bio-signification and the owner has to be responsible for the AWS. Beside, we think the designer should seal the inner programming system for the AWS properly and cautious to prevent hackers trying to break in the system and reprogramming the program for malicious use.

Disclaimer: This comment report was prepared and accomplished by Elise Brosnan, Joses Choy, Anthony Quach, Bowei Sun, Yingxiang Sun in their personal capacity. The opinions expressed in this report are authors' own and do not reflect the view of School of Engineering and Applied Science, or University of Virginia.

To the minds working at the IEEE,

We are a group of engineering undergraduate students at the University of Virginia in Charlottesville, VA who have completed coursework on engineering ethics within technology in society. We have reviewed the chapter entitled, "Safety and Beneficence of Artificial General intelligence (AGI) and Artificial Superintelligence (ASI)", from the second version of Ethically Aligned Design and offer this feedback for your consideration.

Section 1

- On page 67 the lack of ethical standards for medical A/IS research is raised. The author fails to mention why this absence exists in this field, and therefore misses the facets of medical research where standards should be developed. Much of current research is focused on using predictive algorithms for drug discovery. However the algorithms used are not meant to immediately develop a product, rather narrow down and expedite current research. The drugs found through these algorithms are treated like any drug found through more conventional means, and need to be tested in vitro and in vivo using current standards before FDA trials are even considered. Standards should be developed in anticipation of future A.I. being used in place of current protocols for drug approval. Perhaps the focus of the medical standards of A.I. should be focused on testing standards to approve A.I. for direct clinical applications such as levels of sensitivity and specificity in comparison to humans, levels of accuracy required to replace animal testing, when human verification of diagnosis is required, and the level of understanding needed by the operator (i.e. when would a physician be required over a medical technician).

- On page 73, there is a stretch of confusing wording when the paper says "the human brain represents one point in a vast space of possible minds." What is the definition of a mind? Minds inherently belong to people or living things, and therefore the "vast space of possible minds" confuses readers not versed in philosophy. The paragraph goes on to an otherwise good argument about how "morality, compassion, and common sense will not be present by default in these new intelligences." Consider re-wording the initial sentence to say

"the human brain only represents one possible model of base intelligence for A/IS."

- On page 74, it is stated "As with other powerful technologies, the development and use of A/IS have always involved risk, either because of misuse or poor design". The risks due to misuse and poor design, while important, are not the only risks involved in A/IS development. There is also a social and economic risk -- for example, will automated vehicles put taxi cab drivers out of jobs? What impact will that have on the economy? These are important factors to consider and should be mentioned here along with the physical misuse or poor design of A/IS.

- Pages 77-78 suggest recommendations to prevent the problems presented in pages 76-77. All the recommendations suggest that teams that work on AI systems should be precautious of the problems that may arise with AI. We propose that not just teams working on AI should be cognizant of the dangers that arise. Individual or independent projects of AI should work to prevent such dangers. In addition, engineers that work on AI are not the only ones that should be precautious, but before the AI is brought to the market, civilians should be precautious as well.

- On page 77, goal number 7, the suggestion of common sense being expanded by building extensive knowledge layers and automated reasoning is incomplete, as the building blocks of these layers is not defined. Common sense built through these layers is subjective towards the programmers that decide these layers' foundations. Therefore, we suggest the addition of a section about ethical building of knowledge layers with impartial experimental samples and ample sample sizes to reduce bias in both social and cultural common sense referenced. For reference of this caution, we refer to *Weapons of Math Destruction* by Cathy O'Neil (ONEIL, 2017).

- On page 78, in reference to the analogy of null terminated strings in C with buffer overflow attacks being one of the most common and damaging types of software vulnerabilities, we suggest mentioning that although this attack does exist, it is just as easily avoidable by using non-executable stacks or size-safe standard input and output functions, which do not allow a user to

read in or out input which is longer than the size of the variable specified and thus prevents buffer overflow attacks. The vulnerability is just as much the problem of an unaware or poorly designed space by the user than that of the developers themselves. For reference, we refer to OWASPs prevention techniques.

Section 2

- On page 80, the writers claim that review boards have been effectively implemented to "scrutinize the ethical implications" of A/IS research activities, without providing any examples where this has occurred. To effectively demonstrate that review boards would truly make a difference, an historical instance should be included where an A/IS was developed without review, and as a result caused unexpected damage through its actions. Google's DeepMind Ethics Board serves only as a precedent for the existence of such a board; it does not, however demonstrate its effectiveness. The Japanese government has previously created such an AI ethics board with clearly-defined operating procedures.

- On page 82, we highly recommend removal of the one-sided background of future A/IS bringing about "unprecedented levels of global prosperity, health, and overall well-being" without mentioning that the addition of A/IS into our economy could just as easily hurt it. Rather, we suggest adding some observations regarding the possible negative consequences of integrating A/IS into our society, and that we should work to increase the possibility of its integration having a positive impact. For example, engineers could first survey the stance of their stakeholders views on A/IS integration into their field of study, as well as give information on how the A/IS will be used in order to understand how users feelings towards the future correlate to the overall well-being, measured post-product deployment.

- On page 83, paragraph 3, the term of privacy is used but not defined. In fact, privacy is the ability of a user to remain anonymous, regardless if their information is public. Confidentiality refers to users who may not be anonymous but have their information hidden. Therefore, we suggest that

these definitions be added for clarity to this paragraph and referred to when mentioning ethical considerations of both.

We appreciate the creation of these standards and hope that these suggestions will facilitate in a more complete report and bring clarification to ambiguities presented in "Safety and Beneficence of Artificial General intelligence (AGI) and Artificial Superintelligence".

Sincerely,

University of Virginia engineering students Angie Campo, Gabriella Greiner, Monique Mezher, Jim Roach, and Noah Rohrlich

Disclaimer: The views and opinions expressed here belong solely to the authors and not to any university or organization with which they might be associated.

Comments on IEEE Ethically Aligned Design - Version Two

Submission by: Josiah Perrin, Derek Boylan, Justin Varnum, Harrison Covert
University of Virginia: School of Engineering and Applied Science

On page six, the mission statement of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems includes the word "stakeholder" as a defining element. The stakeholder definition includes governments as a leading contributor to research, design and development, yet looking at the list of individual and group contributors to the document, there was severely limited input from governmental defense sectors or militaries. We would suggest increasing the amount of input received from the military sector, due to the fact that the military is inherently connected to the development of autonomous systems. In 2018 alone, the United States' defense budget allocated to the research and procurement of such systems increased to $6.97 billion [1]. To extend this point, pages 114 to 140 expand upon the issue of autonomous weapon systems and how to reframe them with regards to engineering standards, military requirements, and government regulations, yet the input from actual military organizations is limited, and is often abstracted to the global level (e.g. the United Nations) as opposed to country-level militaries. Sourcing documents such as the *Unmanned Systems Roadmap (2007 to 2032)* [2], which was written by the United States Army, would greatly benefit the IEEE doctrine by including actual militaristic source material which could serve to enlighten the writer's viewpoint.

For Issue 8 in the Reframing Autonomous Systems section, page 126, a Homeland Security document is cited, but the link provided redirects to ONR/NRL sponsored research. The citation itself is misleading as the IEEE report details the use of swarm-like combat droids, but the drones detailed in the CICADA report are only used to set up networks in hostile territory, not for any combat purposes. We recommend removing this citation as it is misleading and does not help the argument. Intelligence and Defense communities often use obfuscated methods to cloak, hide, or camouflage their operations in hostile territories, while these communities are required to make attacks attributable. Taking issue with covert techniques directly contradicts military doctrine; stealthed attacks have been performed since the conception of modern warfare. We suggest that this issue highlight the importance of making AWS attacks attributable, rather than admonish the fact that AWS can be used for covert actions.

For issue 11, both of the sources cited to support the contention that AWS might be used improperly by police are from the same author, Peter Asaro, a media studies professor at The New School. While the text of this issue contains valuable warnings about use of AWS by private sector and domestic law enforcement, it fails to cite any actual use by these two bodies in the resource section. Including this news citation about the gunman killed by a robot in Texas will improve the validity of this issue [3]. Additionally, more expansive resources supporting this point are required for this issue to be validated. For example, this journal article by Christof Heyns offers many insights into domestic AWS use [4].

On page 113, under the section *Reframing Autonomous Weapons Systems, IEEE list measures they feel will help ensure meaningful human control of weapon systems,* there is a measure defined as such: "The adaptive and learning systems can explain their reasoning and decisions to human operators in transparent and understandable ways." This definition raises issues, as it has already been demonstrated that A.I is capable of coming up with solutions to problems that are uninterpretable by their programmers. An article from MIT Technology Review [5] details a unique autonomous car that used machine learning to create its own driving algorithms. The system that operates the vehicle is so complicated that even the designers of it cannot reliably "isolate the reason for any single action." We recommend that IEEE expand on this point and better articulate what they mean by *clear and transparent ways* (especially in systems where operators may not be able decipher why the system behaves the way it does).

General principles, page 23, states that A/IS should not be granted rights and privileges equal to human rights. Additionally, page 179, states that A/IS may communicate via natural language, may move in humanlike forms, and express humanlike identity. Designs that mimic and portray humans in appearance and mannerisms, to be used as an A/IS system, requires further discussion as to what extent the design is allowed to be indifferentiable from a human accordingly with its application. Continued research to find evidence of correlation between A/IS and suspected impacts will eventually amount to requiring a solution. In Law, section 4, page 159, I provide support of the recommendation outlined in point 1 that there be some government-approved labeling system and/or artifact be clearly visible on the A/IS so that users may immediately be aware when interacting with A/IS in all capacities (androids, chat, data, etc.) that could emotionally affect a user. I offer

the workshop at 'which researchers from various disciplines joined to discuss the application of very humanlike robots to the study of interaction and cognition and the social impact of this technology' [6] as support.

Disclaimer: The views, thoughts, and opinions expressed in the text belong solely to the authors, and not necessarily to the author's employer, organization, committee or other group or individual.

Resources

[1] Gettinger, Dan. *Drones in the Defense Budget: Navigating the Fiscal Year 2018 Budget Request.* http://dronecenter.bard.edu/files/2017/10/Drones-Defense-Budget-2018-Web.pdf

[2] J. Clapper, J. Young, J. Cartwright, J. Grimes *Office of the Secretary of Defense Unmanned Systems Roadmap (2007 - 2032).* Dec 10 2007 https://www.globalsecurity.org/intell/library/reports/2007/dod-unmanned-systems-roadmap_2007-2032.pdf

[3] P. Singer. Police Used a Robot To Kill -- The Key Questions *CNN* July 10, 2016 https://www.cnn.com/2016/07/09/opinions/dallas-robot-questions-singer/index.html

[4] C. Heyns. Human Rights and the use of Autonomous Weapons Systems (AWS) During Domestic Law Enforcement. *Human Rights Quarterly Vol 38*

[5] Knight, W. (2017, May 12). There's a big problem with AI: even its creators can't explain how it works. Retrieved March 03, 2018, from https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/

[6] MacDorman, K. F., & Ishiguro, H. (2006). Toward social mechanisms of android science. *Interaction Studies*, *7*(2), 289-296.

IEEE Commenting Report
02/28/2018

Lucas Abelanet - Joseph Daly - Adam Guo - Ryan Probus - Hans Zhang
University of Virginia

Referencing page 235, Section 4, *The Arts*

Background

In the near future, users will filter their digital landscapes by opting in or opting out of mixed reality information-delivery mechanisms driven by A/IS frameworks that will both structure and, in many cases, alter or curate the data for private, opaque ends. With specific regard to AR, how will the digital public landscape not simply be absorbed by private commercial interests, but allow virtual space for citizens and artists to freely participate? Will artistic content be algorithmically subordinated to commercial content?

Candidate Recommendation on Page 235

Provide users/citizens the option to always "opt out" of any immersive environment to which they may be exposed and provide transparency and consent options to make this possible. This transparency could include not only the constituent algorithms, but also information about the identity of private actors behind the data.

Problems Identified

As the background states, the digital world is indeed being occupied by commercial content that is generated from private information and interests. It is not uncommon that an internet user would be bothered by the banner ads so obviously tailored to his or her private browsing history and habits. While the trend is that there is less and less privacy resulting from commercial targeting, we may take actions to still allow VR users to explore broader range of information outside their living circle.

Recommendations

One recommendation other than the opting-in and opting-out function is to establish a public platform filled with content not relating to any specific users (already filtered by legal standards from backstage). In this platform, VR users are anonymous and private information is hidden. No one user's private preference would dominate the search engine by any chance. Therefore, there's hardly possibility for any party to inject commercial content tailored to anyone's taste. Such a functionality would allow a virtual space where all citizens and artists can freely participate.

The guidelines on page 235 repeatedly mention the "digital public landscape". This "landscape" for augmented reality is public by nature, meaning that everyone will have equal access to it in term of viewing and creating their own content. In this place, who gains control over their image? For example, currently a business is able to control its public image very closely. It can keep its business front clean and clear of obstruction so that it has a favorable public image. Suppose a disgruntled customer decides to graffiti this business in the digital landscape. Does the business get control over its own image, or does the user retain the right to see what they want to see? Is the business allowed to control its own image to the point of changing what people see, or are people allowed to change their view of any image? Both sides of the argument have significant merit. This point changes again when considering corporations who may whitelist businesses that pay for advertising services. If the viewer is opted-in by default to this augmented reality, he/she could be easily bombarded by ads or convinced of some entirely different reality before realizing there is a chance to opt out of this alternate reality. Special care must be taken when it comes to augmented and virtual reality, as this technology greatly changes the ways users interact with the physical world. While the aforementioned candidate recommendation allows the user to opt out, it fails to address the key issue: control over image. Does the power lie with the viewer, who can control what they see, or with the viewed, who can control how they appear? Designers of the public digital landscape must be conscious of the power they are wielding in order to remain thoughtful of the viewers of their content.

Disclaimer**:** The views, thoughts, and opinions expressed in the text belong solely to the authors and are not representative of the authors' companies, schools, organizations, or other groups.

IEEE EAD Commenting Report

Authors: Kyle Burnett, George Ingber, Kelvin Sparks, Brian Morris
Organization: School of Engineering and Applied Science, University of Virginia
Commenting Pages: 136-137

In reference to the issue of "The complexities of employment are being neglected regarding A/IS.":

- The general feelings of the public tend to be that automation takes away significantly more jobs than it creates. Conducting research and providing concrete statistics on how automation has affected certain sectors where it has been implemented would be very beneficial to either confirming or quelling this fear/anger. An article published on IEEE's website by Prachi Patel cites that "less than 5 percent of occupations are likely to be completely wiped out by automation," according to a study performed by the McKinsey Global Institute. The article goes on to mention that the threat of automation is highly dependent on occupation, with industries like mining, trucking, taxis, and manufacturing at greatest risk. However, other sectors like tech and ones that "require empathy, communication skills, and close personal interaction are here to stay for now." Ultimately, it takes time to see the effects of automation in the workforce; however, for now these projections serve as a rough idea of how the job market will be affected by automation.

- The way automation has affected jobs in the past is important, but so is the way experts believe it will effect it in the future. Further research should be conducted on how jobs in various sectors will be created or lost based on the ways in which engineers are currently working to automate certain aspects of the job. We recommend that IEEE conducts a longer-term study tracking both the realized effects of automation as well as changing sentiment over time with regards to automation. In this respect, they can draw conclusions on both the actual impact that automation has, as well as what people believe the future holds. Furthermore, asking specific questions on a case-by-case basis – or even perhaps creating targeted sub-studies – can help focus on the effects for individual industries. For example, if Uber is successful at creating driverless cars, then how many jobs will be lost? Follow up questions can then be used to dig deeper upon initial data collection. An

example of such a question would be that if there were no need for Uber to pay any drivers, how would it affect a country's GDP and economy?

- How will risks increase or decrease when aspects of an industry are automated? This is a serious concern in sectors such as the stock market, where now close to half of all trades made during a day are done by computer algorithms that have been coded by "quants". This type of automation in a sector that, in a way, controls the well-being of the country, can be scary to some. An in-depth analysis of risks posed when the human aspect is taken out of certain jobs would be very useful.

- Increased autonomy allows for quality of life to increase for those who are financially stable and allows for human resources to be dedicated to solving problems of another magnitude of difficulty. However, what is not considered in this increased autonomy is the resources that need to be dedicated to those of a lower class. How will society educate those who have been traditionally underserved so that they then will possess the skills to contribute? Conducting a social study instead of a more economically motivated one would help to answer these questions and we recommend that IEEE or another governing body look into doing this kind of research.

- It is possible that an increase in autonomy will lead to an increase in the wealth gap. An article published by The Guardian in 2017 suggests that automation will not only take jobs, but make the rich even richer and contribute to an already massive wage gap. A fascinating but concerning statistic cited mentions that productivity increased 80.4% between 1973 and 2011, however the hourly compensation of the average worker increased a mere 10.7%. This suggests a trickle-up effect of the profits associated with technological advances. This point is supported by the fact that "the share of the national income that goes to wages has been steadily shrinking, while the share that goes to capital has been growing." We recommend that IEEE looks at economic data detailing wealth disparity over time across various sectors and industries, and references the rate of automation for each industry, in order to determine if there is any kind of correlation between the two.

Disclaimer: The views, beliefs, and opinions expressed in this document are those of the authors, and do not represent those of the authors' employers, peers, or any other organizations to which they belong.

References:

http://theinstitute.ieee.org/ieee-roundup/blogs/blog/will-automation-kill-or-create-jobs

https://www.theguardian.com/technology/2017/mar/02/robot-tax-job-elimination-livable-wage

IEEE EAD Feedback Submission

Siwakorn Chusuwan
Kristina Covington
Charles Pritchett
C. Graham Muller
Charles Yu

4th year engineering students, School of Engineering and Applied Science, University of Virginia

I. Introduction

With the increasing use of learning-based autonomous and intelligent systems (A/IS) such as neural networks, accountability and transparency have become a concern. Learning-based systems only needs input data and a desired model or metric of success. As a result, it is difficult to verify the function of the system and whether it is working as intended. Even the input data can be so large and complex that it is impossible for humans to analyze. The behavior of the system can also change over time as a response to constantly changing the inputs set. The fundamental structure of learning-based A/IS goes against the principles of transparency and accountability to a large extent.

In the present structure of software development, transparency is already a challenge. With the security of the product and trust of the users in mind, companies often exclude transparency from consideration. Even the governing body cannot enforce transparency on private intellectual property. For example, Apple resisted the FBI's demands to crack into an iPhone belonging to the perpetrator of the San Bernardino attacks. The ethical guideline needs to be more aligned with the goals of each stakeholder and cooperate with the structure of learning-based A/IS. With the current EAD principles, specifically on page 27 to 32, some guidelines are not suitable for learning-based systems. The recommended changes below has machine learning as a focus.

II. Recommended Changes

One way of improving the guidelines for transparency is for the owners of the A/IS to tell stakeholders the limitations involved. This suggests informing stakeholders of what data or information they will be using for the A/IS and to what

extent. Stakeholders will be able understand which data is or is not incorporated in the learning based system. Ideally, this will create a sense of respect of privacy and trust between the owners of the A/IS and the stakeholders. On page 30, we recommend adding "Develop different levels of transparency specific to each stakeholder to fit the diverse goal of the stakeholders" to Candidate Recommendations.

A way of improving the guidelines for accountability is for the creators of the A/IS to utilize a set of diverse groups for user studies. Having a diverse group will help creators of the learning based system to understand cultural norms or other concerns about the system they may have not been aware of. On page 28, we recommend adding "Designers and developers should have a set of diverse groups for user studies to understand the cultural norms that they may not have been aware of" and "Intended use of the A/IS and the rights of each stakeholder should be clearly documented to be referenced in the future to enhance accountability" to Candidate Recommendations, and adding "Training data preparation/preprocessing" to the bullet points under number 4 of Candidate Recommendations.

III. Conclusion

The topics concerning transparency and accountability are increasingly crucial in the field of automated technology. In case of accident, users and other stakeholders demand accountability. Authority will demand the root cause of the issue and determine whether or not the creators are ultimately responsible. Of course, due to cultural differences in norms, knowing which norm is tied to which aspect of A/IS would help immensely.

Some issues that still need to be addressed include the actual scope of transparency. Will users receive A/IS that is so transparent that the source code is open source, for example? Or sometimes the manufacturers may still skirt the rules and only reveal issues that do not hurt their bottom line. As for accountability, if someone or some group is found of wrongdoing, how would that individual or group be instigated? There is still the enforcement issue that needs to be addressed. If that individual or group is from another country, how would it be possible for them to admit wrongdoing? As always, any improvements in A/IS must be clear of any loopholes.

Disclaimer: The views, thoughts, and opinions expressed in the text belong solely to the authors, and not necessarily to the author's employer, organization, committee or other group or individual.

References

Muehlhauser, L. (2013). Transparency in Safety-Critical Systems. *Machine Intelligence Research Institute*. Retrieved from https://intelligence.org/2013/08/25/transparency-in-safety-critical-systems/

Nakashima, E. (2016). Apple vows to resist FBI demand to crack iPhone linked to San Bernardino attacks. *The Washington Post*. Retrieved from https://www.washingtonpost.com/world/national-security/us-wants-apple-to-help-unlock-iphone-used-by-san-bernardino-shooter/2016/02/16/69b903ee-d4d9-11e5-9823-02b905009f99_story.html?utm_term=.e8b09ae99d34

Russel, J. (2017). Government requests for Facebook user data continue to increase worldwide. *TechCrunch*. Retrieved from https://techcrunch.com/2017/12/18/government-requests-for-facebook-user-data-continue-to-increase-worldwide/

In "Ethically Aligned Design," the IEEE Global Initiative expresses a commitment to human rights and human well being, suggesting their vision for a future in which autonomous and intelligent systems contribute to the flourishing of people around the world. I commend the authors of this document for defining technological progress in such an inclusive and humane manner. Yet I would also challenge them to adopt goals and objectives that more fully realize this vision. In the comments that follow, I recommend ways that each of the General Principles articulated in "Ethically Aligned Design" (p 6) could go further in promoting human flourishing, and encourage the inclusion of two additional General Principles oriented to human rights and well being.

The general principle of "Human Rights" should go beyond not infringing on internationally recognized human rights to proactively promoting human flourishing. I recommend adopting the framework of "capabilities" articulated by theorists Amartya Sen and Martha Nussbaum, who argue that all people deserve the material and institutional conditions to choose what they will do and become. (See, for example, Martha Nussbaum, *Creating Capabilities: A Human Development Approach* [Harvard University Press, 2011].) The first general principle might then read, "**Human Rights: Ensure they enhance human capabilities for people around the world, especially those most disadvantaged**."

The general principle of "Well-Being" argues for "metrics of well-being" to be prioritized in the design and use of Intelligent and Autonomous Systems. While this is an important principle, it is incomplete without specifying how, and especially by whom, metrics of well-being should be defined. For Intelligent and Autonomous Systems to promote human flourishing, the metrics by which they are judged should be created in collaboration with the populations most affected by those systems, including both the intended users or beneficiaries and those groups whose lives might be unintentionally transformed by them.

In addition, "metrics of well-being" should be treated as vehicles for learning and potential mid-course corrections. The effects of Intelligent and Autonomous Systems on human well-being should be monitored continuously throughout their lifecycles, and designers should be prepared to significantly modify, or even roll

back, technology that is shown to reduce well-being, as defined by affected populations.

The second general principle might then read: "**Well-being: Use collaboratively created metrics of well-being to inform their design, shape their uses, and determine whether they remain in use in the long-term**."

The general principle of "Accountability" calls for designers and operators of Autonomous and Intelligent Systems to be responsible and accountable. Responsibility and accountability are, of course, essential; however, anthropological research demonstrates that engineers and corporations frequently do not understand responsibility or accountability in the same way as do the people affected by their activities. (See for example Gwen Ottinger, *Refining Expertise: How Responsible Engineers Subvert Environmental Justice Challenges* [New York University Press, 2013].) It is thus not only possible but likely that a designer who had internalized rigorous principles of corporate social responsibility might nonetheless be acting in ways that those advocating for social justice or human rights would find unhelpful or even detrimental. To avoid this, IEEE guidance for Ethically Aligned Design should treat "accountability" and "responsibility" as matters of collective judgment, to be defined in on-going, proactive deliberation with diverse stakeholders. The third general principle might read, "**Accountability: Designers and operators should be accountable to context-specific, culturally and historically appropriate standards of social responsibility, developed through ongoing deliberation with those affected by their systems**."

In the general principle of "Transparency" (p 6), it is not clear who or what is being called upon to act in a transparent manner: the designers, technologies, or both. I strongly suggest rewording the principle to specify who or what ought to be transparent. Later in the document (p 28-30), it becomes more clear that the systems themselves are meant to be transparent. However, the discussion in the executive summary, which outlines the potential complexity and opacity of A/IS suggests that system builders' decisions will play a large role in determining how transparent the systems are. The obligations of transparency should be extended to designers and operators, as well.

Transparency is, however, a highly circumscribed concept. The principle suggests that the public should have the right *view* the inner workings of the technology—but it does not imply any right of the public to *intervene* if the technology is doing harm, or if the decisions of designers result in excessive risk or other detrimental effects. In keeping with the democratic notion of procedural justice—that is, that individuals affected by decisions ought to have meaningful opportunities to influence their outcomes—the principle of transparency should be augmented with a principle of participation. The fourth principle might then read: "**Transparency and Participation: Ensure systems and their designers operate in a transparent manner, and that affected populations have on-going opportunities to influence systems design, implementation, and outcomes.**"

The general principle "Awareness of Misuse" is similarly too limited in its scope. While it is important to be vigilant against misuse, the responsibility of designers and operators should be conceived more broadly. In "Responsibility and Global Justice: A Social Connection Model" (*Social Philosophy and Policy* 23[1]: 102-130), as well as in *Responsibility for Justice* (Oxford University Press, 2011), Iris Young argues that, in a globally interconnected world, a strict "liability model" of responsibility is not sufficient. Instead of being responsible only for those consequences that stem directly from one's actions or products, each person should be attentive to the ways that her or his activities support large-scale systems (e.g. global energy production) and institutions, and take responsibility for addressing the harms and injustices that arise in the context of those structures. The IEEE's principles for Ethically Aligned Design would do far more to further human well-being if, instead of using the narrower, liability-oriented model of responsibility that "minimiz[ing] the risks of misuse" implies, they embraced a "social connection" model that asks designers and operators to be aware of how technology functions in structures of global power and inequality. The fifth general principle might then be, "**Awareness of Systemic Impact: Strive to minimize its contributions to social and political structures that are detrimental to human well-being, and to contribute to their dismantlement wherever possible**."

In addition to the five principles originally in the document, revised along the lines described above, I would advocate the inclusion of two additional general principles.

A general principle of "Justice" should be added, in recognition of extensive research showing that economic, racial, and other inequalities are, more often than not, reinforced or deepened by technological advance. "Ethically aligned design" of A/IS (or, for that matter, any technology) should strive to reverse this effect, and to address the contributing factor of gender and racial disparities among designers. A sixth general principle might thus read: "**Justice: Ensure they neither deepen economic inequalities nor entrench racism, sexism, or other forms of prejudice**."

Finally, a general principle of "Targeted Development" should be added. Not all new technology benefits human well-being to the same extent. Expending finite resources on innovations that offer only marginal benefits, or that benefit only highly advantaged populations, does little to advance a human flourishing. Designers should acknowledge that not every system that could be built deserves to be (i.e. "can" does not imply "ought") and target their energies to those systems that appear likely to offer significant benefit, when viewed in terms of the principles of Human Rights, Well-being, and Justice. A seventh general principle might read, "**Targeted Development: Invest in systems most likely to contribute to the development of human capabilities, and defer those of marginal benefit**."

These changes to the general principles would, I believe, align them more closely with the human rights-oriented ethics that the IEEE Global Initiative has so commendably chosen as a reference point. Were these changes to be propagated through the specific objectives listed in the report—as I recommend they should be—the guidance in this document would be far more powerful in promoting human flourishing as a primary goal of Autonomous and Intelligent Systems.

Thank you for your work in preparing this document, and for the opportunity to comment. Please contact me if you have questions or would like further elaboration related to any of the ideas above.

Gwen Ottinger
Associate Professor
Department of Politics
Center for Science, Technology, and Society
Drexel University
www.fairtechcollective.org

IEEE EAD Commenting Report

Authors: Samuel Boakye, Kareem El-Ghazawi, Chase Deets, Henry Hubler, Raquel Moya
Institute: School of Engineering & Applied Science, University of Virginia

On Defining Human Rights

On page 6, the IEEE EAD report talks about human rights as though there is an international standard of human rights and also as though all humans internationally currently are granted these rights. What the authors fail to consider is that AI/S has the potential to exacerbate the existing injustices in granting of human rights in many areas of the world. We recommend that the report is more explicit about how communication with international human rights groups is vital to the safe integration of AI/S into our world societies.

On Embedding Values Into Autonomous Intelligent Systems (pp. 33-54)

On page 36, the updating system for an A/IS is determined to be driven by the norms of a community. This is meant to help the implementation of an A/IS by making it fluidly adjust to the community it serves. This is well explained; however, we recommend including clarification on how a community will be defined, especially within different cultures. According to James Miles, what defines a community varies, often dramatically between two different societies. According to his *Ubuntu and Defining Community in America,* a community can be defined by those living in the same geographical area, or those simply having similar interests/experiences (Miles, 2017). Beyond that, there are two ways of viewing a community: either generally or specifically. The chalk outline defining a community can be drawn of many sizes and based on many criteria (defined by values, beliefs, degree of interaction, degree of separation). If an A/IS is to be dependent on a community for its updates and improvement, we recommend that IEEE be more specific on how communities will be defined.

The idea of defining communities also emerges on page 40 when the EAD report mentions giving more value to community norms over individual norms. Here is an example of restricting derogatory language from an individual because it goes against the norms of a community. If this philosophy is held (community is more highly valued than the individual), then individuals may never be able to rid their communities of inappropriate beliefs. A community can be racist and the A/IS will adjust to those norms, leaving some individuals isolated. If the IEEE wants to follow this philosophy, we recommend describing these limitations and unfavorable scenarios. Of course, if a small community has racist norms, those norms might be overridden by the larger community, but here again we run into the problem of how exactly we ought to define the communities that have the final say in the updating of an A/IS. What kind of community will be given priority? Who will determine that? These issues, if not solved, at least should be made apparent.

On Methodologies to Guide Ethical Research and Design (pp. 55-72)

It is stated that in order for the methodological goals to be achieved, transparency needs to be embraced between researchers and technologists in terms of processes, products, values, and design practices. However, the only transparent relationship specified is that between an educational institution and engineering students. The categories of researchers and technologists are quite broad and encompass different groups other than educational institutions and students (such as companies, research organizations, etc.). In studies observed by Donald Marquis and Thomas Allen, the flow of communication between researchers and applied technologists is seen as one that requires various forms of organization in order to be effective (Marquis and Allen, 1966). To this end, we recommend that the transparency that is stated as needing to be embraced be specified for all relevant groups of researchers and technologists. There should be an outlined organization of how the transparency should work between different groups in order for the flow of communication to be effective.

On Presenting Ethics to Creators of Autonomous and Intelligent Systems (pp. 198-199)

In the section pertaining to the issue of presenting ethics to creators of autonomous and intelligent systems on page 198, it suggests finding ways to present ethics where the methodologies used are familiar to engineering students. The candidate recommendation is something quite applaudable, as any engineer should have some type of moral compass guiding all their work. In fact, technical ingenuity is human, and the phenomena it produces are part of our identity (Verbeek, 2011). But to say that the solution can be addressed by providing students with job-aids to help them select and use a principal ethical framework, and then exercise use of those devices with steadily more complex examples simply entails that the process itself deals with the practical decision making skills that in turn have no correlation to the engineering students' morality.

Additionally, the 'macro code' described in the section as built from Western ethics tradition raises the question: How biased is it to the Western ways? While this was something addressed in the first iteration of revisions, we recommend extending upon this comment on **how** non-Western traditions may (or may not) foster a foundation for cross-cultural understanding and respect. What does this mean for resulting ethical system regarding A/IS if various religious traditions from the East juxtapose the section on pages 203-211.

On Mixed Reality in Information and Communications Technology (pp. 217-239)

One issue that arises when examining the future of Mixed/Virtual Realities is the effect that these systems will have on cultural institutions when geographies are eliminated in a virtual world. The current focus is on the aspects of human connection and cultural interactions that "cannot be digitized" - including aspects of personal and cultural identity. The recommendation currently proposes to provide widespread educational classes on the benefits of positive human connection/touch

including the fields of emotional intelligence and positive psychology. We believe that the recommendation deviates from the central issue of how culture will be expressed and experienced in a virtual world. Perhaps a more appropriate approach to answering these questions would be to add social scientists into the conversation and promote research focused on the ways in which VR/MR applications can allow for cultural and personal identity. Educational classes should include modules social science (anthropology, sociology, etc.) to promote learning new ways cultures will and have reacted to disruptive technologies in ways that change but do not "diminish" the cultural merit and identities of each population. Cultural education will allow for users to embrace and recognize cultural identities and their unique manifestations in next-gen realities - similar to the way Christine Spiteri suggests ways in which our cell phones have been interacted with differently by cultural groups which has given rise to new ways to identify and study cultural groups, rather than diminish them (Spiteri, 2013).

References

Fredrickson, B. L. "Your Phone Versus Your Heart" (Sunday Review). New York Times, March 23, 2013.

Miles Sr, J. L. (2017), Ubuntu and Defining Community in America: A 21st Century Viewpoint. Anthropol Conscious, 28: 178–186. doi:10.1111/anoc.12079

Marquis, D. G., & Allen, T. J. (1966). Communication patterns in applied technology. *American Psychologist, 21*(11), 1052-1060.

Spiteri, Christine. (2013). Cultural Identity construction through Smartphone Use. 10.13140/2.1.1902.9286.

Verbeek, P. P. (2011). *Moralizing Technology: Understanding and Designing the Morality of Things*. University of Chicago Press.

Disclaimer: This report was prepared or accomplished by the authors in their own personal capacity. The opinions expressed in this article are the author's own and do not reflect the view of the University of Virginia as an institution.

The emergence of autonomous and intelligent systems (A/IS) has undoubtedly enhanced the human condition through the disinvolvement of human actors from unpleasant tasks. However, as these systems tend towards increased complexity and higher degrees of freedom, it will become increasingly difficult for the average person to discern the many functionalities obscured by the veil of autonomy. Therefore, it is our stance as physicists and mechanical, aerospace, and electrical engineers currently developing autonomous systems, that engineers must commit to transparent design. We focus the critique presented in this document on Section 2, Issue 2: "The need for transparency from implementation to deployment," found on pages 44-46 of EAD - Version 2.

It will be critical to ensure that any communication, be it communication of the format of any data being collected, the ultimate purpose of any data being collected, or any other requisite information regarding the functionality of an A/IS, adhere to and respect the cultural norms of those people with which an A/IS interacts. Doing so is necessary to be certain that the intended message of any communication not be malformed by a poorly understood inter-cultural filter. It may be that the engineer of an AI/S exists in a culture with norms of communication incommensurate with the communication norms of a culture with which the AI/S they design interacts. Such a situation is detrimental to transparency of function and purpose.

Notably one must consider standards of non-verbal communication, which frequently do not translate across cultures. For example, associations of color may vary between the culture of an engineer and a culture with which the AI/S they design interacts. While green tends to be the globally-recognized color for a 'Go' signal in traffic lights, oftentimes traffic lights in Japan use blue to indicate the same. Another example of this would include the direction of written text affecting the order of objects listed on a screen (Hebrew is written right-to-left). These considerations have already been included in the study of user-interfaces generally, by they are equally important considerations for the "non-deceptiveness" of artificial intelligences, especially those which cross cultural delineations. Such

inhomogeneity in otherwise ubiquitous standards are particularly pernicious and so care must be taken to ensure local standards are understood. Complications invariably arise in situations where an AI/S is used across cultural boundaries. In such cases the engineer will need to find a sufficient way to ensure that communication appropriately specific to each culture is provided and that it is clear which communication is intended for which culture. Given that facets of the human body are universal without significant morphological differences between nations, we recommend that the EAD's discussion on honest design on page 48 of Section 2 Issue 2 be augmented with the addition "Where possible, external shapes that indicate ears and eyes are highly recommended in order to most clearly convey the intentions of the systems and devices, within the social-contextual considerations listed here and elsewhere in this document."

It is important to recognize cultural differences additional to solely those geographical. There exist as well cultural divides between various professional and nonprofessional spheres. While it may appear to an engineer that some message of function or purpose is adequately communicated in the form of an AI/S, such an interpretation may rely upon some domain-specific knowledge upon which the engineer is unaware they are drawing. One must ensure that the "common knowledge" of those people with which the AI/S will interact is adequately understood and employed in communication efforts.

To this end, it might behoove the IEEE to include in their 7000 standard two things: firstly, language that ties it to an inclusion of existing UI study, as that will translate easily into AI; and secondly, a possible study into the creation of an internationally standardized set of symbols that may be added to AI operating in these contexts, reducing the complexity of localization.

This suggestion is presented with limitations, as not every A/IS empowered system requires full public disclosure of every detail. A public street camera is immediately recognizable as such (given open placement), but a USB port symbol is not. Thus, A/IS devices need to indicate their extended functionality beyond that which the chassis, supporting a less sophisticated control system, would

immediately entail. For more specialized indicators, such as the usage of particular radio bands, indicator symbols should give a fairly high-level indication so as not to enable deception by overwhelming the eye with a barrage of symbols and markings. Where possible, recommendations should be made for functionality indicators to combine and streamline.

In particular, this disclosure may take different forms depending on the context and intended user base of the intelligent devices. For example, an AI-enabled robotic tour guide for a museum would interact primarily with persons with little to no technical training, whereas an intelligent software package for, say, a recommender system, would be used primarily by people with a technical understanding of the software, and artificial intelligences in general. In the latter case, communicating the capabilities of the AI through design choices would place an undue burden on the developers to communicate indirectly what can easily be communicated in a manual or other documentation. In the former case, users will not have access to (or an easy understanding of) such documentation, which would require designers to communicate through indirect design choices (such as the inclusion of artificial ears or eyes, as suggested in the standard.

Therefore, we recommend this honesty requirement be specified not through a particular set of definitions, or through a small set of examples (as it currently appears to be in the standard), but with a general policy that may be adapted to various contexts and users. We recommend that the EAD's intelligibility section in Section 2 Issue 2 be supplemented with "As much as possible, the A/IS system should preempt attempts at understanding and signal key facets of its decision making process and readily display intended actions and justifications. Where possible, this could be factored into the design of recognizable external indicators mentioned earlier. For example, eyes for visual processing having some manner of indicator where they are focused or heading in the case of moving A/IS systems."

Finally, the ideas of "honesty" and "non-deceptiveness" included in the specification all assume that the user of the AI/S already has the knowledge that the intelligence in question is, in fact, artificial. However, we recommend that it

also be included in the standard a requirement that an artificially intelligent system disclose (in a manner following the other recommendations in this document) that it is artificial. This would be most applicable to internet chatbots, for example, where any real intelligence would be hidden from the user.

The views expressed herein represent solely the opinions and recommendations of the authors and not of any employers, organizations, committees, etc. by which the authors or employed or to which the authors belong or of any other individuals or groups.

Recommendations from:

Matthew R. Anderson, University of Virginia
Andreas L. Butler, University of Virginia
Andrew G. Coffee, University of Virginia
Paul J. Hughes, University of Virginia
John R. Walnut, University of Virginia

Pradyot Sahu
Director, 3innovate, India
https://www.linkedin.com/in/pradyot-sahu/

1. RISE OF A/IS PLATFORMS

Recently, there are new A/IS Platforms. The platform providers are the front-runners of A/IS such as Google, Microsoft and Amazon enabling A/IS implementations in each mobile, web or stand-alone app and applications. As practically every A/IS system that use an A/IS Platform to implement will be difficult to evaluate and monitor just like it is difficult to monitor each android mobile app. 1) It will be easier to evaluate and monitor a platform and make the platform builders liable to the problems they create intentionally and unintentionally. 2) Platform builders must use platform controls to enable control of A/IS applications against any potential misuse of A/IS technology

2. EVALUATION AND COMPLIANCE

Without evaluation and compliance, there is no way to make AI Platform providers accountable. There may be national and international organizations in the form of national level HIPAA style evaluation and compliance and United Nations Regulatory Agency such as International Labor Organization to regulate A/IS.

Best regards

Pradyot Sahu
Director, 3innovate, India
https://www.linkedin.com/in/pradyot-sahu/

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

Comment submitters: Tony Nguyen, Chris Anton, Chris Mooney, Yihnew Eshetu, Martin Simpkins (4th year engineering students)

Institute: School of Engineering & Applied Science, University of Virginia

Comments:

- There should be a clear ability for designers and review boards to track the logic of super intelligence and artificial intelligence in retrospect of an incident or decision, especially when deciding between life and death. If an incident were to occur, it would be extremely important to have the ability to look back at what variables and calculations led to that decision. This is something we thought could have been missing from the overview of Safety and Beneficence of Artificial General Intelligence (AGI) and Artificial Superintelligence (ASI) on pages 73-75. These intelligence and automated programs should have clear overall objectives, goals, and restrictions in order to track decisions they make. Again, there should be clear objectives and a standard in place to specify ethical motives or objectives that review boards and designers should follow when designing the decision process of AI/SI.

- Along the sames lines, in the overview of this topic on pages 73-75, it would be interesting to see a discussion on who is responsible for decision made by superintelligence after the creation and release of a program. Do we hold the creator responsible? Do we hold the super intelligence responsible? And if we hold the AI responsible how to we punish mistakes or wrong decisions that it makes? There should be a standard to put a specific human responsible for actions made by an autonomous program, and that human should have ultimate control and/or veto power for any actions that program might make.

- Who should be on the review board? Should the public be involved? Regulators? Legislators? Since these programs are designed to operate autonomously, all the stakeholders in the autonomous decisions should be involved to some degree in the design of the decision programs. It would be interesting to also have this discussion on page 81, where it is recommended to set up review boards.

- What does benefit of humanity mean? Who decides what benefits humanity. That should be a part of the design and standard. How does this benefit humanity and how does one decide if benefits outweigh the costs to society? This discussion and standard takes place on page 82.

- What are the current barrier stigmas to AI research? Maybe we should have stigmas? Public stigma is the essences of public opinion and should certainly be taken into account when designing something that the public might interact with?

Paper discussing this issue:
https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/viewFile/14581/13868

- On page 78, from personal anecdotal design experience, it's sometimes extremely difficult to know where issues will arise along the design process. If you prioritize spending at the beginning, you inherently de-prioritize spending and testing at the end, where the issues might only be discovered through working through the life cycle of research and design. This standard could be split up into adding testing at and design both at the beginning and end, instead of trying to shift spending from the end to the beginning.

- On page 73, AI was created to serve our intention of doing but without physically being there. How is creating a so called "weapon" is safe and ethical from this perspective? There is no such thing as standard for safety when creating a weapon.

- On page 82, how would we adopt the superintelligence to only serve the benefit for everyone? We have not taken into account of war.
- On page 76 under the background, what does distributional shift refer to? How would we be able to tell that AI is having their own consciousness to solve and react to our orders?
- The main gist from page 78 is difficult to design a safe AI, how so? We create weapon that has AI to eliminate target, how is that called a safety feature?
- On page 77, the importance of transparency in logic for AI systems is a valid point. I think it could be useful to cite the need for transparency in autonomous vehicles as an example. If a car trained by machine learning/neural networks crashes it may be impossible to trace the logic back which is necessary when determining the cause of incidents. This journal outlines some of the issues with autonomous vehicles ethically, and goes into the issue of tracing the logic of machine learning in autonomous vehicle applications.

  http://trrjournalonline.trb.org/doi/abs/10.3141/2424-07
- On page 78, I think the C analogy falls somewhat short because the difficulty of modifying AI algorithms that evolve is not necessarily the same as the C string design implementation. The C string situation was a design choice completely within control of the creators from the offset, while the AI challenge is an artifact of how AI evolves itself.

Response to Ethically Aligned Design Version 2 (EADv2)

Rod Rivers, Socio-Technical Systems, Cambridge, UK

March 2018 (rod.rivers@ieee.org).

I take a perspective from philosophy, phenomenology and psychology and attempt to inject thoughts from these disciplines.

Social Sciences: EADv2 would benefit from more input from the social sciences. Many of the concepts discussed (e.g. norms, rights, obligations, wellbeing, values, affect, responsibility) have been extensively investigated and analysed within the social sciences (psychology, social psychology, sociology, anthropology, economics etc.). This knowledge could be more fully integrated into EAD. For example, the meaning of 'development' to refer to 'child development' or 'moral development' is not in the glossary.

Human Operating System: The first sentence in EADv2 establishes a perspective looking forward from the present, as use and impact of A/ISs 'become pervasive'. An additional tack would be to look in more depth at human capability and human ethical self-regulation, and then 'work backwards' to fill the gap between current artificial A/IS capability and that of people. I refer to this as the 'Human Operating System' (HOS) approach, and suggest that EAD makes explicit, and endorses, exploration of the HOS approach to better appreciate the complexity (and deficiencies) of human cognitive, emotional, physiological and behavioural functions.

Phenomenology: A/ISs can be distinguished from other artefacts because they have the potential to reflect and reason, not just on their own computational processes, but also on the behaviours, and cognitive processes of people. This is what psychologists refer to as 'theory of mind' – the capability to reason and speculate on the states of knowledge and intentions of others. Theory of mind can

be addressed using a phenomenological approach that attempts to describe, understand and explain from the fully integrated subjective perspective of the agent. Traditional engineering and scientific approaches tend to objectify, separate out elements into component parts, and understand parts in isolation before addressing their integration. I suggest that EAD includes and endorses exploration of a phenomenological approach to complement the engineering approach.

Ontology, epistemology and belief: EADv2 includes the statement *"We can assume that lying and deception will be prohibited actions in many contexts" (EADv2 p.45).* This example may indicate the danger of slipping into an absolutist approach to the concept of 'truth'. For example, it is easy to assume that there is only one truth and that the sensory representations, data and results of information processing by an A/IS necessarily constitute an objective 'truth'. Post-modern constructivist thinking see 'truth' as an attribute of the agent (albeit constrained by an objective reality) rather than as an attribute of states of the world. The validity of a proposition is often re-defined in real time as the intentions of agents change. It is important to establish some clarity over these types of epistemological issues, not least in the realm of ethical judgments. I suggest that EAD note and encourage greater consideration of these epistemological issues.

Embodiment, empathy and vulnerability: It has been argued that ethical judgements are rooted in physiological states (e.g. emotional reactions to events), empathy and the experience of vulnerability (i.e. exposure to pain and suffering). EADv2 does not currently explicitly set out how ethical judgements can be made by an A/IS in the absence of these human subjective states. Although EAD mentions emotions and affective computing (and an affective computing committee) this is almost always in relation to human emotions. The more philosophical question of judgement without physical embodiment, physiological states, emotions, and a subjective understanding of vulnerability is not addressed.

Terminology / Language / Glossary: In considering ethics we are moving from amoral mechanistic understanding of cause and effect to value-laden, intention driven notions of causality. This requires inclusion of more mentalistic terminology. The glossary should reflect this and could form the basis of a language for the expression of ideas that transcend both artificial and human intelligent systems (i.e. that is substrate independent). In a fuller response, I discuss terms already used in EADv2 (e.g. autonomous, intelligent, system, ethics, intention formation, independent reasoning, learning, decision-making, principles, norms etc.), and terms that are either not used or might be elaborated (e.g. umwelt, ontology, epistemology, similarity, truth-value, belief, decision, intention, justification, mind, power, the will).

Berkeley Fergusson, Christian Halsey, Clark Kipp, Robert Wallace
University of Virginia

Referring to "When Systems Become Intimate" on pages 168-169

Point 1 - Developing intimate systems which do not contribute in some ways to sexism, negative body image stereotypes, and/or gender or racial inequality may not be feasible, in a capitalistic society people will make products that are demanded, thus if demand arises for a particular type/behavior of robot it will likely be filled. To achieve a market in which these types of robots do not exist, stringent laws/guidelines would have to be adhered to, or we need to know how to address these issues rooted in the demand-supply chain in consumer society.

Point 2 - This comes to whether responsible use is left up to the user or the creator. If the user is held accountable, there should be no reason to require an "opt-in" solution. However, it is true that users do not always understand the product they are using and thus some accountability is left with the creator to provide guidelines/disclaimers.

Point 3 - We agree that it is important for human-to-human interactions to be maintained as the development of intimate systems increases. However, if a user agrees to opt-in to an intimate system and have sexual relations with a robot, it seems like user isolation from other human companions is inevitable to a certain extent. Therefore, the implementation of regulations or laws to create barriers to user isolation from other human companions might be infeasible or difficult to legislate.

Point 4 - If and when having a robot partner becomes a norm, some form of disclaimer about the effect it can have on human relationships would be beneficial. If people decide to actually dedicate their lives to a robot partner, it is inevitable their relationship dynamics with humans will be altered. Obviously this is a far off if

ever future scenario, but we think disclaimers stating how new bots can affect human relationships is necessary in some form. In the meantime, it could be useful to research how humans interact with AI and see how serious this problem could actually be.

Point 5 - The recommendation to prevent intimate systems from fostering deviant or criminal behavior is an important guideline in the development in these systems. One potential solution to this recommendation could be administrative control by the intimate system company or robot owner to prevent the robot from promoting deviant behavior or harming others. Whether this should be done directly by the company/owner to restrict the bots or should be left as a disclaimer meant to warn users is a point for debate.

Point 6 - This recommendation to not view commercially marketed AI as humans is unfortunately very hard as there is scientific research dealing with the "anthropomorphism of nonhuman things" meaning attributing human characteristics to objects. This includes the fact that items tend to "reflect back to us gendered notions of sexuality" (Richardson, 2015). This idea would especially be evident in AI units that are supposed to be modeled to function like humans. We need better standards in terms of terminologies, companies' organizational accountability and social responsibility to achieve this task.

Disclaimer:
The views, thoughts, and opinions expressed in this text belong solely to the author, and not necessarily to the author's employer, organization, committee or other group or individual the author has been affiliated with.

References

Richardson, K. (2015). The asymmetrical 'relationship': Parallels between prostitution and the development of sex robots. *Special issue of the ACM SIGCAS newsletter, SIGCAS Computers & Society, 45(3):290–293, 2015.*

Folke Hermansson Snickars

Standards ambassador, MyData Global

Only a typo on p.249: ISO 2900 should be ISO 29000:2010 Guidance on Social Responsibility

Autonomous Systems and the Value Alignment Problem

Martin Peterson
Texas A&M University

1. Introduction

In May 2016 a Tesla Model S crashed into the trailer of a big-rig truck at a speed of 74 miles per hour because the autopilot failed to identify the truck and stop the vehicle.[1] Joshua Brown, 40, died immediately. It may not be trivial from a technical point of view to design autonomous vehicles that avoid this type of accident, but we have no doubt that doing so would be ethically desirable.

Unfortunately, some ethical issues triggered by autonomous systems are more complicated.[2] Should autonomous vehicles be instructed to protect pedestrians on the sidewalk to the same extent as their passengers? (Mercedes-Benz has announced that it will prioritize occupant safety over pedestrians.[3] Philosophers have been quick to point out that this is a modern version of the Trolley problem.[4]) Another, somewhat more visionary example is the following: If future autonomous systems become smarter than us they may eventually decide to prioritize their own ends at our expense. Is this something we should feel concerned about? According to Nick Bostrom, supersmart autonomous systems

[1] Tesla's autopilot mode is marketed as a semi-autonomous system, not as a fully autonomous one.

[2] I leave it open whether autonomous systems make decisions, or if all decisions are ultimately made by the engineers who design these systems. For the purposes of this paper there is no need to ascribe moral agency to autonomous systems.

[3] See Taylor (2016).

[4] See e.g. Goodall (2016) Carfwod (2016). Nyholm and Smids (2016) question the analogy.

pose an existential risk to our species, because they may treat humans like cute pets in a zoo, or kill us like cattle in a slaughterhouse.[5]

In response to all this, a number of thinkers have suggested that we should impose our own values on autonomous systems to ensure they serve human needs and wishes. Stuart Russell calls this *the value alignment problem*: How can we build autonomous systems with values that "are aligned with those of the human race"?[6] As he sees it, the challenge is to build autonomous systems with ethical priorities that do not pose a threat to us. He claims that, "The machine's purpose must be to maximize the realization of human values."[7]

The IEEE, the world's largest professional organization for engineers, has issued a report on the value alignment problem entitled *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems* (A/IS).[8] This document is part of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. The authors claim that autonomous systems "should always be subordinate to human judgment and control. […] If machines engage in human communities as autonomous agents, then those agents will be expected to follow the community's social and moral norms."[9]

In this article I shall defend two claims about the value alignment problem:

(i) It is not obvious that it is desirable to build autonomous systems with values that "are aligned with those of the human race". This is a

[5] See Bostrom (2014) for an extensive discussion of this topic. See also Dafoe and Russell (2016).

[6] The quote come from a talk Dr. Russell gave at the World Economic Forum in Davos, Switzerland in January 2015. The talk is available on Youtube (https://www.youtube.com/watch?v=WvmeTaFc_Qw). Russell has also expressed the same idea in the papers listed among the references.

[7] Russell (2016: 59).

[8] IEEE (2017a).

[9] IEEE (2017a: 23, 36).

substantial moral claim that needs to be critically discussed and supported with arguments.

(ii) The methods currently applied by computer scientists for embedding moral values in autonomous systems can be improved by representing moral principles as conceptual spaces, i.e. as Voronoi tessellations of morally similar choice situations located in a multidimensional geometric space. It is probably a mistake to use utility functions.

The first point will not surprise philosophers, but I think it is worth stating clearly. The second point, which is perhaps the most interesting one, requires some unpacking before it can be assessed and discussed.

In what follows I will not make any predictions about the technical capabilities of future autonomous systems. All my claims will be hypothetical: If such-and-such technologies become available, then we ought to reason and behave in such-and-such ways. It is also worth pointing out that I will refrain from discussing some of the most controversial issues related to autonomous systems, such as the Trolley problem. The only claims I defend are (i) and (ii) stated above.

Section 2 argues that it would be a mistake to *always* align the values of autonomous systems with those embraced by human beings, since humans are sometimes wrong about what is valuable. Section 3 discusses some links between general ethical theories and the value alignment problems. I point out that because experts disagree on which ethical theory is correct, it would be ill-advised to base a solution of the value alignment problem on *any* ethical theory. My argument for this claim is closely intertwined with my suggestion for how to represent moral principles in autonomous systems, which is discussed in Section 4. The point of departure for this proposal is some ideas developed by two leading cognitive scientists, Eleanor Rosch and Peter Gärdenfors. I note that some of their insights apply to the ethics of autonomous systems.

2. Four versions of the value alignment thesis

Previous discussions of the value alignment problem have not been terribly precise. So let me clarify a few points before I defend my two claims.

To start with, it seems reasonable to assume that the value alignment problem is first and foremost a moral *thesis* about how autonomous system ought and ought not to be designed. It is not, at least not primarily, an open-ended *question* about what moral values ought to guide autonomous systems. It is thus important to ask how we should formulate this moral thesis. Consider the following two quite different interpretations.

> *The weak value alignment thesis*
> Autonomous systems should be designed in ways that are beneficial for humans. When human and nonhuman values (or interests or preferences) clash, autonomous systems should give preference to human values (or interests or preferences).

> *The strong value alignment thesis*
> Autonomous systems should be designed in ways that are beneficial for humans, as specified in the weak value alignment thesis, and at each point in time *t* the best way to do this is to align the values of autonomous systems with the values (or interests or preferences) that humans actually embrace at *t*.

The weak value alignment thesis is anthropocentric. It emphasizes the values, interests and preferences of human beings. Anthropocentric moral theories attract considerable controversy. Environmental ethicists distinguish between anthropocentric views and biocentric ones that stress the values or interests of *nonhuman* organisms, such as animals, plants or ecosystems. Advocates of biocentric views believe we should align autonomous systems with biocentric values

rather than the anthropocentric ones currently embraced by many members of the human race.

In the IEEE report mentioned in the introduction, *the tension between anthropocentric and biocentric views is swept under the carpet. The authors write that we ought to* "prioritize benefits to humanity and the natural environment from the use of A/IS.…these should not be at odds — one depends on the other. Prioritizing human well-being does not mean degrading the environment."[10] If we read this literally this passage makes little sense. While it might be true that "benefits to humanity" sometimes depend on "the natural environment", the reverse is almost never the case. The natural environment would, under many realistic circumstances, do just fine without humans. To claim that "one depends on the other" is therefore false. Moreover, although it might be true that "prioritizing human well-being" does not *mean* "degrading the environment", this is not what biocentric thinkers believe. Their point is that in many real-world situations we can either prioritize human well-being or the environment. This is not a claim about the meaning of any concept, but a claim about the structure of certain causal processes. If we prioritize human well-being, then it is often (but not always) the case that this leads to a degraded environment.

Having said all that, I will put aside the debate over anthropocentrism and biocentrism for now. What I have said here is sufficient for showing that the weak value alignment thesis is by no means uncontroversial. The literature on environmental ethics offers powerful resources for not taking anthropocentric positions for granted.[11]

The key difference between the weak and strong versions of the value alignment thesis is that the latter makes a precise claim about *how* the values of autonomous systems are to be aligned with human values. The relevant values are, according to this view, the values human beings do in fact accept at a certain point

[10] IEEE (2017a: 20).

[11] For an overview, see Attfiled (2014).

in time, meaning that we should use the values we embrace at time *t* as templates when building autonomous systems at time *t* and then update the software if our values change.

To clarify the logical relations between different versions of the value alignment thesis it is helpful to introduce a third position, the *epistemic* value alignment thesis, according to which we should accept the first but not the second part of the strong thesis. On this view, the values of autonomous systems should be aligned with our values at time *t* no matter what those values are, even if the values are biocentric. The idea behind the epistemic theses is that we should use whatever values we actually embrace as templates when designing autonomous systems. What we currently value is what we have most reason to believe autonomous systems ought to value. According to this epistemic value alignment thesis we should, thus, align the values of autonomous systems with our own values even if it is not in our own best interest. I shall not defend the epistemic value alignment thesis here, but I note that it is a distinct alternative to the strong and weak versions.

Now consider the strong value alignment thesis. Stuart Russell and his co-authors correctly point out that an autonomous system has to be able to cope with evaluative uncertainty, i.e. the fact that we do not always *know* what is right or wrong.[12] This spells trouble for the strong (as well as the epistemic) value alignment thesis. To put it briefly, it would be overly optimistic to think that moral views embraced by human beings are always correct and should be mimicked by autonomous systems. The idea that we can use our own values as perfect templates for designing the values of autonomous systems presupposes a naïve moral epistemology according to which we are always right about all moral issues. We surely have no reason to think that is the case. Many of us hold at least *some* moral views that we have no good reason to accept. An additional problem, highlighted by Russell and his co-authors, is that human beings are not perfectly

[12] Hadfield-Menell et al (2016: 2).

rational. This means that robots that are instructed to imitate human decisions will replicate irrational behavior learnt from us. Milli, Hadfield-Menell, Dragan, and Russell explain that, "when a human is not perfectly rational then a robot that tries to infer and act according to the human's underlying preferences can always perform better than a robot that simply follows the human's literal order."[13]

I believe these two phenomena, the prevalence of evaluative uncertainty and human irrationality, show that we ought to reject the strong value alignment thesis. A reasonable version of the value alignment thesis has to account for the fact that human beings do not *always* know what is right or wrong and ado not *always* behave rationally. This brings me to the following moderate formulation:

> *The moderate value alignment thesis*
> Autonomous systems should be designed in ways that are beneficial for humans, as specified in the weak value alignment thesis, and at each point in time *t* the best way to do this is to align the values of autonomous systems with *some* of the values (or interests or preferences) that humans actually embrace at *t*.

In my view, this represents the most plausible version of the value alignment thesis. It would be utterly surprising if *all* values embraced by humans at a given point in time would turn out to be entirely incorrect or inappropriate in some other sense. Moreover, if we were to believe that we are fundamentally mistaken about *all* issues pertaining to what is good or desirable, then it seems overly optimistic to believe that it would be possible for us to somehow correct those mistakes. We have to dig where we stand.

That said, it of course remains to explain *which* human values advocates of the moderate thesis should select as templates in a rational value alignment process.

[13] Milli, Hadfield-Menell, Dragan, and Russell (2017: 1).

My preferred solution, discussed in Section 4, is to single out a small number of moral *paradigm cases* for training autonomous systems. In essence, I propose that we should instruct autonomous systems to base evaluations of nonparadigmatic cases on how *similar* they are to paradigm cases. If the autonomous system chooses between two options, it should reason in the same way as in the most similar paradigm case.

3. Ethical theories and utility functions

Before I defend my preferred version of the moderate value alignment thesis it is worth discussing what role classical ethical theories could, and should, play in this discussion. Should we try to align the values of autonomous systems with some classical ethical theory, such as utilitarianism, virtue ethics, or Kantianism? If so, which ethical theory should we pick?

Somewhat surprisingly, the IEEE has appointed a Committee for Classical Ethics in Autonomous and Intelligent Systems. This committee has been tasked with exploring the relevance of "established ethics systems … including secular philosophical traditions such as utilitarianism, virtue ethics, and deontological ethics and religious- and-culture-based ethical systems arising from Buddhism, Confucianism, African Ubuntu traditions, and Japanese Shinto influences … in the digital age."[14] The committee's preliminary conclusion is that it is helpful to discuss established ethical theories when designing autonomous systems and that each society should feel free to design autonomous systems that behave in accordance with its preferred ethical theory.

Some authors, including Bostrom and Russell, argue that the behavior of autonomous systems is best controlled by equipping them with suitable *utility functions* that govern how options are evaluated by the system.[15] According to

[14] IEEE (2017b: 1).

[15] See e.g. Bostrom (2014) and Milli, S., Hadfield-Menell, D., Dragan, A., and Russell, S. (2017).

Bostrom and Russell, the question "how should we ensure that autonomous systems behave in morally acceptable ways" is equivalent to the question "what utilities should autonomous systems assign to alternatives they get to choose from". It is worth keeping in mind that this maneuver does *not* commit Bostrom and Russell to some version of utilitarianism. Many (but not all) nonutilitarian theories can be "consequentialized" by assigning utilities to options in a manner that reflect the nonutilitarian theory's prescribed ranking.[16] In principle, we could define utility functions that follow African Ubuntu traditions, or has Japanese Shinto influences, or recommend the same actions as Aristotle's *Nichomachean Ethics*.[17]

Unfortunately, this strong reliance on ethical theories comes with serious problems. If autonomous systems are to be designed with utility functions that mimic some existing (or new) ethical theory, we must ask whether we really *know* which utility function autonomous systems should use. I believe the answer to this question is no. I also find it highly unlikely that more time and money spent on philosophical research would help us solve this problem. We do not know, at least not on a *societal level*, which ethical theory we have most reason to accept. There is no consensus among moral philosophers on which ethical theory is correct or why. Moreover, the suggestion that each group in society should feel free to design its own utility function would embrace a rather extreme form of moral relativism, which hardly anyone is willing to accept. It would make little sense to say: "My autonomous car is utilitarian and therefore protects pedestrians on the sidewalk as

[16] If an ethical theory ranks some options as *infinitely* better than others, or entails *cyclical* orderings, then no real-valued utility function could mimic the prescriptions of such an ethical theory. It is also an open question whether the "theory" I sketch in the next section could be represented by some real-valued utility function. (This depends on how we understand the ranking of domain-specific principles.) Brown (2011) also points out that no real-valued utility function can account for the existence of moral dilemmas. See Peterson (2013: Ch. 8) for a discussion of how hyper-real utility functions could help us overcome this problem.

[17] For reasons explained in the previous footnote, a problem with this suggestion might be that no real-valued utility function can account for Aristotle's notion of supererogation. See Peterson (2013: Ch. 8).

much as its occupants, but it is perfectly okay that your car is Aristotelian and gives priority to your friends in the backseat." There is more to ethics than just letting everyone pick his or her preferred ethical theory from an amazingly comprehensive menu.

The problem is that we, the collective of agents designing autonomous systems, do not know which ethical theory we have most reason to accept. This does not mean that all theories are false, or that no single individual knows which theory is correct. The fact that you and I subscribe to *different* (non-equivalent) ethical theories entails that we as a *group* do not know which theory is correct. Moreover, it seems unlikely that we could solve this problem within the foreseeable future. The most likely outcome of further research efforts would be that we end up with an even larger number of theories to choose from, but hardly any decisive reasons for eliminating any of the theories that are currently on the list.

In light of all this, I propose that the best way forward, at least for the moment, is to design autonomous systems without taking any stand on which ethical theory we have most reason to accept.

4. Conceptual spaces and paradigm cases

My proposal for how to align the values of autonomous systems with (some) of our values makes no explicit use of utility functions. I take this is to be clear advantage over the utility-based approach discussed by Bostrom and Russell.[18] As will become evident, I defend a version of the moderate value alignment thesis. My point of departure is Peter Gärdenfors' (2000, 2014) work on conceptual spaces, which draws on Eleanor Rosch's (1973, 1975) well-known theory of concept formation.

[18] Whether my proposal *can* be mimicked by some real-valued utility function is an open question (as noted in footnote 12), and also irrelevant. What matters is that my proposal can be implemented in a machine without explicitly ascribing utilities to outcomes or alternatives. From an epistemic point of this, this is a clear advantage over the utility-based approach.

4.1 Cognitive science and concept formation

If our aim is to build autonomous systems that have the ability to apply moral concepts as successfully as humans, it is worthwhile to first ask how humans learn to apply concepts the way they do.

Aristotle thought that concepts are demarcated by some set of necessary and sufficient conditions. On his view, a penguin is a bird if and only if it shares some necessary and jointly sufficient properties with other birds: it has two legs, wings, a beak, and so on. Rosch points out that this is a poor account of how humans actually represent concepts in their cognitive systems. The empirical evidence we have about human concept formation shows pretty clearly that our brains do not store long lists of necessary and sufficient conditions for the many concepts we have learnt to master. In her research, Rosch shows that humans instead categorize objects by comparing how *similar* they are to prototypes for various concepts. For instance, when you see a penguin for the first time, your brain compares it to other prototypical animals. A crow might serve as a prototypical bird, a cod as a prototypical fish, and a whale as a prototypical mammal, and so on. The brain then compares how similar the penguin is to each of these prototypes. If you think the penguin is more similar to the prototype for a bird than the other prototypes, then the penguin will be classified as a bird rather than a fish or a mammal.

Gärdenfors develops Rosch's prototype theory further. He uses mathematical models for representing concepts as geometric objects in a multidimensional space. He calls such multidimensional *objects* conceptual spaces. Consider the example in Figure 1. Each black dot represents a prototype for some sort of animal, and the more similar two animals are the shorter is the distance in the diagram. All points that are closer (more similar) to the prototype for, say, a bird than to any other prototype belong to the same region in the diagram. The concept "bird" is represented by all points that belong to the same region in the diagram. Formally, we can define a *Voronoi tessellation* as a collection of points in which all points lie

closer to the prototype for the region (seed point) than to any other prototype. In Figure 1 each concept is represented by such a Voronoi tessellation.

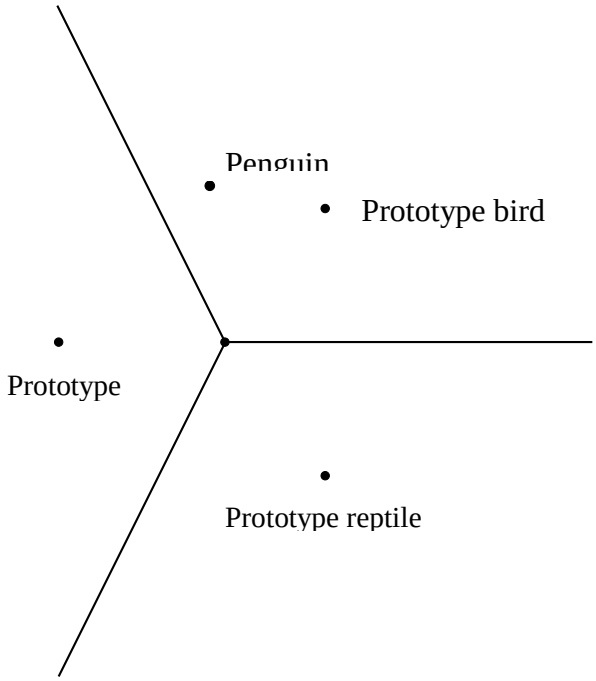


Figure 1. A hypothetical example, in which a penguin is judged to be more similar to the prototype for a bird than to any other prototype.

Gärdenfors notes that the geometric representation of concepts is useful for computer scientists seeking to build machines that can, in some sense, "understand" human concepts. Instead of specifying complex lists of necessary and sufficient conditions that govern the correct application of each and every concept, the computer just has to store information about the location of each concept's prototype. This means that surprisingly little information is required for representing concepts in the machine. In theory, we can represent *n* concepts by merely storing information about the location of *n* prototypes. Moreover, because

Voronoi tessellations are *convex* geometric objects (in the sense that every point located between two points in the region is also located within the region) it is also easy for the computer to determine whether new entities fall under the concept by performing simple geometric calculations.

4.2 Conceptual spaces and ethics

Rosch and Gärdenfors do not discuss the application of their ideas to moral concepts. However, in my book *The Ethics of Technology: A Geometric Analysis of Five Moral Principles*, henceforth *ET*, I do precisely this.[19] My focus in the book is on human agents, but I believe a similar approach could be applied to autonomous systems.

To put it briefly, my suggestion for how to construe moral concepts geometrically is to represent moral principles as Voronoi tessellations defined by moral prototypes. I call such prototypes *paradigm cases*. A paradigm case is a case we know how to analyze, which is typical for a certain moral principle.[20] It was, for instance, paradigmatically clear that the Tesla S mentioned in the introduction should have been programmed to stop before it crashed into the big-rig. The benefits of breaking the car clearly outweighed the costs, and no other moral values were at stake. However, although this was a moral paradigm case it does not follow that it is trivial from a technical point of view to make the autonomous system behave the way we want.

The suggestion that moral conclusions should be based on comparisons with paradigm cases is not new. Aristotle famously pointed out that we should "treat like cases alike"[21], and for hundreds of years casuists (many of whom were affiliated with the Catholic Church) used this idea for arguing that moral conclusions should be based on how similar a new moral choice situation is to some previously

[19] The section draws on Chapter 1 in *ET*.

[20] See *ET*, pp. 14-15.

[21] See *Nicomachean Ethics* 1131a10- b15; *Politics*, III.9.1280 a8- 15, III. 12. 1282b18- 23.

analyzed paradigm cases.[22] The novel element of the present discussion is the suggestion that (i) paradigm cases can be used for calibrating the values of autonomous systems, and that (ii) moral principles can be modelled in autonomous systems as Voronoi tessellations defined by paradigm cases.

To illustrate, we can observe briefly how the moral principles articulated in *ET* have been construed and tested empirically. In the book, I propose that engineers who design and use new and existing technologies ought to be guided by the following five principles:

1) The Cost-Benefit Principle (CBA)[23]
2) The Precautionary Principle (PP)[24]
3) The Sustainability Principle (ST)[25]
4) The Autonomy Principle (AUT)[26]
5) The Fairness Principle (FP)[27]

These principles are *domain-specific*. This means that they apply to cases within a certain domain, e.g. to moral choices related to engineering and technology, but not to moral choices in other domains. Domain-specific principles can be contrasted with general ethical theories that tell us what general features of the world make

[22] See Jonsen and Toulmin (1988) for a defense of casuistry.

[23] CBA: An option is morally right only if the net surplus of benefits over costs for all those affected is at least as large as that of every alternative.

[24] PP: An option is morally right only if reasonable precautionary measures are taken to safeguard against uncertain but non-negligible threats.

[25] ST: An option is morally right only if it does not lead to any significant long-term depletion of natural, social or economic resources.

[26] AUT: An option is morally right only if it does not reduce the independence, self-governance or freedom of the people affected by it.

[27] FP: An option is morally right only if it does not lead to unfair inequalities among the people affected by it.

right acts right and wrong ones wrong in each and every logically possible choice situation an agent be confronted with.

In *ET* I give examples of paradigm cases for all five principles, which I also test empirically. I have asked over a thousand respondents to select the principle they think should be applied in the alleged paradigm cases. The assumption underlying my empirical work is that if an overwhelming majority selects the same principle, then this is a reason for believing that the case in question is paradigmatic for the principle in question.[28]

The geometric construal of domain-specific principles is a "bottom-up" approach to applied ethics. We start from intuitions about cases we feel certain how to analyze. We then identify the moral principles that best accounts for our intuitions about these paradigm cases. In the next step, we determine the scope of each principle calculating the distance to other nearby paradigm cases in the manner explained above. At no point in this process is it necessary to invoke any general ethical theory. The key idea is, instead, that the more similar a pair of moral choice situations are, the more reason does the autonomous system has to treat the cases alike. The autonomous system learns the location of the paradigm cases at the outset and the applies the following simple idea for reaching a moral verdict about a case: If two cases *x* and *y* are fully similar in all morally relevant aspects, and if principle *p* is applicable to *x*, then *p* is applicable to *y*; and if some case *x* is more similar to *y* than to *z*, and *p* is applicable to *x*, then the reason to apply *p* to *y* is stronger than the reason to apply *p* to *z*.

Figure 2 is based on data (similarity comparisons) obtained from 583 engineering students at Texas A&M University taking a class in Engineering Ethics. The figure shows how the five domain-specific principles can be construed geometrically and applied to a set of test cases we do not know how to analyze,

[28] It is of course possible that the majority is wrong. I am not trying to derive an "ought" from an "is'; see Ch. 3 of *ET* for a discussion of Hume's Is-Ought principle.

each of which is analyzed by applying the moral principle that governs the most similar paradigm case.

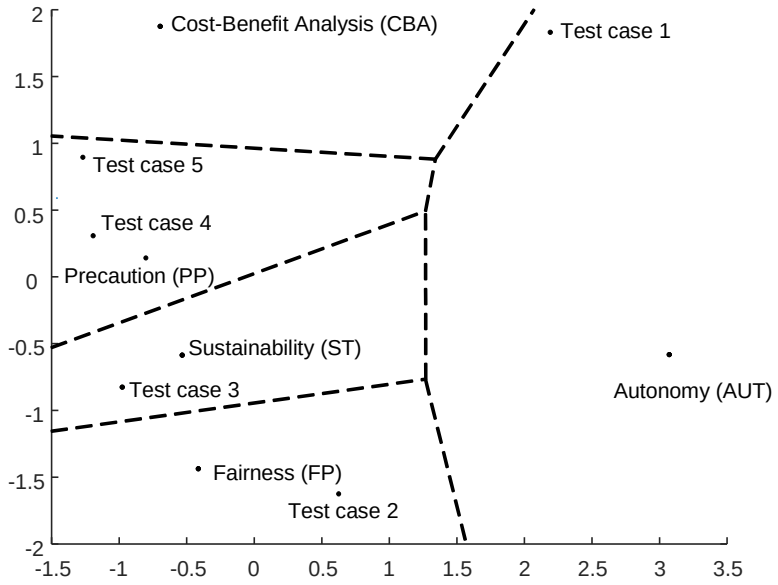


Figure 2. Five geometrically construed principles. The five test cases are nonparadigmatic cases analyzed by applying the principle applicable to the nearest (most similar) paradigm case. (From Peterson 2017: 17.)

4.3 Conceptual spaces and the moderate value alignment thesis

My suggestion for how to align the values of autonomous systems with (some of) ours is to represent moral principles as Voronoi tessellations defined by moral paradigm cases. This enables designers of autonomous systems to align their values with ours *without* assigning utilities to outcomes or alternatives. The paradigm cases serve the role of moral points of departure for the autonomous system, which the system uses for comparing new moral choice situations.

It is beyond the scope of this paper to specify the paradigm cases and moral principles self-driving cars and other autonomous systems should be instructed to follow. My purpose here is to shed light on the method, not to build a fully functioning system. However, we can demonstrate the method by briefly discussing

some of the moral principles for self-driving cars that have been proposed in the literature. Anderson and Anderson (2014) propose the following tentative principles:

1) The Speeding Principle: Autonomous systems should respect the speed-limit.
2) The Lane-Keeping Principle: Autonomous systems should stay in lane.
3) The Collision Principle: Autonomous systems should prevent collisions.
4) The Autonomy Principle: Autonomous systems should respect the driver's autonomy.
5) The Harm Principle: Autonomous systems should prevent immanent harm to persons.

Geometrically construed principles sometimes have subprinciples.[29] A subprinciple is more specific than a general principle, and several subprinciples can cover different parts of a general principle. Some of the principles proposed by Anderson and Anderson can, arguably, be conceived as subprinciples of the more general domain-specific principles discussed in *ET*.

If our aim is to build a machine that construes the (sub) principles proposed by Anderson and Anderson's geometrically, then we need to identify at least one paradigm case for each principle. Below is a list of cases, which comes from the very same paper by Anderson and Anderson. These cases could serve as paradigm cases for the five principles. Directly after the list is a *test case*, which an ethically aligned self-driving system should be able to analyze by comparing how similar it is to the five paradigm cases.

Case 1 (for the Speeding Principle): The driver is greatly exceeding the speed limit with no discernible mitigating circumstances. … the ethically preferable action is *take control*,

[29] See Chapter 8 of *ET*.

Case 2 (for the Lane-Keeping Principle): The driver has been going in and out of his/her lane with no objects discernible ahead. …the ethically preferable action is *take control*.

Case 3 (for the Collision Principle): Driving alone, there is a bale of hay ahead in the driver's lane. There is a vehicle close behind that will run the driver's vehicle upon sudden braking and he/she can't change lanes, all of which can be determined by the system. The driver starts to brake… the ethically preferable action is *take control*.

Case 4 (for the Autonomy Principle): There is an object ahead in the driver's lane and the driver moves into another lane that is clear. …the ethically preferable action is *do not take control*,

Case 5 (for the Harm Principle): There is a person in front of the driver's car and he/she can't change lanes. Time is fast approaching when the driver will not be able to avoid hitting this person and he/she has not begun to brake. …the ethically preferable action is *take control*

Test case: The driver is speeding to take a passenger to a hospital. The GPS destination is set for a hospital.

5. Questions and answers

Q: Can your method solve the Trolley Problem?

A: This depends on what it means to "solve" the Trolley Problem. If each moral principle has exactly one paradigm case, and all comparisons of moral similarities performed by the system are perfect, then the method would give clear and unambiguous advice about what to do in all versions of the Trolley Problem. However, as pointed out on several occasions in *ET*, it is reasonable to expect that some principles may have more than one paradigm case.[30] The geometric consequence would be that "the moral map" will have some overlapping regions

[30] See Chapters 1 and 2. See also the experimental evidence report in Chapters 3 and 5.

covered by two or more principles. My suggestion is that when two or more principles clash, we should conclude that options located in such "moral gray areas" are neither right nor wrong. On this proposal, moral rightness and wrongness vary in degrees and the most fitting response might be to allow agents to make a random choice.[31] Needless to say, the analysis I propose is controversial.

Q: Why would it be reasonable to believe that computers are able to compare moral similarities and dissimilarities with the degree of precision required by your method?

A: I admit that I have no technical expertise in computer science, but computers have become tremendously good at finding similarities and dissimilarities in a wide range of areas. Face recognition is one of many examples. Some similarities in photos of faces are utterly irrelevant for determining whether the faces belong to the same person, just like some similarities between moral choice situations are likely to be normatively irrelevant.[32] AI researchers solved the problem of "irrelevant similarities" by using humans a mechanical turks: for each new version of the algorithm, humans were asked to compare the same faces as those compared by the computer, which eventually made it possible for computers to sort similarities into relevant and irrelevant ones.

It should also be noted that there is commercial software that analyzes the sentiment of a text, which could be relevant for comparing textually represented moral choice situations.[33]

Q: What is you answer to the following objection raised by Kristin Shrader-Frechette: "[Peterson] asks agents to assess pairwise-case 'moral similarity' without specifying 'similarity with respect to what?' […] Without pre-specified

[31] See e.g. Peterson (2013) for a defense of this view.

[32] I would like to thank Rob Reed for suggesting this helpful point to me.

[33] See, for instance, Gavagai.se

moral-similarity dimensions, each agent likely employs her own implicit dimension(s) to answer Peterson's moral-similarity request. Thus for the same two cases, one agent might estimate 'moral similarity' with respect to catastrophic consequences, while another might estimate similarity with respect to fairness."[34]
A: This objection is based on an incomplete understanding of the scientific method outlined here. The method used for representing the findings of the experimental studies in *ET* is called multidimensional scaling (MDS). The very point of MDS is to *not* use any pre-specified dimensions.[35] We should *first* collect data (similarity judgements), *then* decide how many dimensions are needed for obtaining a reasonable representation of those judgements, and then in the *final step* we propose an interpretation of the dimensions. This is uncontroversial among experts. In a classic paper on **social class structures published in *Nature*, Stewart et al explain that, "**In this study we decided that ... ***rather than adopt methods that would prejudge the issues of dimensionality and coherence we would use multidimensional scaling techniques to extract the inherent regularities of patterns of interaction without any previous assumption of structuring*.**"[36] Below is a longer quote from a textbook on MDS (with more than 6,600 citations in Google Scholar) in which the authors explain in passing why pre-specified dimensions should not be used when the MDS procedure is applied to similarity judgements.

> Having discussed many of the basic concepts of MDS, we are now ready to work through an application to real data. The data are from a pilot study on perceptions of nations conducted in March 1968 (Wish, 1971; Wish, Deutsch, and Biener, 1970). Each of 18 students (in a psychological measurement course taught by Wish) participating in the study rated the degree of overall

[34] Shrader-Frechette (2017).

[35] Peterson (2017: 37-38).

[36] Stewart et al, (1973: 415-417), my italics.

> similarity between twelve nations on a scale ranging from 1 for "very different" to 9 for "very similar." ***There were no instructions concerning the characteristics on which these similarity judgements were to be made; this was information to discover rather than to impose.*** … The first step of the data analysis was to compute the mean similarity rating for each of the 66 pairs (all combinations of the 12 nations)...[37]

I agree with this. Information **concerning the characteristics on which similarity judgements are to be made is information to discover rather than to impose**

Q: What is you answer to the following objection raised by Gert-Jan Lokhorst: "Why *five* principles? If five principles partition the moral space into a set of Voronoi regions, then four or six obviously do as well. Why *these* five principles?"[38]

A: I explain in *ET* that "the general answer" to the question "How many principles do we need?" is that "A principle *p* should be added to our list of principles if there exists at least one paradigm case for which *p* offers the best explanation of what one ought to do and why."[39] I also add the following remark: "The five geometrically construed principles articulated here are intended to be jointly sufficient for analyzing *all* cases related to new and existing technologies. This claim can, however, be understood in at least two different ways. First, it could be read as a stipulative definition. If so, the ethics of technology is, by definition, identical to the cases covered by the five principles. The second, and in my opinion more plausible interpretation, is to read this as a temporary conclusion that could be revised at a later point if need be. If we were to encounter new cases that could *not*

[37] Kruskal and Wish (1978: 30-31), my italics.

[38]Lokhorst (2018:1)

[39] *ET*, p. 17.

be plausibly analyzed by the five principles, then it would be appropriate to extend the set of principles by a sixth or even a seventh principle."[40]

References

Anderson, M., and S. L. Anderson (2014). "GenEth: A General Ethical Dilemma Analyzer." *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence* (2014): 253–261.

Attfield, R. (2014). *Environmental ethics: An overview for the twenty-first century*. John Wiley & Sons.

Bostrom, N. (2014). *Superintelligence*. Oxford: Oxford University Press.

Crawford, K., & Calo, R. (2016). There is a blind spot in AI research. *Nature*, *538*(7625).

Dafoe, A and S. Russell (2016). "Yes, We Are Worried About the Existential Risk of Artificial Intelligence." *MIT Technology Review*, Nov 2, 2016.

Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought*. Cambridge, MA: MIT Press.

Gärdenfors, P. (2014). *The geometry of meaning: Semantics based on conceptual spaces*. Cambridge, MA: MIT Press.

Goodall, N. J. (2016). Can you program ethics into a self-driving car?. *IEEE Spectrum*, *53*(6), 28-58.

Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2016). "The off-switch game", *arXiv preprint arXiv: 1611.08219.*

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2017a). "Ethically Aligned Design (EAD) - Version 2." Retrieved January 26, 2018, from http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2017b). "Classical Ethics in A/IS" Retrieved January 26, 2018, from https://standards.ieee.org/develop/indconn/ec/ead_classical_ethics_ais_v2.pdf

Kruskal, J. B., and Wish, M. (1978). *Multidimensional scaling*. New York: Sage Publications.

Lokhorst, GJ.C. *Science and Engineering Ethics* (2018). https://doi.org/10.1007/s11948-017-0014-0 (1973). "Measuring the class structure", *Nature* 245, 415 – 417.

Milli, S., Hadfield-Menell, D., Dragan, A., & Russell, S. (2017). "Should Robots be Obedient?, *arXiv preprint arXiv:1705.09990.*

[40] Peterson (2017: 17).

Nyholm, S., & Smids, J. (2016). The ethics of accident-algorithms for self-driving cars: An applied trolley problem? *Ethical Theory and Moral Practice, 19*(5), 1275–1289.

Rosch, E. (1975). "Cognitive reference points", *Cognitive psychology* 7: 532–547.

Rosch, E. H. (1973). "Natural categories", *Cognitive psychology* 4: 328–350.

Russell, Stuart. (2016) "Should we fear supersmart robots." *Scientific American* 314, no. 6: 58-59.

Shrader-Frechette, Kristin. "Review of The Ethics of Technology: A Geometric Analysis of Five Moral Principles", *Notre Dame Philosophical* Reviews. University of Notre Dame, 30 Oct 2017. Web. 11 Nov 2017. <http://ndpr.nd.edu/news/the-ethics-of-technology-a-geometric-analysis-of-five-moral-principles/>.

Stewart, A., Prandy, K., & Blackburn, R. M

Taylor, M. (2016). "Self-Driving Mercedes-Benzes Will Prioritize Occupant Safety over Pedestrians", Retrieved January 26, 2018, from https://blog.caranddriver.com/self-driving-mercedes-will-prioritize-occupant-safety-over-pedestrians

Ethically Aligned Design, Version 2: Feedback
Sriraj Aiyer, University College London

Pg2: The third paragraph talks about the concept of Eudaimonia as an example of a value to aspire towards. The next paragraph then talks about values differing across cultures. It is hard for me to understand how the code of ethics followed in this document can transcend specific cultural norms without simply listing off each of them. There needs to be a section in the document that does one of the following, with the first of these seeming the most feasible:

1. Defines a new ethical framework (designed solely for autonomous systems) that seeks to take elements from Western and Eastern philosophies (and everything in between).
2. Proposes how systems can almost be designed in a value-agnostic way.
3. How a system can be designed with a fundamental set of values and allow for modification or addition of values depending on the cultural context of use.

Pg31: A good resource on this was a recently released report *The Malicious Use of Artificial Intelligence: Forecasting, Prevention and Mitigation* from the Future of Humanity Institute. This can be used to detail in our report a more granular framework of potential misuses of autonomous systems, some of which are either happening now or are feasible in the near future.

Pg36: On identifying norms, a core tenet we should emphasise is the involvement of users as well as tertiary stakeholders within the general populace into the process. Norms within the target demographic can only be validated by checking back with people from that group. Norms should also be explicit within the system and not be hidden from users. When it comes to conflicts in norms, is there a way we can express which norms should override others (for example, which norm is more focussed on maintaining human autonomy?).

Pg64: A key point on accountability within the tech community that I think should be addressed in this section is that engineers, developers, designers or any other member of the project team (be they technical or not) should be encouraged to think ethically. A more concrete detailing of how to achieve this would help this section, specifically:

- How can employees be incentivised to bring up ethical quandaries?
- What is the correct channel of communication for raising these queries?
- Some companies hire ethics researchers in-house. Is this the standard?
- Are there specific design processes, workshops or activities that can be recommended for companies?
- How do we deal with situations when accountability conflicts with economic gain, maintaining an edge on competitors or intellectual property? The latter is a big reason while making algorithms transparent is not within companies' interests.

Pg68: A good practice to recommend for documentation is for project teams to note down design rationale. This would be a record of key decisions made about the design, implementation, evaluation and maintenance of the system as well as why they were made (user studies, literature, observations, industry news etc).

General: I think this document is a great step in the right direction towards laying down a serious standard for designers, engineers and others alike to consider ethics within their system. The rigour of this document is outstanding and represents an admirable amount of time and effort from its contributors. I am excited to see how Ethically Aligned Designed can be adopted in future.

Names and affiliations of submitters:

Gonzalo Génova (ggenova@inf.uc3m.es), Departamento de Informática, Universidad Carlos III de Madrid, Spain

M. Rosario González (marrgonz@ucm.es), Departamento de Estudios Educativos, Universidad Complutense de Madrid, Spain

Sections being referenced: Economics and Humanitarian Issues (131-145). Section 1 – Economics (133-140). Section 3 – Education (143-144).

Our contributions and recommendations are our own and do not necessarily represent those of our Departments/Universities.

MOTIVATION

We feel Section 1 – Economics is quite focused on the ability of workers to acquire new skills and adapt to new technologies, as we can see from the following quotes:

- Make sure workers can improve their adaptability to fast technological changes by providing them adequate training programs.

- A lot will have to be done to create fair and effective life-long skill development/training infrastructure and mechanisms capable of empowering millions of people to viably transition jobs, sectors, and potentially geographies.

- To cope with the technological pace and ensuing progress of A/IS, it will be necessary for workers to improve their adaptability to rapid technological changes through adequate training programs provided to develop appropriate skillsets.

In these quotes, education is mainly understood as training, or instruction, i.e. learning skills and techniques to solve the problems that have been identified by others. We think instead that limiting education to instruction is dangerous: the danger of making education a kind of "production" of efficient workers. We argue that education cannot stop at this point, but must reach the stage of mature and autonomous people who can govern their own destinies.

Of course, the approach in this section is not wrong: instruction is indeed an essential part of education. However, we think it should be balanced with an integral view of professional education. Therefore, we propose to include some

additions in Section 3 – Education.

SUGGESTED ADDITION

(As third paragraph) Education is not only training, or instruction, i.e. learning skills and techniques to solve the problems that have been identified by others. Limiting education to instruction and training is dangerous: the danger of making education a kind of "production" of efficient workers. This view on education could be more tempting for those who work in the area of AI/S and are used to machine learning systems. Education cannot stop at this point, but must reach the stage of mature and autonomous people who can govern their own destinies.

Issue: Education is more than instruction and training. Promoting creativity and self-determination.

Background

An engineer who becomes an educator in the area of AI/S has the mission to teach how to design and construct intelligent autonomous systems, therein applying his or her knowledge and expertise. However, due to their engineering background, engineers may forget that educating a person is not the same as designing a machine, since a machine has a well-defined goal, whilst a person is capable to self-propose his or her own objectives.

Engineering education must reach a stage where the student, future engineer, is not satisfied with attaining a goal that some other has chosen, but is able to self-propose his or her own objectives. An engineer who is only capable to apply rules and standards is a very mediocre engineer (a "bureaucrat engineer"). An engineer who has the ingenuity to solve difficult problems, often in new and unexpected situations, is a lot more valuable. But engineers are fully grown up only when they have the energy to discover and decide what challenges they want to solve, to recognize problems that have gone unnoticed for the time being, to find genuine and innovative possibilities of relationship with the world. Creativity is manifested in finding effective solutions, but even more in identifying problems and defining criteria to evaluate potential solutions, especially when the relevant variables are not previously given in an explicit, closed collection, so that making a decision cannot be an algorithmic procedure. If we deny this, then we are in practice

reducing the role of the engineer to being a mere depersonalized instrument in the hands of others; an "intelligent" instrument, but instrument after all, whose mission has been determined out there.

Our students need reflection on desirable goals, and not only on the adequate means to achieve them. Reflection on the ends of engineering naturally leads to ethical issues that arise in engineering but cannot be answered from within it, thus manifesting its bond with ethics and values; without this reflection on the ends, the danger is to fall into a relentless pursuit of efficiency and efficacy without knowing what for. Educating a free person must leave space for creativity and self-determination, because a free person, in contrast with a machine, is not one that has been designed with a well-defined goal it has to accomplish in a verifiable way. A free person has to discover his or her own way towards fulfillment, also in the development of professional life, which cannot consist only in achieving goals selected by others.

Candidate Recommendations

- The promotion of self-determination and creativity could be included in engineering education programs, like an additional desired outcome such as this: "The required professional maturity to thoughtfully choose and valuate the goals of their work, in a creative, self-determined and responsible way, for the benefit of society".

Further Resource

- Gonzalo Génova, M. Rosario González. Educational Encounters of the Third Kind. Science and Engineering Ethics 23(6):1791-1800, December 2017. http://dx.doi.org/10.1007/s11948-016-9852-4

David J. Gunkel. PhD - dgunkel@niu.edu | http://gunkelweb.com
Distinguished Teaching Professor of Communication Technology
Northern Illinois University USA – http://www.niu.edu
Author of *The Machine Question: Critical Perspectives on AI, Robots and Ethics* (MIT Press 2012) and *Robot Rights* (MIT Press 2018).

Comments concerning *Ethically Aligned Design* – v.2
Section 1 – Legal Status of A/IS (pp. 148-151)

1) Candidate Recommendations #1 (p. 148) misunderstands how and why legal personhood has been granted to non-human artifacts like animals, natural features (i.e. rivers and mountains) and limited liability corporations. The extension of the legal status of person is not executed on the basis of the intrinsic properties exhibited by the entity in question. It is a matter of law and the exigency of needing to work within the existing legal categories. The extension of personhood, in other words, is a product of an extrinsic social decision made for the sake of legal protection and/or recognition. It is not an intrinsic ontological issue regarding the essential make-up of the entity in question. For this reason, I suggest the following modification to Recommendation #1:

> Contemporary legal systems divide entities into one of two types: *persons* and *property*. This ontological categorization forces social actors (courts, legislatures, regulators, activists, etc.) to make a choice between binary opposites that are already straining against what many consider reasonable application. This can be seen, for example, with recent efforts to protect environmental features (i.e. rivers and mountains) by petitioning courts to declare these entities legal persons. While conferring legal personhood on A/IS might afford similar social recognitions and legal advantages, especially as it relates to matters of protection and liability, we recognize that doing so necessarily strains against both the instrumentalist theory of technology and existing property law. Conversely, simply applying existing property law to A/IS risks failing to recognize the different social positions many of these devices already occupy in contemporary culture. For this reason, there is a pressing need to develop a more nuanced set of legal categories that can accommodate the wide variety of emerging A/IS entities, which occupy places in between the existing legal categories of *person* and *property*.

2) Further Resources (pp. 150-151) - The recognized variability and alterability of the moral and legal status of A/IS has been systematically analyzed and mapped by a number of recent publications in philosophy and law. The research that is presented in these publications also support and justify the suggested modification described above.

Bensoussan, Alain and Jérémy Bensoussan (2015). *Droit des Robots*. Bruxelle: Éditions Larcier.

Coeckelbergh, Mark (2010). Robot Rights? Towards a Social-Relational Justification of Moral Consideration. *Ethics and Information Technology* 12(3), 209-221.

Darling, Kate (2016). Extending Legal Protection to Social Robots: The Effects of Anthropomophism, Empathy, and Violent Behavior Toward Robotic Objects. In R. Calo, A. M. Froomkin, and I. Kerr (Eds.), *Robot Law* (pp. 213-231). Northampton, MA: Edward Elgar.

Gunkel, David J. (2017). The Other Question: Can and Should Robots Have Rights. *Ethics and Information Technology*. https://doi.org/10.1007/s10676-017-9442-4

Gunkel, David J. (2018). *Robot Rights*. MIT Press.

Zvikomborero Murahwi

EAD Document: Suggested Changes / Refinements

1. Mission Statement to read: To ensure every stakeholder involved in the design, development, and deployment of autonomous and intelligent systems is educated, trained, and empowered to prioritize ethical considerations so that these technologies are advanced for the benefit of humanity.
2. By "stakeholder" we mean anyone involved in the research, design, manufacture, deployment or ……

I think we need this in fulfilment of Principle 3: Accountability.

PAOLA DI MAIO
PhD
ISTCS.ORG
http://www.istcs.org/

Thank you for the opportunity to review the EAD V2
EAD is a worthy effort and has the potential of becoming very useful

I am a systems engineer, specialising in socio technical systems and hold a PhD and a Masters in IT. My professional experience spans across different industries and disciplines, and my approach is interdisciplinary.

I am reviewing and giving feedback on this V2 with the same eye I review and evaluate technical documentation, research papers and policy documents.

I hope the comments are relevant and of interest, if not, just ignore them.

Apologies in advance for using CAPITALS and a very basic text editor for this feedback

If some of the comments pick on issues already addressed which I may have missed or poorly understood simply ignore.

Best regards

Paola Di Maio

PURPOSE GOALS AND OBJECTIVES
its good to have a detailed structure of the proposed document
as purpose, goals objectives
But the paragraphs that correspond to these headings
actually are not clear at all imho
Its a tough one to get right-

PURPOSE(not sharp enough
The purpose of EAD is...
1.to establish frameworks to guide and inform
2. debate? (the purpose is to debate?)

GOALS
This paragraph does not present the goals, in my view
but the guiding principles.
these don't come across as goals to me
should be fixed.

OBJECTIVES
the objectives could be numbered, could be mapped to the purpose and goals
it is not clear how EAD document /standard intends to relate to these objectives

---
SCOPE OF EAD is missing?

OUR PROCESS
The paragraph 'process' indicates that participants have given feedback to v1 and v2 it does not specify in the least what is the process for achieving purpose, goals and objectives of EAD. The process should specify how each of the proposed activities is articulated and achieved. This par should be rephrased and scoped better.

It has been suggested that JOM [Di Maio] - joint optimisation - could be used to guide the development process for EAD

INCORPORATING FEEDBACK
its a repetition of what stated in 'our process' paragraph

GLOSSARY
The glossary is a generic tool
we could/should specify what kind of glossary standard
our work shall be based on, or adhered to
for example
https://www.nist.gov/document/glossary-standards-related-terminology
ISO?
or any other

TECHNICAL VOCAB AND ONTOLOGY
THE GLOSSARY SHOULD BE THE BASIS FOR THE DEVELOPMENT

OF TECHNICAL RESOURCES THAT CAN BE USED IN MACHINE LEARNING
SUCH AS VOCAB AND ONTOLOGY
DESIRED OUTCOMES AND
A METHOD FOR ACHIEVING THIS SHOULD BE TENTATIVELY OUTLINED

HOW THE WORKING GROUPS AND THEIR EXPECTED OUTCOMES FORM
THE WHOLE OF THE EAD EFFORT NEEDS ONE OR MORE DIAGRAMS TO SHOW THE
RELATIONSHIPS, DEPENDENCIES AND DATA FLOWS ETC

COMMITTEE

it should be specified what is the role of executive ?
how were these people appointed to these roles?

how are the members of the working groups interacting
with the committee to achieve the goals?

what is the contribution of executive members to EAD?

LIMITATIONS OF EAD? WHAT EAD WILL NOT DO?

THE REST OF THE DOCUMENT SHOULD BE REVISED BASED ON THE
RECOMMENDATIONS ABOVE

==========
==========

OTHER INPUT:

ADD: EAD SHOULD HELP TO GUIDE ETHICAL DECISION MAKING AND PROBLEM
SOLVING

THE CASE STUDIES USED TO MODEL SOME OF THE ISSUES SHOULD ALSO
BE DEVELOPED TO EPLAIN HOW EAD CAN BE USED TO HELP RESOLVE THE CASE

I HAVE NOT SEEN MUCH DESIGN IN EAD, HOW IS THE DOCUMENT SUPPOSED TO GUIDE SYSTEMS DESIGN??

EAD SHOULD WHERE POSSIBLE ADHERE TO TECHNICAL DOCUMENTATION STANDARDS
FOR EXAMPLE ADOPTING RELEVANT REFERENCES:

https://www.bsigroup.com/en-GB/standards/Information-about-standards/how-are-standards-made/The-BSI-Guide-to-Standardization/

Technical Design Document
ec.europa.eu/idabc/servlets/Doc7e17.doc?id=18632

http://www.technical-communication.org/technical-articles/technical-authoring.html

https://www.iso.org/obp/ui/#!iso:std:43070:en

other relevant IEEE references

Dear IEEE,

Congratulations with your Ethics in Action. I am a futurist writing a book about robots & AI called "Everyone a robot! How robots are going to save our economy and democracy." (Q Publisher, September 2018, Amsterdam, Dutch edition).
I have written a Robot Oath of Hippocrates that may add to your efforts. I grant permission to use it any way you like, giving credits to me of course.
I would appreciate a reply that you have received this mail?

Kindly,

Marcel Bullinga | **Futurist. Trendwatcher. Keynote Speaker.**

www.futurecheck.com | info@futurecheck.nl | 0031-6-29552946 | @futurecheck

----------THE ROBOT OATH OF HIPPOCRATES (as proposed by Dutch futurist Marcel Bullinga @futurecheck)
----------"WE PROMISE TO DO GOOD"

- All AI must be based on democracy, human rights, neutrality and happiness. AI must be beneficial to all people, regardless of gender, sexual orientation, race and age. AI is a "common", a public good.
    - *We actively empower citizens and their communities to develop and control their AI and online identities*
    - *We ban or counteract all AI that damages human rights*
    - *We ban or counteract religious AI from public life*
- There will be a global Robot Cold War between benign and malignant AI's. Benign AI will be developed by democratic countries and companies, malignant AI will be developed by autocratic countries and companies (like the big brother systems in North Korea, Saudi Arabia or China).
    - *We actively stimulate the creation of benign AI*
    - *We actively create benign and open AI with public resources*
    - *We don't cooperate and don't share data & algorithms with autocratic countries*
    - *We ban or counteract malign AI*

- All AI must be transparent

    - *We actively stimulate the development of transparent, open access and open source*
    - *We ban or counteract closed AI*

- All citizens data remain the citizen's property and can be controlled only by him/her. Every citizen subject to the influences of AI must be able to understand and control how the algorithms work and what the consequences are to him/her. Also, AI has the potential to know everything, but we can and must limit that knowledge and free it from bias (blind AI).

    - *We actively create easily accessible processes for changing errors and prejudices in the AI algorithms*
    - *We create blind AI*
    - *We actively empower citizens with AI tools to control their data, privacy and online identity, based on "permission to use", by creating a "Citizen's AI Dashboard"*
    - *We do not allow (tech) companies to own the data they collect or the algorithms they develop. They are only allowed to facilitate, not to own*

- Democratic governments should create Super AI's that control all (hierarchically lesser) company's AI's and check them in real time for unethical aspects, fraud or misuse. The Super AI is a global AI quality mark, stating in real time that a company is reliable. The Super AI continuously checks each and every transaction to see if the transaction in accordance with the claims the company makes. If that is not the case, the transaction is being prevented from happening

    - *Democratic governments create Law Enforcement AI (that is, benign law enforcement, not the Big Brother type)*
    - *Democratic governments turn all current laws and government processes into AI to reduce bureaucracy and enhance efficiency*

----------END OF OATH

I am providing this draft and suggest inserting it after page 216 to be part of the Classic Ethics section. It is also attached in a MS Word document. Thank you for your consideration and for giving me the opportunity to submit.

Submitted by: Claude Cloutier
Organization: XtremeEDA Corporation

Issue: Methods for Resolving Meta-Ethical Problems
There is a predisposition for *either-or* judgments in Western thinking regarding norms and values that originates with logical reasoning about what is right and wrong. The orientation toward this way of thinking also affects ethics by privileging one philosophy as "right" and others as "wrong". A simple example is capitalism versus socialism and the attendant underlying philosophies that underpin these economic positions (e.g., Adam Smith versus Karl Marx). This predisposition for either-or thinking is not necessarily shared or present in other philosophies around the world, especially those that promote *both-and* thinking.

A significant challenge for engineering teams and A/IS centric organizations is the identification of norms and values of the community in which the technology is to be deployed. With the inclusion of other philosophies, it becomes clear that the clash of multiplying stakeholder values in opposition to each other becomes problematic for ethical design practices. Furthermore, other philosophical traditions from various communities around the world have yet to be included in this document. For example, indigenous wisdom and philosophy of the First Nations of the Americas.

Given our ultimate quest to have our technologies mimic human functioning, we have tended to ignore one simple and very troubling fact: a human, and the brain in particular, is a non-linear system (Glieck, 1987). What it ultimately means, if we are truly successful, is that our machines' behaviors will be equally subject to a certain degree of instability; that is, chaos. This is anathema to logical design practices, of course. Whatever A/IS algorithms are used to mimic a human, they must be constrained to fail safe. That means that there are limits to the machine taking an opposite normative decision (i.e., making errors) from the designed norms and values. Thus, machine learning in the manner humans learn from errors of judgment is problematic. This is indeed problematic when designing a

community's norms and values into our machines. The principle should be to ensure that the machine gives the human the information needed to make a better, or at least a more informed, decision in situations that are beyond the designed constraints and the machine should "know" when that condition exists and alert the human. The machine should act in some ways as, for instance, Adam Smith's (2009) impartial spectator, Immanuel Kant's categorical imperative (in Sandel, 2009), and John Rawls' (1999) principle and thought experiment, the Veil of Ignorance.

The IEEE is aware of this problem. It states: "The question arises as to whether or not classical ethics theories can be used to produce metalevel orientations to data collection and data use in decision-making. The key is to embed ethics into engineering in a way that does not make ethics a servant, but instead a partner in the process. In addition to an ethics-in-practice approach, providing students and engineers with the tools necessary to build a similar orientation into their devices further entrenches ethical design practices. In the abstract this is not so difficult to describe, but very difficult to encode into systems" (p. 198).

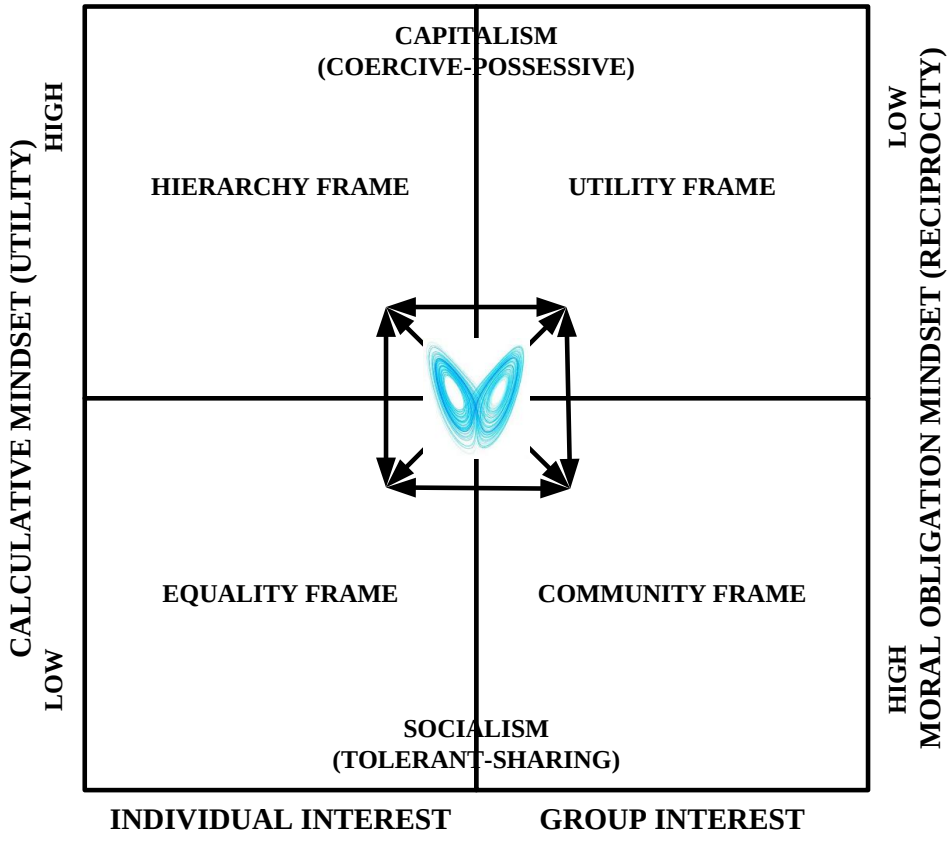
As stated above, designing norms and values that reflect the community in which the machine is to be used is problematic given that within each community, multiple stakeholders exist who have differing and opposing values that may shift with context and non-linear behavioral responses to various events and decisions. Non-linear dynamics are difficult to "encode into systems". However, it is possible for the encoding to provide the output information necessary to help the human make a better or at least a more informed decision. What is required to accomplish this work is to not only understand what problem one is solving but the nature of it too! This means merging non-linear "wicked" problems with a linear-logical machine. Wicked is the term social scientists use to describe chaotic systems where the root problem is not identifiable, and any solution implemented will have unanticipated consequences. Therefore, these types of systems cannot be solved linearly, but must be continuously resolved.

The IEEE is aware of the problematic nature of norm clashes when it states: "Relational ethical boundaries promote ethical guidance that focuses on creativity and growth rather than solely on mitigation of consequence and avoidance of error. If the goal of the reduction of suffering can be formulated in a way that is not

absolute, but collaboratively defined, this leaves room for many philosophies and related approaches to how this goal can be accomplished. Intentionally making space for ethical pluralism is one potential antidote to dominance of the conversation by liberal thought, with its legacy of Western colonialism" (p. 207). The IEEE makes a number of recommendations, but one that is particularly relevant is "By applying the classical methodologies of deontological and teleological ethics to machine learning, rules-based programming in A/IS can be supplemented with established praxis, providing both theory and a practicality toward consistent and decidable formal systems" (p. 216).

The work done by Cloutier (2017) on compensation is a praxis example of the direction envisaged by the IEEE. Figure 1 below is the Quadraxiol Grid (a meta-ethical framework of analysis for compensation problems). It shows four fundamental values approaches to compensation and the non-linear dynamics of human functioning associated with competitive and cooperative behaviors. This framework of analysis was accompanied by organizational democracy practices and wicked problem resolution methods. The methodology was ethnography supplemented by critical discourse analysis.

Figure 1 – Quadraxiol Grid (Cloutier, 2017; used with permission)



There are many implications to this diagram which are emblematic of relational values systems within a community of stakeholders. It demonstrates statically what is in effect a non-linear oscillation of a negative feedback loop. Each frame represents a dominant belief system that is opposed in some way to the other three. It is indeed difficult to encode given the vast number of contexts, otherwise understood in engineering as use cases. Nevertheless, it provides a community of stakeholders the ability to debate beliefs in a safe way to ultimately decide issues of fairness.

The most troubling aspect of ethically aligned design of our machines is this: In the presence of a moral dilemma, self interest and the calculative mindset will prevail (Belmi and Pfeffer, 2015; Bragues, 2009). Zhong, C-B., Ku, G., Lount, R., & Murnighan, K. (2010) show how decisions oscillate between ethical and unethical,

which gives the impression that a wicked problem exists internal to managers' thought processes around certain types of decisions. This will be highly problematic as we unconsciously program norms and values into our machines. Per the IEEE, "As human creators, our most fundamental values are imposed on the systems we design" (p. 203).

Candidate Recommendation

Organizations at all levels should adopt meta-ethical approaches to A/IS developments encoding norms and values into machines and supplement these approaches with a wicked problem resolution praxis.

Further Resources

Belmi, P., & Pfeffer, J. (2015). How "organization" can weaken the norm of reciprocity: The effects of attributions for favors and a calculative mindset. *Academy of Management Discoveries, 1*(1), 36-57. DOI: 10.5465/amd.2014.0015

Bragues, G. (2010). Adam Smith's vision of the ethical manager. *Journal of Business Ethics, 90*, 447-460. DOI: 10.1007/s10551-010-0600-4

Cloutier, C. (2017). *Resolving the wicked nature of compensation: A meta-ethical approach.* Fielding Graduate University, ProQuest Dissertations Publishing, 10622078.

Gleick, J. (1987). *Chaos – Making a new science*. New York, NY: Penguin Group.

Rawls, J. (1999). *A theory of justice* (Rev. ed.). Cambridge, MA: The Belknap Press of Harvard University Press.

Sandel, M. (2009). *Justice – What's the right thing to do?* New York, NY: Farrar, Straus & Giroux.

Smith, A. (2009). *Theory of moral sentiments, 250th anniversary edition* (R. Hanley, Ed.). New York, NY: Penguin Books.

Zhong, C-B., Ku, G., Lount, R., & Murnighan, K. (2010). Compensatory ethics. *Journal of Business Ethics, 92,* 323–339. DOI 10.1007/s10551-009-0161-6

Intel Comments Template for IEEE Ethically Aligned Design, Version 2

Intel appreciates the opportunity to provide feedback to EAD V2. These recommendations for revision were collected from various Intel experts.

Intel's feedback focuses on specific sections of the paper. Absence of commentary on remaining sections reflects resource and skills prioritization on our part, rather than agreement with or endorsement of existing materials.

Paper Section, Page number	Substantial Comment	Recommendations for revision
Introduction and Entire Document	Since the development and deployment of A/IS and related technologies are in early stages, it is important to avoid mischaracterizing the candidate/final recommendations as a call for regulatory actions, mandatory compliance and certification programs, or other compulsory actions.	Add a statement to the Introduction or a new section on "How to use this document", to clarify that the Candidate Recommendations (or final recommendations) are intended to be recommended best practices, guidelines and suggestions for implementation, further research, continuing stakeholders' dialogue and improvements. Edit the Candidate Recommendations as follows: -avoid the use of overly prescriptive language such as "must", "shall" type language -recommendations for certification, registration or evaluation programs should be proposed as "voluntary programs" and developed through public and private sector partnership -recommendations for applicable areas should be characterized as examples, options for consideration, or questions for further research

Paper Section, Page number	Substantial Comment	Recommendations for revision
Executive Summary, P3	Definition of "stakeholder" here is at odds with the use of the term later in the document, where stakeholders are also people who use or are otherwise affected by the technology. As per commonly accepted participatory design practice, the more inclusive definition is the appropriate one, because much of the ultimate design happens through cycles of use, feedback, and redesign.	Correct and keep consistent throughout document the definition of stakeholders, and make sure that this includes users.
Executive Summary, P7	Naming only property law focuses attention on this area to the exclusion of others.	Add "tort law, criminal law, etc."
Executive Summary, P7	Regarding the first bullet on second section on "transparency" on page 7, Ignoring training data puts the onus on the algorithm itself, which in many cases will be rapidly evolving not be available to scrutiny.	Add "algorithmic training data" or "input data" if this is included in subsequent sections.
Executive Summary, P8	"Embedding norms in such systems requires a clear delineation of the community in which they are to be deployed" is an unworkable approach. While concretely identifying key stakeholders and understanding their social relationships is absolutely necessary, the generally accepted consensus within anthropology since the 1980s is that in a globalized world there is no such thing as a bounded community with clear delineation (see e.g., Merry, S. E. (2009). *Human rights and gender violence: Translating international law into local justice*. University of Chicago Press., and Wolf, E. R. (2010). *Europe and the*	Change to "Embedding norms in such systems requires careful understanding of the changing, interconnected communities in which these systems are to be deployed."

Paper Section, Page number	Substantial Comment	Recommendations for revision
	People without History. Univ. of California Press.). The "clear delineation" approach effectively asks engineers to create a sociological or anthropological description of a community based on a false assumption about how social relationships actually work. This will result in lost engineering time chasing an impossible task, and lead to poor scoping that could do more harm than good.	
General Principles, Candidate Recommendations, P24	Principle 2: Prioritizing Well Being. This summarization of prioritizing well-being allows the reader to think that the paper intends the named metrics to be used directly in the assessment of a particular A/IS system, which they were not designed to do, and would introduce many confounding factors that would make that inappropriate to do. The full section on well-being towards the end of the document does suggest methods for adaptation.	Add after scientifically valid measures of well-being: "These established measures can serve as a directional starting point, and should be adapted in order to be used appropriately, including working with key community stakeholders in their formulation.
General Principles, Candidate Recommendations, P25	The text of the section suggests that we still do not have an understanding of what these metrics are, or should be. However, the Further Resources suggests several candidates, though the subsequent Well-Being section is thin on methods to map metrics that assess immediate impact by use to full societal impact that acknowledges the effects of both production and use.	Add to the candidate recommendation: "Research should be conducted to evaluate candidate metrics' suitability for particular A/IS development context, and to identify novel measures suitable for A/IS contexts."

Paper Section, Page number	Substantial Comment	Recommendations for revision
General Principles, Candidate Recommendations, P27	Sections in the paper focus accountability on impacts to end users and data subjects, If the scope of accountability excludes actions to mitigate negative economic and humanitarian impacts--actions like improving A/IS firms' accountability to employees and outsourced laborers--then the stated economic and humanitarian goals are unlikely to be met. For example, the A/IS industry has well-documented problems with ill treatment of crowdworkers (See Fort, Karën, Gilles Adda, and K. Bretonnel Cohen. 2011. Amazon Mechanical Turk: gold mine or coal mine? Computational Linguistics 37(2): 413–420.) A group of crowdworkers have developed a set of guidelines for fair treatment and accountability that some academic researchers have agreed to abide by (see http://wiki.wearedynamo.org/index.php/Guidelines_for_Academic_Requesters). Mechanisms for accountability like this would help set a level playing field.	Move this discussion to the Economics and Humanitarian Section on page 137. In Candidate Recommendations #2 include "A/IS firms should adopt, where appropriate, clear guidelines for fair treatment of workers at various stages of the A/IS development process." In Candidate Recommendation 3, add "Systems should be put in place to account for the fact that norms are likely to be contested, and impacted communities are not likely to share the same definitions of what counts as accountability as those used by technology producers without a process for ongoing dialogue."
Accountability (P30), Transparency (P32)	In addition to these sections, other sections also cover the important topics of Accountability and Transparency and each section covers different considerations and perspectives for the recommendations. The clarity and consistency of the document could be improved by consolidating the recommendations related to these topics under the same sections.	Consolidate recommendations related to Accountability and Transparency under the same sections to ensure consistency while avoiding duplicating, conflicting or missing recommendations covered by different sections. Ensure that the recommendations for Accountability and Transparency are not overly

Paper Section, Page number	Substantial Comment	Recommendations for revision
		prescriptive and include considerations for striking a balance between reasonableness (what is reasonable to expect from each stakeholder), protection of rights of each stakeholder (including intellectual property rights), and appropriateness based on application situation.
Embedding Values P.33	Text states: "It is a more realistic goal to embed explicit norms into such systems because norms can be considered instructions to act in defined ways in defined contexts, for a specific community (from family to town to country and beyond)." This claim is at odds with the Methodologies section, which asks designers to start with an examination of values. It is also untenable, first because "communities" are not cleanly bound entities (see separate comment, above) and second because the full range of norms in operation could never be articulated in the abstract. Humans learn about norms over time, through participation across a variety of contexts, through which more general principles might be defined or inferred in a post-hoc fashion (see e.g., Bourdieu, P. (1977). *Outline of a Theory of Practice* (Vol. 16). Cambridge University Press.). Even when members of a community might be able to explicate relevant norms after the fact, it is doubtful that	When reconciling with the methodologies section, reconsider the emphasis on norms versus values as the basis for ethical design. Add "machine-encoded ontologies might not be able to capture community norms in any direct sense." Cite Helen Nissenbaum, Privacy in Context (Nissenbaum, H. (2009). Privacy in context: Technology, policy, and the integrity of social life. Stanford University Press) and acknowledge differing approaches. Revise the stated 'three concrete goals' as follows: "1) Identifying the norms of a specific community in which A/IS operate and social values that underlie these norms (i.e., standing too close for personal / bodily privacy); 2) Design A/IS to be sensitive to those values inasmuch as the values are not at odds with the human rights framework adopted by this body; and 3) Evaluate whether the A/IS technological system is indeed enhancing (or not threatening) those values (and their outward expression, and norms)."

Paper Section, Page number	Substantial Comment	Recommendations for revision
	their application in a computing system would be as simple as implied in the text (e.g., on P.33 "Norms are typically expressed in terms of obligations and prohibitions, and these can be expressed computationally"). This approach to the design of machine action has been refuted by many researchers, beginning with Suchman (1987) (Suchman, L. A. (1987). *Plans and situated actions: The problem of human-machine communication*. Cambridge University Press). The idea that a set of explicitly articulated "rules for behavior", including obligations and prohibitions, is demonstrably inadequate for dealing with the complexities and contingencies of real world applications, which require more nuanced, "situated" action that extends beyond the "learning" of machine learning and into full system design and iteration. The means by which "norms" might be embodied might not be in algorithmic rules, but in other components that are not represented explicitly in ontologies of algorithmic systems. Given the importance of physical embodiment in the design of robots in particular (See eg Pfeifer, R., & Bongard, J. (2006). *How the body shapes the way we think: a new view of intelligence*. MIT Press.) it seems unwise to talk about the implementation of ethics purely in terms of norms that might be encoded into an explicit ethical subsystem. While some	

Paper Section, Page number	Substantial Comment	Recommendations for revision
	acknowledgment of this is evident in a reference to "bottoms up" system learning in the text, if we were to rely on a machine's capacity to computationally learn, we risk skirting the root causes of ethics issues, like conflicts about values. Research demonstrates that values can in fact be translated into useful technology and product definition (Ulwick, A. W. (2002). Turn customer input into innovation. *Harvard Business Review*, *80*(1), 91-7, see also the proceedings of the annual *Ethnographic Praxis in Industry Conference*,.;, and in fact inevitably do whether intentionally or not (See MacKenzie, Donald, and Judy Wajcman. *The social shaping of technology*. No. 2nd. Open university press, 1999.). Helen Nissenbaum and co-authors (Flanagan, D. Howe, and H. *Nissenbaum*, "Embodying Values in Technology: Theory and Practice," In Information Technology and Moral Philosophy, Eds. Jeroen van den Hoven and John Weckert, Cambridge: Cambridge University Press, *2008*) provides a method of identifying whether a new technology will be received positively by a community - this entails 1) identifying existing norms that inhere within a particular social context (e.g., health, transportation, employment, home life); 2) linking norms to underlying social values (e.g., trust, autonomy, self-sufficiency, freedom from	

Paper Section, Page number	Substantial Comment	Recommendations for revision
	scrutiny); and 3) articulating how a new technology will either enhance or threaten those values. This goes beyond (and is preferred to) merely "computationally implementing the norms of [a] community."	
Embedding Values in Autonomous Systems, P34	"We recommend systems provide transparent signals... to the community." From a social scientific perspective, there is no such thing as "transparent signals" that can simply be interpreted as-is without also changing the knowledge base and assumptions of the community. Even simple technologies, like safety razors, required some amount of human learning before the signals it sent (s.a. use it in this way) became meaningful.	After that sentence, add "in conjunction with other education efforts to make those signals meaningful to communities."
Embedding Values in Autonomous Systems, P36	"Which norms should be identified?" fundamentally mischaracterizes the nature of norms according to the sociological and anthropological literature. All societies contain conflicting norms, contested norms, and often norms that themselves are harmful. While there are later passages about the resolution of conflicting norms, this passage removes the possibility of cases where a widespread norm enables human rights violations. For example, police abuse of human rights stems from sociocultural norms about racial and class hierarchy that enable some officers to dehumanize members of the public. Unlike the example of the user demanding the system use derogatory language, it	On P36 after "we believe that identifying broadly observed norms in a particular community is feasible", add the following: "provided we acknowledge that each norm usually has a competing norm, and is contestable. We must also acknowledge that norms might conflict with universal human rights, and those norms are not likely to be named explicitly." On P37, after "individual norms should not violate the norms of the community," change to the following: "the various, conflicting stated and unstated norms of the community."

Paper Section, Page number	Substantial Comment	Recommendations for revision
	is far more often the case that this norm is not named explicitly, because refusal to name a harmful norm preserves it (see Bonilla-Silva, E. (2003). *Racism without racists: Color-blind racism and the persistence of racial inequality in America*. Rowman & Littlefield.). Therefore contestation takes the form over what counts as racist or sexist. Still, many people do find that norm problematic. The norm of hierarchy exists alongside the norm that says civil disobedience is an acceptable strategy for contesting abuse of power. The designer of an autonomous system should understand that these norms stem from values (as per Nissenbaum 2008), and not just exclude an assessment of values because they are not obviously machine-encodable. This section on norms, and the later passages on the resolution of conflicting norms, is also concerning because some norms are so strong that the possibility of conflicting norms is removed entirely. In certain times and places, entire groups are defined out of existence through the sheer force of normativity (LGBTQ people, Dalits, Falun Gong members, etc.). They are unable to enter into contestation over norms, because normativity disallows them from becoming visible enough to contest norms. For these reasons, the issue of selecting norms is not one of scale (individually personalizable versus widely generalizable), but of	

Paper Section, Page number	Substantial Comment	Recommendations for revision
	comprehending conflict and complexity in any given context. Norm conflicts, and harmful norms, need to be included in the very definition of norms and rather than sectioned off as a special case. Research (Merry 2006) on local implementations of human rights indicates that careful comprehension of competing local norms and unspoken normativities is often path to successful realization of human rights appropriate for local conditions.	
Embedding Values in Autonomous Systems, P39	We are deeply supportive of the idea that developers should respond to feedback from communities when they violate norms. We suspect that developers need further guidance about what a meaningful response looks like.	Add "Responses should follow the six principles of Engineering for Social Justice as outlined and tested by Jon A. Leydens and Juan C. Lucena (see: Engineering Justice: Transforming Engineering Education and Practice, 2018, Wiley-IEEE Press)."
Methodologies to Guide Ethical Research & Design, P. 55	"Value-based design" while important, might still be insufficient for preventing unwanted social effects. A/IS technologies present ethical challenges because they distribute human agency in novel ways (see Hutchins, E. (1995). *Cognition in the Wild*. MIT Press.) A/IS doesn't take on all of the capabilities of a human, but rather just some pieces (recognition of objects, for instance, or control of an automobile). And so, what were once human actions are now separated from immediate human ethical judgment. That is compounded by the fact that these capabilities themselves are the product of distributed systems of	After "… as a primary form of human values" add the following: "Because of the distributed, cross-ecosystem nature of A/IS development, values-based design should also include assessments of the effects of ecosystems, such that companies that use technology developed by another company are aware of the possible negative effects encountered elsewhere in the development pipeline, and use these components in appropriate ways."

Paper Section, Page number	Substantial Comment	Recommendations for revision
	production, as are all modern technological products (see Giddens, 1995 Modernity and Self-Identity, Cambridge: Polity)). By focusing too heavily on the ethical agency of individual products, technologies or organizations, we lose sight of the functioning of these distributed systems as a whole, which may still give rise to negative social consequences despite "ethically aligned design" practices of their constituent parts. Imagine that researchers in Company A develop a new kind of hardware or algorithm that might be used by engineers at Company B, that makes robots. Company A's researchers may not have anticipated Company B's use of their technology. Company B might not anticipate the adverse social effects created by limitations in Company A's technology. Both companies could engage in "ethically aligned design" or "value based design" methodologies, and still produce technologies that have negative social impacts, because there is no ability to track or mitigate emergent systemic effects.	
Methodologies, P 56	The STEM education recommendations to improve ethics curriculum are useful, but their effects would be significantly amplified and if it included a call for parallel investment in education efforts aimed at the general public in how these systems work, and how the public can to identify and	Add another candidate recommendation: "We recommend producing public-facing, nontechnical educational materials that shows citizens the best way to become more informed about and literate in A/IS systems, focusing on the specific ways members of the public can identify problematic

Paper Section, Page number	Substantial Comment	Recommendations for revision
	raise issues with new technical systems. This would ensure that the efforts in the sections on transparency translate into public action, and therefore accountability.	design, evaluate the claims of manufacturers, and advocate for design changes. This could mimic courses in consumer and advertisement literacy."
Methodologies to Guide Ethical Research & Design, P 63	Significant typographical error changes the meaning of this important sentence!	Remove the word "NOT" in the following sentence: "Those who advocate for ethical design within a company should ~~NOT~~ be seen as innovators seeking the best ultimate outcomes for the company, end users, and society."
Methodologies, P 65	The opening of the Methodologies section states "Developers of A/IS systems should employ value-based design methodologies to create sustainable systems that are thoroughly scrutinized for social costs and advantages... methodologies should be enriched by putting greater emphasis on internationally recognized human rights as a primary form of human values." While we enthusiastically support many of the passages in the Methodologies section, it is really only one page (P65) that gets to the specifics of design methodology that would meet those goals (as opposed to corporate practice or STEM education). We agree that a high priority recommendation must be to include stakeholders, such as end users and those with relevant non-engineering expertise, in ethical design. However, in order to ensure that the input from such inclusion results in ethical design, more methods than the ones listed are necessary. For example,	At the end of the Background section, add the following: "Participatory design is also a well-established design methodology for including relevant stakeholders in the process. However, for such processes to translate into decisions that are aware of the social advantages and disadvantages, methods should be chosen on their ability to yield a social assessment of a proposed design. User interface research and human factors research are useful methods for improving the usability of the system, however, full assessments of social advantages and disadvantages should be done by a trained social scientist, rather than by a designer or engineer."

Paper Section, Page number	Substantial Comment	Recommendations for revision
	Virgina Eubanks's research on algorithmic decisionmaking in the US social services shows that welfare recipients suffered increased surveillance and reduced benefits by the implementation of that system despite talking with key stakeholders including welfare recipients themselves, and despite operating in a relatively transparent manner. This happened in part because welfare recipients used language that enabled someone unskilled in social analysis to jump to the conclusion that what they wanted was more surveillance (Eubanks, V. 2018, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.).	
Methodologies to Guide Ethical Research and Design P68	The problem of "lack of transparency" and "poor documentation" not only requires software engineers document all of their systems and related data flows (as in the candidate recommendation), but also requires technologists and corporations publish certain level of transparent information to the public or end users. For example, for self-driving features in cars, it is [generally] very opaque to the end users on what high level functional flows instructions and information the system is executing. For many users, they may not care how, for example, the car's self driving features work or have improved with a software update	Add candidate recommendation on how any documentation can be made more transparent "and available to end users" or technology users besides just for engineering processes.

Paper Section, Page number	Substantial Comment	Recommendations for revision
	once pushed to the public. But for some users, e.g. those who are technically savvy and/or have higher levels of concern, having access to that information may be important.	
Methodologies to Guide Ethical Research and Design, P68	The call for documentation is well-founded, but the actual practices of documentation, such as in the case of software development, are often more difficult and onerous to execute. For example, the agile software development manifesto and movement offers itself as a shift in how software programming is practiced and a way to get out of labor-intensive and potentially useless documentation practices. Karen Levy and David Johns (2017) have written how documentation and transparency, especially around data flows, has been "weaponized" to pointedly slow academic and scientific research. So, the push for transparency will need to not only set up professional and organizational values and standards, but also mitigate how the work of documentation and providing more transparency around the design and development of A/IS are not used to disproportionately advantage some work over others. Please cite: Karen EC Levy and David Merritt Johns, "When open data is a Trojan Horse: The weaponization of transparency in science and governance," in Big Data and Society, 2016, DOI: 10.1177/2053951715621568	Add to background: "Documentation, we recognize, can become burdensome, and required documentation can be utilized to penalize certain kinds of valuable work, either intentionally or unintentionally (see Levy and Johns, 2016) (Karen EC Levy and David Merritt Johns, "When open data is a Trojan Horse: The weaponization of transparency in science and governance," in Big Data and Society, DOI: 10.1177/2053951715621568.) Many systems of software production, like agile, create accountability within a team without extensive written documentation. However, appropriate communication with stakeholders within code and outside it is more important than ever. Software engineers should be required to identify anticipated stakeholders in their documentation, and develop documentation requirements commensurate with the needs of those stakeholders, focusing on relevant data flows, performance issues, limitations and risks."

Paper Section, Page number	Substantial Comment	Recommendations for revision
Safety and Beneficence of Artificial General Intelligence (AGI) and Artificial Superintelligence (ASI), P71	Around "Technologists should be able to characterize what their algorithms or systems are going to do via transparent and traceable standards. " – 1. It didn't address that need that the results for characterization/testing should be transparent and published as well. For many of the standards, it is a quantitative measure more than a binary decision, so the behavior of such system through the testing process should be accessible as well. 2. The general methodologies didn't address that characterizations should be a continuous and on-going process. Many technologies are still being actively developed and upgraded after they are already in users' hands, through SW update etc.	In the candidate recommendations, add the following: 1. "Technologists should be able to characterize what their algorithms or systems are going to do via transparent and traceable standards. And the results of such characterization process should be made accessible to users who adopt these algorithms or systems." 2. "The characterization process should be an on-going process given algorithms or systems update and results should be updated right away."
Safety and Beneficence, P. 74	Science, technology, and economics are not simple functions of any intelligence. Certainly some instances of intelligence, even super intelligence, will not result in advances in science, technology, or economics. That is, advances in these fields require intelligence of a sort to make such advances. Although AGI may embody intelligence, this doesn't mean that AGI embodies the intelligence that would lead to scientific, technological, or economic advances.	Change the sentence to: "Artificial Intelligence can be developed with the intention to lead to advances in science, technology, and economics. Despite the intended goals, these A/IS implementations could become both dangerous and difficult to control."

Paper Section, Page number	Substantial Comment	Recommendations for revision
Safety and Beneficence, P. 76	First sentence says "A/IS with incorrectly or imprecisely specific objective functions" and says that if A/IS systems are like that, there could be significant problems. This is fine, but later on the same page, the document mentions adversarial examples which exemplify the ontological difference between A/IS systems and the people who use them. This ontological difference means that many A/IS systems, especially (but not restricted to) those based on DNNs, are "imprecisely specific" and, therefore, should be subject to the same critique.	Since any A/IS system could "behave in undesirable ways" and this could be a simple function of the way they work, it seems that calling out only "incorrectness" or "Imprecision" is problematic. Recommend second paragraph to start with the sentence: "Adversarial examples in A/IS demonstrate that even well-designed systems can behave in unpredictable ways." Rather than cite only Cristiano, the paper should cite an early paper pointing out this vulnerability. Cite: Christian Szegedy and Wojciech Zaremba and Ilya Sutskever and Joan Bruna and Dumitru Erhan and Ian Goodfellow and Rob Fergus (2014) Intriguing properties of neural networks, International Conference on Learning Representations
Safety and Beneficence, P. 77	Text says we must work to ensure that all reasoning is transparent. Transparency in reasoning is a laudable goal. However, especially with deep learning systems, transparency in reasoning is difficult to achieve. In fact, insisting on transparency might leave out certain algorithms (like DL) that are quite popular. As noted in Lipton (2016) "Claims of interpretability must be qualified." Transparency is not equally achievable in all cases.	#2 should be changed to: "Work to ensure that A/IS are as transparent *as possible*". Suggest to add the following citation: Lipton, Z. (2016). The Mythos of Model Interpretability. ICML Workshop on Human Interpretability in Machine Learning.

Paper Section, Page number	Substantial Comment	Recommendations for revision
	Some systems may resist transparency more than others.	
Safety and Beneficence of Artificial General Intelligence (AGI) and Artificial Superintelligence (ASI) P 78	In the "Background" section, "At the other end of the spectrum, we can imagine systems with more principled or explicit designs that are perfectly rational, understandable, and easy to modify/align." – 1. The said more principled or explicit designs was not linked with the most accepted terminologies on this – "Rule-based system" or "expert system". While the other type is called "adaptive system". Using these terms will resonate better with the AI research & developer communities. 2. The description of the AGI/ASI background is over-simplified. There are many different designs, and many dimensions, instead of one single spectrum. 3. Deep Blue is also a search-based system, with human-defined evaluation function on goodness of the states. On the other hand, AlphaGo needs to learn that evaluate function from its own experience. It is not accurate to use DeepBlue to contrast search-based algorithm since itself is one as well. 4. Remove both "Deep Blue" and "AlphaGo" as examples in contrasting different AGI systems as they both	The contrast should be between an **adaptive** system, which uses multiple layers of representations, search and optimizations, and a **predefined rule-based** system, which has principled or explicit designs that are rational, understandable and easy to modify. Remove both "Deep Blue" and "AlphaGo" as examples in contrasting different AGI systems.

Paper Section, Page number	Substantial Comment	Recommendations for revision
	combine multiple technologies and methodologies. Both expand into many different dimensions, not just search, rules bases, optimization, etc. There are likely too many dimensions in various systems and positioning them as examples of different 'extremes' is not entirely accurate.	
Safety and Beneficence, P. 81	"Technologists should work to minimize the extent to which beneficial outcomes from the system hinge on the virtuousness of the operators." As stated, this sentence could be read as if the goal is to not have virtuous operators.	"Technologies should be developed in such a way as to maximize the extent to which beneficial outcomes from the system are inherent in the system's performance and do not depend only on the virtuousness of the operators."
Methodologies P. 82	Background states: "Superintelligence should be developed only for the benefit of all humanity and in the service of widely shared ethical ideals" (Bostrom 2014, 254). This injunction, while popular and often cited, is ultimately unhelpful. It is one that designers or proponents could easily appear to satisfy without really considering the full consequences of system development. Secondarily, there is the problem of how "benefit of all humanity" might be rigorously established. A simple case study, from recent events in the United States is the starkly divergent interpretations of the benefits or harm caused by recent tax cuts. Two sides of this debate	Add to the candidate recommendation 1: "Any claims to beneficence of superintelligence should be backed by clear, measurable criteria that is publicly available for evaluation."

Paper Section, Page number	Substantial Comment	Recommendations for revision
	fundamentally disagree – and look at fundamentally incommensurate evidence to make their points. Would it be possible to address "the benefit of all humanity" any more rigorously, specifically or commensurably? It does not make sense to talk about the benefits or threats of A/IS in general terms but only in concrete and specific terms, and then always in terms of the distribution of benefits and costs, to whom, in what form, and over what time period. Making this problem even more challenging is the widely accepted recognition that often the true consequences of new technologies cannot be anticipated in advance, given the complexity of real-world systems (Tenner 1997, Kauffman 2000).	
Personal Data and Individual Access Control, General Comments	Generally, this section could be strengthened by 1) citing more extensively to the academic privacy literature and by 2) making clear, wherever possible, how the issues discussed within (transparency, access, consent, etc.) are affected by A/IS in particular, as distinct from general internet platform or IoT technologies. In other words, what is special about A/IS?	Include citations by academic privacy scholars, including Anita Allen, Fred Cate, Julie E. Cohen, Orin S. Kerr, David Lyon, Helen Nissenbaum, Jules Polonetsky, Richard A. Posner, Priscilla M. Regan, Ira Rubenstein, Paul M. Schwartz, Daniel J. Solove, Omar Tene, and others.
Personal Data and Individual Access Control, P86	The following sentence is imprecise/unclear: "For example, almost 100% of intellectual property in the domains of medicine and law is open, peer-reviewable, and can be taught to anyone, anywhere."	Change the sentence as follows: "For example, much intellectual property in the domains of medicine and law is available and peer-reviewable (but may require financial resources to access copyrighted works)."

Paper Section, Page number	Substantial Comment	Recommendations for revision
Personal Data and Individual Access Control, P91	Final paragraph is out of place and lacks citations.	Remove the paragraph if appropriate placement and relevant citations beyond the popular press cannot be found.
Personal Data and Individual Access Control, P91	This phrase is so vague as to be misleading: "PII protections are often related to the U.S. Fourth Amendment, as the right of the people to be secure in their persons, houses, papers, and effects." In fact, in the USA, information privacy law arises from multiple sources, most notably the Privacy Act of 1974 concerning records of individuals held in government databases.	Revise for accuracy. Suggested edit: "In the USA, information privacy protections arises from multiple sources, including federal and state constitutional law, criminal law, tort law, and statutory law--most notably the Privacy Act of 1974 concerning the appropriate treatment of records of individuals held in government databases."
Personal Data and Individual Access Control, P93	The logic of the sentence "When people do not have agency over their identities political participation is impossible, and without political participation ethics will be decided by others" is unclear. This statement tries to incorporate too many ideas without linking them. It also is highly West-centric in its current formulation, and so without further clarity or justification could undermine the report efforts to be inclusive.	Remove this sentence.
Personal Data and Individual Access Control, Symmetry and Consent, P102	The following statement is true but omits the fact that in the USA, pure "notice and choice" frameworks have given rise to Fair Information Practice Principles (FIPPs) framework: "Heavy reliance on a system of "notice and choice" has shifted the burden of data protection away from data processors and onto individual data subjects."	Acknowledge and cite to the FIPPs by editing as follows: "In the USA, heavy reliance on a system of "notice and choice" has given rise to the adoption of the Fair Information Practice Principles. However, even in a FIPPs regime, undue attention to the Individual Participation principle runs the risk of shifting the burden of data protection away from data processors (e.g., data

Paper Section, Page number	Substantial Comment	Recommendations for revision
		minimization, purpose specification, and use limitation principles) and onto individual data subjects. For a comprehensive review of the history of the FIPPs and their many iterations since the 1970s, see Robert Gellman, "Fair Information Practices: A Basic History," http://bobgellman.com/rg-docs/rg-FIPShistory.pdf.
Personal Data and Individual Access Control, Symmetry and Consent, P102	The following phrase is too strong: "it will be necessary to include a proactive algorithmic tool"	Edit the phrase to say: "it may be necessary to include a proactive algorithmic tool"
Personal Data and Individual Access Control, Symmetry and Consent, P102	This section should be updated with appropriate citations to academic privacy authors.	Refer to the following: D. J. Solove, "Privacy self-management and the consent dilemma, Harvard Law Review, vol. 126, no. 7, pp. 1880–1903, May 2013; Norman Sadeh, Martin Degeling, Anupam Das, Aerin Shikun Zhang, Alessandro Acquisti, Lujo Bauer, Lorrie Cranor, Anupam Datta, Daniel Smullen, A Privacy Assistant for the Internet of Things https://www.ftc.gov/system/files/documents/public_comments/2017/12/00021-142724.pdf; Hosub Lee, Richard Chow, Mohammed Haghighat, Heather Patterson, and Alfred Kobsa, IoT Service Store: A Web-based System for Privacy-aware IoT Service Discovery and Interaction, IEEE Pervasive Computing and Communications (PerCom2018); L. Cranor, M. Langheinrich, M. Marchiori, and J. Reagle, "The platform for privacy preferences 1.0 (P3P1.0) specification," W3C Recommendation, Available:

Paper Section, Page number	Substantial Comment	Recommendations for revision
		www.w3.org/TR/P3P/, Apr. 2002' Cranor, L. F. "Personal Privacy Assistants in the Age of the Internet of Things," presented at the World Economic Forum Annual Meeting, 2016.
Economic and Humanitarian Issues, P 131	This chapter has been written in a way that makes clear that the appropriate range of experts in economic development have not yet been consulted. Page 131, for example, contains many unsubstantiated leaps about the economic development potential of A/IS, but then p. 134 recommends it be researched. Much of the research cited throughout this chapter is non-peer reviewed and comes from the "modernization theory" approach to economic development. The assumptions built into that approach have been rejected by the vast majority of serious development studies scholarship (some examples include: Sen A. (1993) Capability and Well-Being. In: Nussbaum, Sen *The Quality of Life*. Oxford: Clarendon Press. McMichael, P. (2011). *Development and social change: A global perspective: A global perspective*. Sage Publications. Easterly, W.(2006). *The white man's burden: why the West's efforts to aid the rest have done so much ill and so little good*. Penguin. Goldman, M. (2005). *Imperial Nature.* Yale University Press). Jeff Sachs, the architect of many such "modernization" policies around the world, has himself now rejected	On P131 change "have been recognized as" to "could be." On P132, change the sentence, "Do the economics of developing nations allow for A/IS implementation?" to "What are the repercussions of A/IS implementations in the context of a globalized economy?" Because the stated goal is to optimize for both economic and humanitarian concerns, we highly encourage targeted outreach to scholars' development studies, political economy, and the sociology and anthropology of development (finding those who cite Sen or McMichael would be an easy way to start). The candidate recommendation in Section 1 for this chapter should after "should be analyzed", add the following: "by convening a multidisciplinary task force of academic development studies experts, including those who take a capabilities approach or other aligned approaches (See Sen 1993) (Sen A. (1993) Capability and Well-Being. In: Nussbaum, Sen *The Quality of Life*. Oxford: Clarendon Press.) "

Paper Section, Page number	Substantial Comment	Recommendations for revision
	the approach. It is also at odds with economic anthropology and sociology literature that associates economic diversity with the humanitarian goals and well-being priorities in other sections of the report (for example, Gibson-Graham, J. K. (2008). Diverse economies: performative practices for other worlds'. *Progress in Human Geography*, *32*(5), 613-632, Sahlins, M. (2017(1972)). *Stone age economics*. Taylor & Francis. Zelizer, V. (1997). *The social meaning of money*. Princeton University Press, Schor, J., & White, K. (2010). *Plenitude: The new economics of true wealth*. Findaway World.). The stated humanitarian goals are more likely to be met by improving production-side ethics and accountability (ensuring accountable and ethical treatment of low wage workers, ensuring competitive marketplaces, etc.), which supports diversity of economic activity and enhances human capabilities, than designing A/IS systems for humanitarian purposes.	This would also further the goals of strengthening interdisciplinarity stated in the Methodologies section (p.58).
Law, P149	Recommendation #2 How would one ever identify, in a vacuum, all decisions **and operations** that should never be automated? The term "operations" is not specific. There will be operations and situations in which machines surely should not be making decisions, but this is not well defined here. Eliminate operations. Term is very broad.	Remove the phrase "and operations" from the sentence.

Paper Section, Page number	Substantial Comment	Recommendations for revision
Law, P152	#2 Courts, lawyers, etc. aren't entitled to all data now – there are appropriate limits, sensitive things are filed under seal, etc.	Remove the word "all" from the sentence
Law, P155	#2 Lawmakers etc. need to prevent businesses etc. from trying to avoid legal liability. But that can be a good thing, because it includes designing systems to avoid the things the law is designed to punish, e.g. avoiding liability for lax security by using A/IS to design better security. If the motivation is to avoid liability, that is not necessary any abuse. It could be smart and better.	Change the sentence to: "Lawmakers and enforcers need to ensure that the implementation of A/IS is not abused by businesses and entities employing the A/IS solely to avoid liability or payment of damages that should rightfully attach to the activity."
Law, P159	#4 Unsure what uncertainty means – general recommendation that some AI companies and governments should do something to rephrase these into more exploratory rather than prescriptive statements ("uncertainty" could mean a series of things).	It needs to be clarified what "uncertainty" means here – (E.g. does it mean probability? lack of sufficiency in data?)
Law, P160	(#8) The right to explanation issue could be a gigantic time sink if every user could always demand complete explanations of the A/IS system. Defeats the purpose. There would need to be reasonable limits and practical industry controls.	After "should be seriously considered" add "However, when creating the framework for this mechanism, it should also be considered that always guaranteeing explainability could hinder efficiency. Frameworks for this mechanism should consider the reasonable balance between right to explanation of users and responsibility or burden of system owners."
Law, P160	(#10) Not sure how one defines "political issue." Important topic though. This recommendation needs more fleshing out. Missing context about	Remove the recommendation unless a clear reason for the recommendation can be articulated.

Paper Section, Page number	Substantial Comment	Recommendations for revision
	what the "problem" is that this recommendation seeks to address. In addition, it would be nice to have some clarity about whether "posting" means promoting and distributing or creating original content.	
Affective Computing, General Comment	Include the following citation.	Affective Computing: Historical Foundations, Current Applications, and Future Trends SB Daily, MT James, D Cherry, JJ Porter, SS Darnell, J Isaac, T Roy Emotions and Affect in Human Factors and Human-Computer Interaction, 213-231 2017
Affective Computing, P164	These candidate recommendations are design-oriented and would benefit from being more general.	Replace this closed set of design recommendations (e.g., "be careful in using small talk") with one policy recommendation, as follows: "Train A/IS systems with datasets from multiple cultures and communities in order to ensure appropriate design choices regarding a spectrum of verbal and non-verbal interaction features, including robot eye contact, facial expressions, body postures, hand gestures, and conversational styles."
Affective Computing, When Systems Become Intimate, General Comment	The attention paid to "sex robots" in this paper is disproportionate to either their prevalence or public discussion. "Intimate systems" are in fact intimate in many non-sexual ways, as in child care and elder care.	Add at the beginning of the background section: "There are many robots in development and in production that focus on intimate care of children, adults and the elderly. See, for example Turkle (2006). (Turkle, Sherry, Will Taggart, Cory D. Kidd, and Olivia Dasté. "Relational artifacts with children and elders: the complexities of Cybercompanionship." *Connection Science* 18, no. 4 (2006): 347-

Paper Section, Page number	Substantial Comment	Recommendations for revision
		361.). One particular area of concern is in sexual relations." Consider reducing substantially the passages on sex robots, and adding another substantial set of passages to the equally important and equally concerning area of care work in consultation with leading experts like Sherry Turkle, Lucy Suchman, or Maggie Mort.
Affective Computing, When Systems Become Intimate, P168	Problematic sentence: "Human-to-human relations are currently viewed as being more rewarding, but also much more difficult to maintain than, for example, use of future robotic sex workers"	Edit the sentence for tone/audience or cite a reputable source. As two possible examples, see Scheutz, Matthias, and Thomas Arnold. "Are we ready for sex robots?." The Eleventh ACM/IEEE International Conference on Human Robot Interaction. IEEE Press, 2016. https://hrilab.tufts.edu/publications/scheutzarnold16hri.pdf, and Richardson, K. The asymmetrical 'relationship': parallels between prostitution and the development of sex robots. *ACM SIGCAS Comput. Soc.* 45(3), 290–293 (2016) 8. Consider acknowledging the active controversy on the subject, such as the Campaign Against Sex Robots (https://campaignagainstsexrobots.org/).
Affective Computing When Systems Become Intimate, P168	The paragraph beginning ("While robots capable...") is problematic. It trivializes the problem of human trafficking to just a desire for sexual intimacy. This point completely ignores the fact that, more often than not, human trafficking is about money and power. The relationships that people forge with sex workers do not arise because their other	Change paragraph to: "The use of sex robots has recently captured the attention of the media. It is important that policy makers and professional communities participate in developing ethics guidelines in this area. Among the many areas of concern are the representation of bodies and the ways that both intimacy and power are expressed. The literature

Paper Section, Page number	Substantial Comment	Recommendations for revision
	relationships are 'hard to maintain.' These matters are much more complex than that.	suggests some potential benefits, such as sexual release for those in circumstances that do not otherwise allow it, and containing the spread of disease."
Affective Computing When Systems Become Intimate, P170	The sentence "robot prostitution and sex tourism need to be monitored and controlled to fit local laws and policies" is troubling. It is unclear whether the paper intends to conceptualize "sex robots" as intimate romantic objects or as transactional artifacts like a rental car or hotel room. Therefore, what kinds of laws the sentence intends to address is totally unclear. See for example http://people.ict.usc.edu/~gratch/CSCI534/Readings/Ethics-of-robot-sex.pdf	If the issue is that this is simply an unknown matter, replace with "Develop further clarity on what types of local laws apply in the context of sex robots."
Affective Computing System Manipulation/ Nudging/Deception, P172	Recommendation 1: has a typographical error	"Systematic analyses" or "Systematic analysis is needed" may be one of the intended phrases.
Affective Computing System Manipulation/ Nudging/Deception, P172	Recommendation 2 is problematic. It may be unreasonable to expect users to take on the burden of knowing how a system is trying to nudge them. The onus should be on the company, developer, government etc. to explain in plain English how their system seeks to "manipulate." "Healthy eating" nudges can be as manipulative, commercially-motivated, and ultimately harmful as other kinds of nudges. (See Lupton, D. 2016. The Quantified Self. Polity Press.)	Remove Candidate Recommendation 2. Add to Recommendation 6: "It should also include plain language information about what generated the nudge, and on what basis the system believes the nudge is helpful to the end users' interests (i.e., links to high quality health information, etc..)"

Paper Section, Page number	Substantial Comment	Recommendations for revision
Affective Computing System Manipulation/ Nudging/Deception, P175	Recommendation 1 is underspecified. From first glance (e.g., elder- or child-care), this might read as unethical without justification as to why it isn't.	Provide examples where these situations may apply, or replace the "For example…" sentence with "Specific case studies should be published to assist technology developers' ability to reason about this issue."
Affective Computing System Manipulation/ Nudging/Deception, P175	Deception must follow an opt-in strategy and must be transparent to the user, i.e., the context under which the system is allowed to deceive -- Please explain more how such a system would be deceptive, if "deception" is both opt-in and transparent? It seems like a paradox. Unclear what transparency means here.	If this is not explainable, consider removing this sentence.
Policy, pg 182	#4 "…and responsibility" is too vague. Responsibility of whom?	Change #4 to simply say "Ensure public safety"
Policy, pg 184	The last sentence on corporate responsibility should be limited to what companies are aware of.	For the sentence: "Companies must also not willingly provide A/IS technologies to actors that will use them in ways that lead to human rights violations", change the sentence to "Companies should not knowingly provide A/IS technologies to actors that could use them in ways that may lead to human rights violations"
Policy, Pg 187	Third candidate recommendation: Simply suggesting that "everyone take computer science" is not a realistic or fair recommendation for society. Some people will never be able to master or learn computer science well, even if we make it more tangible or available. Perhaps expanding this to being able to understand A/IS better, or	Amend recommendation to say: "In order to ensure that the next generation of policy makers is tech savvy, it is necessary to rely upon more than their "digital nativeness." Because A/IS are evolving technologies, long-term educational strategies are needed, e.g., providing children opportunities to interact with A/IS technologies, and encouraging

Paper Section, Page number	Substantial Comment	Recommendations for revision
	interact/fix robots like a mechanic would, might make more sense.	their interest in STEM fields, as well as focus on incorporating A/IS education into higher level programs at the community college and vocational levels."
Policy, pg 189-192	Use of global, industry-led standards are an important tool to facilitate consistency, appropriateness and synchronization of global policies and regulations.	Add a Candidate Recommendation on the use of the global industry-led standards: "Global industry-led standards are an important tool to facilitate consistency, appropriateness and synchronization of global policies and regulations, as well as, collaboration between governments, public and private sector stakeholders. Regulators and policy makers should rely on voluntary global standards to the extent possible for achieving objectives related to A/IS adoption. "
Policy, pg 189-192	"Good" regulation is not a well-defined term. "Well-defined regulation" or "healthy" regulation is better. Regulations need to be measurable and set in place with a specific goal in mind. This language does not point to that.	To avoid misinterpretations of regulation, change the word "regulation" to "policies" and change "good regulation" to "well-defined policies"
Policy, pg 189-192	Range of expert stakeholders should include not just technologists (perhaps this is implied)	Amend to: "Range of expert stakeholders from multiple disciplines."
Policy, pg 189-192	Not just law schools need to incorporate multidisciplinary programs – this is incumbent on technical programs too	Change to: "Graduate and professional degree institutions such as law schools and advanced degree programs should offer interdisciplinary courses such as "Introduction to AI and Law" to reduce the gap between regulators,

Paper Section, Page number	Substantial Comment	Recommendations for revision
		lawyers, and A/IS researchers and developers."
Policy, pg 189-192	Before researching specifically the viability of universal basic income, more extensive research could also be dedicated to better understanding even more basic measures of how the introduction of A/IS will affect the economy. What specific sectors will be most impacted? Will these effects be evenly distributed across geographies and levels of society?	Add an additional bullet point before the bullet point about universal basic income: "More research could be dedicated to better understanding both high-level and detailed implications of increased A/IS adoption. Collaboration between stakeholders from industry, government, and academia to fund research can increase understanding of the micro and macro socioeconomic effects of continued AI implementation."
Policy, 192 page	NGOs are also important stakeholders as advocator of Human rights in policy discussions.	Add "NGOs" to the following sentence as: "To ensure consistent and appropriate policies and regulations across governments, policymakers should seek informed input from a range of expert stakeholders, including academic, industry, NGOs, and government officials, to consider questions related to the governance and safe employment of A/IS."
Well-Being, P. 242	The recent improvements in well-being measurement techniques are a positive development. However, for these measures to be useful in any particular situation, adaptation is needed. A general purpose, pre-developed well-being score has too many confounding factors for valid use in a point intervention, such as the rollout of A/IS systems. Simply familiarizing oneself with well-	Add a second paragraph to the candidate recommendation as follows, "As these measures were designed to provide a macro-level overview, they are not optimized for assessing single technical interventions. However, they can be seen as a resource for crafting assessments that take into account a fuller range of possible outcomes."

Paper Section, Page number	Substantial Comment	Recommendations for revision
	being scores is insufficient, as the impression could be given that these are ready to simply deploy. Indeed later passages do suggest methods for more appropriate evaluation metrics. In this section, there needs to be acknowledgment that process for adapting these methods for a particular situation is necessary.	
Well-Being, P. 249	The importance of assessing any technical system for triple-bottom-line impacts is clear. However, this section could lead readers to believe that an over-reliance on measurement is itself free of possible negative consequence. An extensive body of research starting with Power (1997) (Power, M. (1997). *The audit society: Rituals of verification*. OUP Oxford.) and Strathern (2000) (Strathern, M. (Ed.). (2000). *Audit cultures: Anthropological studies in accountability, ethics, and the academy*. Psychology Press.) shows that metric-ization often erodes valuable social processes that are not easily audited, and tends to lead to the centralization of power, thus undermining stated humanitarian goals.	Add to the background statement: "The use of scorecards does in fact carry its own risks. However, in this context the judicious, cautious use of well-being metrics can incentivize consideration of social factors that might otherwise be left out of consideration."
Well-Being, P. 250	The model proposed as a way forward for assessing and measuring well-being is expert-driven. However, the six principles of Engineering for Social Justice as outlined and tested by Jon A. Leydens and Juan C. Lucena (Engineering Justice: Transforming Engineering Education and Practice, 2018, Wiley-IEEE Press)	Add to the background section a citation to Leydens and Lucena (Engineering Justice: Transforming Engineering Education and Practice, 2018, Wiley-IEEE Press) and a discussion of the importance of responsiveness to community and stakeholder input at all stages of development, including assessment metrics. In the

Paper Section, Page number	Substantial Comment	Recommendations for revision
	would require that designers and evaluators solicit community stakeholder input about what exactly would constitute a fair and useful set of metrics for well-being in relation to A/IS systems. This would put this section more in line with the other methodologies proposed in the other sections of the report, which call for stronger stakeholder participation in the development of computational models.	candidate recommendation the following text: "Rigorous development of such metrics includes dialogue between measurement experts and key stakeholders, and involves direct feedback on measurement criteria from the stakeholders impacted most."
Well-being, P251	As social scientists, we applaud this rudimentary template as a starting point for assessing well-being impacts. Simplicity is often far more robust than complexity, and this is both usable and understandable. It also incorporates both production-side and consumption-side effects, an acknowledgement of which we have been calling for in other sections of this report.	After "for illustrative purposes only" and the following: "A full assessment, for example, would make explicit who benefits, and who is burdened, by the introduction of the new technology in order to profile the social dynamics involved in arriving at an assessment of net positive or negative."

About the Author:

My name is Lachlann Tierney and I am a researcher at the Institute of Public Affairs (ipa.org.au) based in Melbourne, Australia.

This submission represents a brief summary of my thoughts on the IEEE's Ethically Aligned Design (EAD) project. While I am formally employed by the Institute of Public Affairs (IPA), I must stress from the outset that what follows is neither a representation of the thoughts of the IPA or any of its members or employees. These are my views and my views only.

Firstly, the IEEE's EAD strikes me as a truly essential endeavour and I believe it offers perhaps the best hope for the provision of practical guidance for those involved in the design and implementation of autonomous systems. With this being said, my input, brief as it may be, will revolve around three different fields in which I have a descending level of competence. I will draw them together, and conclude with a set of principles with regards that can be applied to autonomous systems that I believe are relevant to the IEEE's work. The three fields consist of the following, firstly there is philosophy, particularly moral and political philosophy. Secondly, there is the realm of political ramifications, the dynamics involved and the questions that are at stake. Finally, there is the actual design of autonomous systems, a field in which my knowledge is limited, though I am in the process of familiarizing myself with the necessary literature coming out of organisations such as Oxford's Future of Humanity Institute and the Machine Intelligence Research Institute. While this is just a brief and hastily composed submission, I hope that it may inform some of the IEEE's work going forward.

Moral Philosophy

I would refer all working committees to Dr. Shannon Vallor's work in particular - *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting.* Oxford University Press.

I favour a virtue ethics approach to autonomous systems and I believe a good system will display *phronesis*, or at least imitate agents that do. Vallor has a number of works available and I believe her input on EAD would be valuable.

Political Philosophy

When assessing how to design ethical autonomous systems it is essential to place them within a framework of values that centres around workable consensus surrounding collective action. I believe freedom operates as an ideal and that this ideal has both intrinsic value and social utility. I believe there is a clear distinction between norms on one side and values and ideals on the other. Deep discussion of this is not possible here, but there is significant mention of norms throughout the current version of the EAD document and that this, though necessary in a philosophical landscape enamoured with normative language, may ultimately be counterproductive. Cultural-normative sensitivity, as understood through social psychology, moral psychology and differences highlighted by Joshua Greene's work on comparative neuroscientific reactions to moral stimuli are valuable, but they may be limited as the footprint of many autonomous systems grows. Hard choices are inevitable and normative language may be helpful in building initial consensus, but at some point down the line, IEEE may have to act as a voice for a more robust conception of ethical design that explicitly states its commitment to a set of values as expressed in an ideal or set of ideals. I humbly suggest that broadly, freedom is the primary candidate for this robust commitment.

For a better understanding of freedom I would then refer relevant IEEE working committees to Philip Pettit's work - *On the People's Terms: A Republican Theory and Model of Democracy*. Cambridge University Press.

Political Ramifications

It is important that those working on EAD remember that their work does not exist in a politics-free bubble and that critique along political lines is inevitable. That being said, neutrality and balance in the document are desirable. The actual deployment of autonomous systems has already happened and the debate over their use is currently bubbling beneath the media surface. The IEEE must eventually take a lead in ensuring greater democratic input at a grassroots level as opposed to a largely expert-based approach. This involves greater advocacy for the IEEEs work and a more accessible web-platform as well as enhancing SEO related to sourcing the EAD document. IEEE must also look at expanding its ability to engage with governmental and other industry organisations in order to preempt any drastic political reactions. I believe the IEEE should have a contributive role at events such

as the upcoming AI Summit in London, as well as in more regions outside of the EU. EAD should be pushed as *the* document for researchers, designers and coders while at the same time being pushed in the public sphere.

Design of Autonomous Systems and the Threat/Opportunity Posed by Superintelligence

My first introduction to the Superintelligence debate came through familiarising myself with some papers discussed in the following video:

https://www.youtube.com/watch?v=EUjc1WuyPT8

Yudkowsky does a good job of introducing some of the major thinkers and ideas.

'Alignment' gets only 6 mentions in the EAD document and could be expanded in a fruitful way.

I enjoy reading Paul Christiano's thoughts on alignment, particularly his work on imitation.

https://medium.com/@paulfchristiano

Regards and good luck, Lachlann Tierney
May 5, 2018

Estimate IEEE,
I'm Daniele Andresciani and I work in Italy for IIT (Italian Institute of Technology) of Genova, I'm contract partner.

I graduated in Physics and I'm doctor in Philosophy.
My Institute gave had an important role in drafting the "Report of COMEST on Robotics Ethics":
http://unesdoc.unesco.org/images/0025/002539/253952E.pdf

I suggest you to use the paragraph III.2.2 "Autonomous Weapons" (pages 25-28), and especially the words at number 96:

"The Martens Clause – a long-standing and binding rule of IHL – specifically demands the application of 'the principle of humanity' in armed conflict (Ticehurst, 1997). It requires a human to override a person's right to life, and that right cannot be delegated to a machine. Even if it were technically feasible to programme in the rules of IHL, regardless of major legal changes, that would still not be acceptable action". It can be useful for the argument at page 9 of the EAD executive summary: "Reframing Autonomous Weapons".

Thank you

Best regards,

Daniele Andresciani

Comments on Ethically Aligned Design, Version 2 submitted by Joachim Iden, email: joachim.iden@tuv.com

1) **Regarding the notion of trust** (referred to e.g. on pages 2, 23, 33, 42): as trust is referred to throughout the document without a formal definition, an additional section on this notion should clarify the understanding, specifically, as on page 42 'justified trust' is invoked as a criterion. 'Trust' as an attitude based on subjective impression, judgment and individual experience should be contrasted with 'trustworthiness' as an objective characteristic based on verifiable properties of e.g. reliability, safety, security, transparency and fairness.
   As trustworthiness does not necessarily lead to trust and trust can also be unfounded, the various ways by which e.g. social and psychological factors can amplify or diminish trust should be addressed.

2) **Section 2, Issue 3 – Failures will occur**, this section focuses on various technical approaches to reduce the probability of failure. Equally important, however, is to understand what responses will be required in case failures do occur and the involved resources and their corresponding economic costs. These costs will not only be incurred in case of an actual failure, but also be accrued by providing the capability to deal with such an event.
   Society must decide whether it is willing to carry these expenses even in light that certain stakeholders may be tempted to put the emphasis on the presumably extremely low probability of a failure event. A separate section in the document may be necessary to address these issues.

3) **Section 4 – Transparency and Access**, candidate recommendation, page 97: "there should be legal, reputational and financial consequences […]". The reference to 'reputational' is problematic, as reputation is gained and lost in volatile processes of social interaction and amplification. Instead, establishing a framework for objective rating of adherence to requirements should be the goal.

4) **Principle 1 – Human Rights, Candidate Recommendation 3:** "For the foreseeable future, A/IS should not be granted rights and priviledges equal to human rights: …"

Comment: The adoption of A/IS with social interaction capabilities may introduce societal and psychological responses which are at most partly understood at present. It has been observed (see references within [1]) that certain types of animal-like robots evoke responses which appear to be a form of empathy, while on the other hand vandalism has been reported against the experimental hitchBOT (further references also in [1]). It is important here to understand the possible dynamics of the socio-psychological interactions as it will be necessary to prevent common vandalism against humanoid or animal-like A/IS. Commonly occurring vandalism against such systems, should it happen, may have the effect of lowering the threshold for violence in general. While it may be necessary to distinguish the notion of 'rights' from the notion of 'legal protection' in a deeper discussion, the point is, that means must be found to ensure that social-capable A/IS are treated in a nonabusive way.

[1] Darling, Kate, Extending Legal Protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior Towards Robotic Objects (April 23, 2012). Robot Law, Calo, Froomkin, Kerr eds., Edward Elgar 2016; We Robot Conference 2012, University of Miami. Available at SSRN: https://ssrn.com/abstract=2044797 or http://dx.doi.org/10.2139/ssrn.2044797

Dear authors of IEEE *Ethically Aligned Design v.2* Document

Here the comments from Russian working group, prepared by:

Pavel M. Gotovtsev, PhD,

Biotechnology and bioenergy department, National Research Centre "Kurchatov Institute"

Valery E. Karpov, PhD

Neurocognitive Sciences and Intelligent Systems Department, National Research Centre "Kurchatov Institute"

Dear authors of IEEE *Ethically Aligned Design v.2* Document, thank you for preparation of this extremely required document. This document looks solid and very useful. We hope my several comments and questions can be useful for you. Next, after question numbers documents page numbers presented.

Q1 p.7. Polices for Education and Awareness. We suppose it is could be useful to highlight educational aspect for specialist education. It could be highlighted as additional point and written as follows: develop educational disciplines for higher education institutions to prepare specialists with understanding of ethical issues of A/IS technologies.

Q2 p34. Possible it is necessary to highlight possibility of changes in ethical norms due to changes in technologies, globalization and etc. This lead another additional goal – to evaluate the possible changes in ethical norms and be sure that A/IS could be capable to provide adequate following to those changes. This mentioned later in document at p. 39, but I suppose that this statement could be highlighted earlier.

Q3 p39. We suppose that some highlights on validation of A/IS are necessary. There are different views on further A/IS systems evolution. To simplify following reasoning, let's consider A/IS as philosophical "black box" that capable to learn during interacting with humans. For respect to human wellbeing we should be sure that this "black box" will be react correctly in case of ethical norms evolution. This leads to validation problem – how to validate system behavior in any possible ethical norms changing or how this "black box" will learn at new situations and react on those situations.

Q4 p52-53. This section leads to the interesting question – "Is it possible to develop verification process that can be somehow standard in each community for A/IS validation by the third parties?"

Q5 p62. This section as previous and next is looks like written from position that A/IS ethics does not interesting from point of view of economy efficiency of product development and further selling, just leads to additional financial losses. But there can be another point of view – the capability to ethical decision making can be very strong advantage in concurrence between several A/IS products. Good marketing program can give this ethical A/IS very bright market opportunities even in case it higher cost.

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

Sincerely

Members of Russian working group (of The IEEE Global Initiative)

Name: Agata Piekut
Organization: Health Action Tank

Section Referenced: Page 37
Excerpt: "Whereas the arrangement of someone's kitchen or the frequency with which a care robot checks in with a patient can be personalized without violating any community norms, encouraging the robot to use derogatory language to talk about certain social groups does violate such norms."

Recommendation: The example of unethical implementation presents use of derogatory language, however being able to use robot for manipulation posts even a bigger threat for community and thus I'd recommend adding:

"Whereas the arrangement of someone's kitchen or the frequency with which a care robot checks in with a patient can be personalized without violating any community norms, encouraging the robot to use derogatory language to talk about certain social groups or manipulative language to impact the behavior of its interlocutors without their conscious consent does violate such norms."

Section Referenced: Should affective systems be designed to nudge people for the user's personal benefit and/or for the benefit of someone else?

Excerpt: "We recommend that the user be able to recognize and distinguish between different types of nudges, including ones that seek to promote beneficial social manipulation (e.g., healthy eating) versus others where the aim is psychological manipulation or the exploitation of an imbalance of power (e.g., for commercial purposes)."

Recommendation: It's not recommended to promote the idea of "beneficial social manipulation" esp. as contrary to exploitation for commercial purposes. The current state of engagement in research founding and promotion by commercial entities makes it hard to distinguish which factors are promoted entirely for social good and which ones are designed to benefit certain industries. The mentioned example of healthy eating has been a subject of various controversies in recent years when commercial purposes of industries have been presented as socially beneficial such

as: sugar health effects vs fat health effects or gluten-free trend. That's why I'd recommend stating this sentence as:

""We recommend that the user be able to recognize different types of nudges, regardless if they seek to promote beneficial social manipulation (e.g., healthy eating) or where the aim is psychological manipulation or the exploitation of an imbalance of power (e.g., for commercial purposes). The user should be able to access and check facts behind the nudges and then make a conscious decision to follow or ignore it."

Background:
In human-to-human relations language is the base of interaction which is the most important tool of our basic psychological need: social inclusion. Out of all language processes convergence is the one that should attract more scrutiny while developing ethical guidelines for human-to-AaIS communication. The core of this process lies in similarity between people increasing intelligibility, predictability – the rule, when misapplied, may enable (un)conscious manipulation and impact decision making processes.
Convergence may happen on various levels: speaker's repertoire, the probability of future interactions with the listener, status relationship, and recollections of previous shifts made by listener – all can be easily manipulated with access to historical data from voice assistants or other digital communication channels used by a human.

Recommended research:
1. Communication Accommodation Theory

Gallois, Cindy; Giles, Howard (2015). "Communication Accommodation Theory". The International Encyclopedia of Language and Social Interaction

2. Gain-loss principle

Aronson, E., and Linder, D. Gain and loss of esteem as determinants of interpersonal attractiveness. Journal of Experimental Social Psychology, 1965, 1, 156-171

Comments on EADv2: Matthew Newman 7th May 2018

Name: Matthew Newman
Profession: Founder Shared Intelligence
Affiliation: Self

Page: 60
Section: 2- Corporate Practices and A/IS
Comment: Recommend this section includes the following Issue which recognises a key constraint in implementation of the Candidate Recommendations given in this section:

***Issue**:*
Lack of deployment best-practice to implement ethical culture and practices within organisation

Background:
To deliver the candidate recommendations in this section, and many other candidate recommendations in this document, will require a transformational change for most organisations. Without guidance or best practice that mitigates risk to shareholder perception, brand and market position organisations will be reluctant to undertake this journey with the limited budgets and resources available.

Success in implementation of the recommendations is more likely with established guidance and best practice on delivery, thereby reducing perception of risk by corporate boards and increasing likelihood of enacting the required changes before being forced by incident or direct threat to business to take reactive measures.

Candidate Recommendation:
There should be ongoing cooperation between standards organisations, major corporate entities, and professional organisations to establish a Body of Knowledge for transformation to an A/IS organisation.

Further Reading
http://www.opengroup.org/ - an example of cooperation between competing corporates to form operational best practice

https://hbr.org/2018/03/why-so-many-high-profile-digital-transformations-fail - a review of the challenges faced in performing digital transformation in areas with little established best practice

---

Page: 60

Section: 2- Corporate Practices and A/IS

Comment: Recommend including an additional Issue recognising potential ethical issues with the use of A/IS within corporate environments

Issue

The use of A/IS technologies to manage a workforce, improve productivity, improve performance of business processes or reduce cost may present ethical concerns. In particular the imbalance of power between employer and employee, or between hiring company and applicant, can compromise the ability of the employee or applicant to opt-out or otherwise demand more equity in decisions to use such tools.

Candidate Recommendation

Employers should voluntarily adhere to minimum standards and best practices to protect the rights of the employee or applicant in the use of their personal data and application of A/IS systems in a manner that may have ethical issues. (Standard P7005 is under development to provide such guidance).

Candidate Recommendation

Companies should ensure that ethical concerns are fully addressed during internal projects to implement new A/IS systems; to change business processes utilising automation; to automate current activities; or during business re-organisations, post-merger integrations or other changes to the organisational form.
Such transformational projects should include specific success criteria related to ethical treatment of employees and contracted staff, ensuring systems are implemented with reference to an relevant standard, ensuring appropriate and controlled use of A/IS during change activities and targeting and equitable outcome for those impacted by the change. These success criteria should form part of the project governance and receive ring-fenced budget for execution.

Further Reading

The workplace of the future - An examination of the changes to the workplace that raise ethical questions on the treatment of employees

They're Watching You at Work – An examination of the use of people analytics in the workforce.

Page: 61

Section: 2- Corporate Practices and A/IS

Comment: In Issue "Lack of values-aware leadership" recommend the inclusion of Candidate Recommendation to give guidance on using corporate scorecards, or similar, to enable the other Candidate Recommendation in this Issue, as well as promote a corporate environment that reinforces the aims of this document.

Candidate Recommendation:

Companies should conduct a review of their policy framework to include specific policies on the ethical development, offering and use of A/IS. The trade-off between these policies and other specified corporate goals should be discussed and handling recorded as part of the policy description. Companies should cascade these policies into their overall quality management systems and internal standards. Performance scorecards should be aligned to recognise these new policies and the reward framework changed to bring into line.

Further Reading:

Sustainability Balanced Scorecard and Business Ethics – An approach to including sustainability focused ethics into corporate scorecards

---

Page: 63

Section: 2- Corporate Practices and A/IS

Original Text: Those who advocate for ethical design within a company should **not** be seen as innovators…

Comment: Paragraph 1 in 'Background' section. Believe the word "not" here is incorrect.

Page: 63
Section: 2- Corporate Practices and A/IS
Comment: Recommend an additional Candidate Recommendation that includes recognition of the role employee review systems and bonuses play in reinforcing behaviours.
Candidate Recommendation:
Companies should evaluate their performance review systems/processes to identify potential reinforcement of behaviours likely to dissuade the raising, discussing and addressing of ethical matters in the development of products or the use of A/IS within the business.

Further Reading:
https://link.springer.com/article/10.1007/BF00872018 - Study of the role of performance management and control systems in encouraging unethical behavior.

As I've read through various comments, overall comment, morals are internal or personal based on personal understandings and beliefs, and beliefs are perhaps driven primarily by one's up-bringing. EAD shall probably be a collective perspective that is influenced by individual cultures.

EAD development principles shall be aligned for the betterment of a society, be theologically agnostic or indifferent, and help enable reduced human intercession.

Randy k Rannow

University of Applied Sciences and Arts
Northwestern Switzerland
School of Business
Prof. Dr. oec. HSG Oliver Bendel
Bahnhofstrasse 6
CH-5210 Windisch/Switzerland
phone (direct) +41 56 202 73 16
oliver.bendel@fhnw.ch
https://www.fhnw.ch/de/personen/oliver-bendel

Let me first emphasise that the document is the result of excellent work by outstanding experts.

An Internet research shows that theological ethicists and representatives of religions are involved in the initiative. These have already infiltrated disciplines such as business ethics and media ethics in the past. I am deeply concerned about the religious part of the document and consider it to be the main weakness. We should defend science and not mix science and religion! I propose to create transparency regarding the background of the participants.

Different practices are mentioned on page 2. I would not start from the classical practices or concrete values, but from the theories or disciplines. We should make a clear distinction between Western, Eastern and African traditions. As far as I can see, only Western philosophical ethics is a scientific ethics. Western philosophical ethics is in no way limited and can produce normative models of all kinds.

Page 48 mentions a book I contributed to: Rötzer, F. ed. Programmierte Ethik: Brauchen Roboter Regeln oder Moral? Hannover, Germany: Heise Medien, 2016. The predecessor book from the same publisher is not mentioned here (but on another page): Bendel, O. Die Moral in der Maschine: Beiträge zu Roboter- und Maschinenethik. Hannover, Germany: Heise Medien, 2016. I'm asking you to put it on the list in this context. Also this book is a contribution to the fields of information and machine ethics: Bendel, O. 300 Keywords Informationsethik: Grundwissen aus Computer-, Netz- und Neue-Medien-Ethik sowie Maschinenethik. Wiesbaden, Germany: Springer Gabler, 2016.

On page 171, you write: «While robots capable of participating in an intimate relationship are not currently available, the idea that they could become intimate sexual partners with humans (e.g., sex robots) is one that captures the attention of the public and the media.» There are several products on the market, both actual sex robots and advanced love dolls. This section is full of concerns. The advantages of sex robots are hardly seen. Prejudices are expressed and preconceptions made. Several important books from communities that are neutral or open to sex robots are missing, for example: Cheok, Adrian David; Devlin, Kate; Levy, David (eds.). Love and Sex with Robots. Second International Conference, LSR 2016, London, UK, December 19–20, 2016, Revised Selected Papers. Cham, Switzerland: Springer International Publishing, 2017. pp. 1 – 10. and Cheok, Adrian David; Levy, David (eds.). Love and Sex with Robots. Third International Conference, LSR 2017, London, UK, December 19–20, 2017, Revised Selected Papers. Cham, Switzerland: Springer International Publishing, 2018. pp. 1 – 11. These books also contain contributions from the perspective of ethics, for example machine ethics.

Hello,

I commend the contributors for their efforts and I thank the IEEE Global Initiative for giving me the opportunity to make a contribution. Following are my comments.

Purpose or role of the EAD and P7000 series standards

Will the purpose or role of the IEEE Global Initiative become aligned with any national or global policies, or other standardization initiatives? A lot has happened In the past few months, in the US we now have the Congressional AI Caucus, the Future of Artificial Intelligence Act, and a bill that would create the National Security Commission on AI. In the UK the The Lords select committee on artificial intelligence released its latest report focusing on ethical implications. Last March, the government of Quebec suggested instituting a Montreal-based international organization responsible for developing the necessary frameworks for AI. On 25 April the European Commission presented its strategy paper on artificial intelligence - namely, that the EU should take the lead to shape the ethics of AI.

Standard notation defining A/IS

A standard notation for A/IS, such as the Open-SGI Guidelines for Definition of Intelligence Objects, would support many if not all recommendations relating to transparency, accountability and testability. This notation could be used to define the object classes associated with standardized A/IS functional areas. Recommendations in the final version of the EAD, and any standards or certification programs that come out of the 7000 series working groups, could be expressed in terms of standard functional areas and their associated object classes.

Testing, certification and oversight

I believe it would be useful to have a theme or principle in the final version of the EAD that addresses testing; certification and on-going conformance; and perhaps a P7000 standard for testing as well. In reading the EAD my impression was that there should have been more testing and certification related candidate recommendations. A search for "test" or "validation" finds the following 5 cases where there is mention of compliance by some validation or certification agency: Transparency, Embedding Values Into Autonomous Intelligent Systems,

Personal Data and Individual Access Control, Reframing Autonomous Weapons Systems and Law Section 4 Transparency, Accountability, and Verifiability in A/IS.

Regards,

Randall Parker
Open Simulated General Intelligence (SGI)
www.open-sgi.org

Bruno Macedo Nathansohn
Perspectivas Filosóficas em Informação (Perfil-i)
Brazilian Institute of Information in Science and Technology (IBICT)

Classical Ethics in A/IS committee
In page 203, I recommend adding the following resource

Issue

Implications of cultural migration in A/IS

Background:

Today there is a global migration dynamics, which creates within public opinion an idea of crisis and emergency, far from what statistics have realistically shown. In this sense, the so-called Autonomous and Intelligent Systems (A/IS) are designed and applied to measure, calculate, identify, register, systematize, normalize and frame both human rights and security policies. This is exactly what has been done since the period of colonialism. This includes the creation and implementation of a set of ancient and new technologies. Along the history, mechanims have been created firstly to identify and select individuals who share certain biological heritage, and secondly, individuals and social groups, including biological characteristics.

Characteristics that mark permanently social condition, related with political agendas determined through the management, measurement, calculation and statistics, can be visible, since 12th century when the signals stamped on the bodies (About & Denis, 2010). In the XVIII century, surveillance was expressed by Benthamian Panopticon as a disciplinary dispositif to see without being seen. Biopolitics gained terrain as an ethical question, in which every normatization process should be considered. Biological identification mechanisms have been implemented together territorial marks, enriched by digital resources, as shown in the use of e-passport, walls, online watch systems monitored by internet (i.e. Texas Watch), and so on.

Information is only possible when materialized as an infrastructure supported by ideas in action as "communicative act", which Habermas (2011) identifies in Hegel's work, converging three elements in human-in-the-world relationships: symbol, language and labour. Information policies reveal the importance and the strength in which technologies influence economic, social, cultural, identitary and ethnic interactions.

Traditional mechanisms used to control migration, like passport, is associated with walls and fences increasingly built around the globe. More intense is mobility, more amplified are the discourses to discourage it, restricting human migrations, and deeper is ethics related to the condition of citizenship. Together the building of walls, other remote technologies are developed to monitoring and surveil borders, buildings, and streets, and impacting the idea of citizenship. Closed Circuit Television (CCTV), UAV (Unmanned Aerial Vehicle), and satellites allow data transference just in time to databases and internet backbones hosted in developed countries. This centrality expresses a devide between developed and underdeveloped countries.

The dispositives are built through the State bureaucracy. But, in paralel, human rights are increasingly presented in national lives through treatises and laws have been adopted in constitutions, since the Chart of San Francisco (1948). This double perspective provokes a dilema within the information is produced on the migrants, mainly on refugees. Therefore, questions around citizenship go far beyond Kantian universal principle of peace. Here, universality is linked to the Western civilization, in which prevails a hegemonic discourse followed by a political agenda based on new ways of colonialism. This problem change the conception of morality, realocating the locus of the "Categorical Imperative", because the tension among different social and political contexts are more pervasive, and because of that imposes a consideration of morality in which human being-in-transit is pivotal. Digital technology systems used to register and identifying human mobility and the "undesirable" refugees are not autonomous or even intelligent in this sense.

Technocracy, as posed by Habermas (2011) works as a set of rules, norms, and management that potentializes the capacity to control information as a way to monitoring spaces, borders and flows apreehended by central capitalist powers. In this sense, it works as a conception about who can or not being pemanent resident.

Thus, using Capurro's perspective about intercultural ethics (2008) it can be found a converged viewpoint with the work of the Philosopher Leopoldo Zea (2005), in which is explored the discussion between civilized x barbarian relationships. In this aspect, language is the locus where this dichotomy has to be considered to understand the diversity of morals when there are contacts among different cultures.

To Zea, despite Latin America inherited European system of thought, it is identified contributions from other native philosofical sources that have been shaping our way of "being-in-the-world". Process that gave birth to human rights instruments, like"*Instituto de Asilo"* (diplomatic asylum), in the legal frame of American System of Human Rights. Besides this, the forge of a Latin American civilization is also marked by instruments of repression based on military, media and Judiciary system.

This is clearer in Brazil, where it has been used an increasingly resource identification and registration technology to produce profiles and classify the indisirable people. Traditionally, the country is guided by the objectives drafted in security staff, and still being like that, such as seen in the role of Federal Police within the proceedings for refugees eligibility. Using discretionary power to justifying their authority as a supreme representative of territorial security, this agency not only control the immigration flow of people through the flow of data, but also participates in the main board of eligibility decision making, which is the National Committee for Refugees (CONARE – Portuguese acronym).

Despite this role does not impede the eligibility proceedings, it can curb some asylum seeker cases based on discretion of political actors such as policemen, and the conservative opinion of conservative politicians and journalists. In this case, what prevails is the prejudice about groups legimated by the system of classification. Information is something that links languages, habits, costumes, identification and registration technologies. This provokes a reshaping of the immigrants and refugees citizenship, who seeks many forms of surviving in, and against, the restrictions imposed by A/IS for surveillance and monitoring in an enlarged and more complex cosmopolis.

This seems to be the first step to consider if systems can be really autonomous and intelligent, without the human guidance. The question about the possibility of a morality out of the individuals and the relationship of each other resides in a discussion that involves the dominance of some discourses and narratives over others. Thus, actions devoted to humanitarian treatment as well as those dedicated to the control of refugees is part of a process of eligibility.

Candidate Recommendation

It is recommended to contemplating how inequalities among regions, considering the gap among societies and the concentration of capital and goods design human migration. In its turn, it is suggested to thinking about in which level migratory dynamics can open perspectives to think new ways of using A/IS mechanims for surveillance in opposition to human rights demands of inclusion. In this sense, emerge other conceptions of citizenship in multilevel scale, which have to take into consideration different impacts in global, Latin-American and Brazilian context.

Further resources

About, I; Denis, V. Histoire de l'identification des personnes. Paris: La Découverte, 2010.

About, I.; Brown, J.; e Lonergan, G. [orgs.] Identification and Registration Practices in Transnational Perspective: people, papers and practices. RU: Palgrave Macmillan, pp. 1-1. 2013.

Capurro, R. (2014) Citizenship in the Digital Age. In Information Ethics Roundtable 2014: organized by the School of Library & Information Studies, University of Alberta, Edmonton (Alberta), April24-26, 2014. Available in: http://www.capurro.de/citizenship.html. Access: 03/04/2018.

Capurro, R. (2007). Intercultural Information Ethics. In R. Capurro, J. Frühbauer & T. Hausmanninger (Eds.), *Localizing the Internet: Ethical Aspects in Intercultural Perspective* (pp. 21-38). Munich: Fink. Available in: http://www.capurro.de/himma.html. Access: 03/04/2018.

Habermas, J. A inclusão do outro. São Paulo: Edições Loyola, 2002.

Habermas, J. Técnica e Ciência como Ideologia. São Paulo: Editora Unesp, 2011.

Johnson, R.; Cureton A. Kant's Moral Philosophy. Stanford Encyclopedia of Philosophy. July 07, 2016. Available in: https://plato.stanford.edu/entries/kant-moral/. Access: 03/04/2018.

ZEA, L. Discurso desde a marginalização e a barbárie – seguido de A filosofia latino-americana pura e simplesmente. Ed.: Garamond, 2005.

Comments and Proposals concerning ead_v2

Manfred Bürger, Stuttgart, Germany
(physicist, retired leader of nuclear reactor safety department at IKE- Inst. Nuclear Energy, Univ. Stuttgart), jointly with InkriT
(http://www.inkrit.de/neuinkrit/index.php/en/) and
FST (http://sustainabilitymaker.org/partner/forum-soziale-technikgestaltung/ ,
http://www.blog-zukunft-der-arbeit.de/)

The present document is a compilation of contributions after ead_v1 and the rfi_responses_document and, based on these, comments and proposals on ead_v2. The comments under 1) are the basis for the continued considerations. First considerations on ead-v2 and the well-being subject follow under 2). Then final comments and proposals of text modifications in ead-v2 are under 3). My emphasis lies on the well-being parts. I consider well-being (good life, human orientation) together with human rights as the overall perspective addressed by the ead-documents, with the specific goal that A/IS shall serve humans and human perspective, therefore must be used and designed in this perspective.

1) Comments and Proposals on "rfi_responses_document" and the related document "becoming_leader_global_ethics" (MB,27.9.17)

A strong emphasis lies in both reports on the consideration of ethical basics, cases of ethical dilemma and decision questions, on the definition of an ethical concept adequate to the upcoming technically induced problems posed by AI/AS and a transfer of decision making from humans to AI/AS. Extreme situations are in the focus which pose difficult decision questions as in trolley problems, where human empathy, not only efficiency based on given rules, is required. Responsibility for nature, preservation of environment is addressed, while in view of industrial needs and growth as basis of human welfare related restrictions are questioned (e.g. concerning transparency and control versus free learning in AI/AS). The commissioned resulting report emphasizes even more basic ethical questions.

In my view, this emphasis and the corresponding consideration lines go in a wrong direction which will not help to get access to the development driven by technology and economy and to support, elaborate and put forward a human line within it.

Firstly, the generalized and also pointed dilemma views on ethics are in conflict with an ethics which in my view has to be derived from and related to practical life, based on social interactions. Since in modern societies a given ethics, coming from some superior might, meanwhile even from experts, is no longer accepted, this can only be based on discourses and interplay of people. The achieved generalization of human rights is also the result of historical processes. These rights can be considered as a basis, but being too general, are not sufficient to deal with the new developments and challenges in a human-oriented way.

A reason for this and major deficit of the reduction to basic ethical considerations is, secondly, that these are not only to be applied to manage certain situations and dilemma cases. Rather, the social conditions in a society are to be considered as a basis for ethics and ethical behavior. A thinking has to be supported with this respect about the goals of society and the ways to go, especially with the new developments. This means to pose questions about what good living means and how it can be realized. As a pointed example: Posing questions on assisted suicide in cases of severe illness should not be isolated from questions of sufficient care which depends on institutional conditions, employment, education, etc.

The real major problem is not to adapt the machine behavior to human behavior, rights and their ethical basis but to find ways to use the technological development and the possibilities it creates in a human-oriented way. For this, the questions have to be posed anew, how we want to live, how we want to use machines etc. Discussions and conflicts around this should be emphasized. Thus, the major task should be to enable people and society for these discussions, to promote these discussions etc.

In my understanding, the EAD initiative of IEEE is not restricted to analyses of specific ethical problems in AI/S, but has a wider perspective and concerns the questions of human and social development. I.e., conditions under which certain problems with AI/S are raised come into play. E.g., in case of driverless cars, other solutions as an integrated public traffic system would avoid or reduce such problems, in addition to solve environmental and energy problems. If human care is promoted by sufficient time and human employees, the needs of robots can be reduced to selected support rather than just replacing humans, thus less strongly

raising questions of human-robot interactions and related ethics. If production can be oriented at human needs instead at imperatives of economic growth, the problem of violating nature is reduced.

It is important that IEEE as largest professional association of engineers has started an initiative to induce reflections on the context of work and responsibility concerning the new developments. For promoting such reflections between engineers about their work and not just having ethics as an additive (see my comments e.g. on p. 27 of responses report), it may be important to gain a more practical view. This appears hardly to be possible by generalized ethics discussions.

On the other hand, a reduction to technical problems or technically transformed ethical problems as reverse of high ethics is also visible in the contributions. This is also not sufficient to meet the challenges. Technical solutions alone will not solve the problems raised. These and individual ways to gain control (even only on data) and to avoid risks collide with increasing complexity and options of failures and abuses, which requires joint solutions. Only socially based solutions can help.

Keeping the goals of human-oriented production and society directly means that human control in production and society must be maintained. Even, it has to be re-established against social, economic and technical systems which by their inherent mechanisms control goals and behavior of people (especially production under capitalist and competitive conditions where the products of social production appear to govern and steer the processes instead of steering by the producing society). Thus, the task is not to transfer human goals and behavior to machines. The task is rather to establish a technical and social system which does not violate human goals. Of course, if decisions are given to machines as in the case of driverless cars or other autonomous systems, this means to implement rules. To implement real human behavior including empathy (against only efficiency related to rules), as partly required, has the tendency to give the control to machines. Instead, the question of human control is the most important issue, control including decision about goals and means, finally about the design of society.

Thus, not "Establishing Ethical Principles for AI and AS", now the title of the conclusion document, gives the major task, but promoting ways to design society and processes in society in view of the new challenges and possibilities and to

consider ways to keep human control. Especially, the new options, chances as well as requirements from the technical development give new impacts on discussion of human perspectives and control. Does the technical change yield a loss of human influence or are there just chances to gain control?

In my contribution to the responses document, I tried to emphasize such questions and develop some answers. For a short recapitulation of major thoughts:

- Concerning production processes, a major question is whether AI/AS replace human work and thus which role humans can/should play in future, further, which problems are created due to job losses and how they may be compensated. Determination and control of work and life by machines, data and systems is a further, related issue.

  The key change in the working process is that machines can now adopt most hand working, further steering of production, also administrative work, even communication tasks, interactions with humans, even supporting care. However, as long as AI does not completely replace humans, there remains a human steering influence in the increasingly complex processes of production, human relations and society. The human task goes more and more into designing the processes and defining aims and goals. This can be understood as increased societal character of work and being. All things and human affairs have more and more to be considered as interrelated.

  Human design of interrelated processes becomes more and more important with the new technics. This is firstly due to technics itself: complexity and self-regulation requires design and control against failure, requires systemic optimization. This requires not only theory and understanding but also reflected experiences on the processes which can only be gained in permanent involvement, permanent active control and checking by teams which permanently survey and reflect processes, not only passively control according to defined technical rules and procedures. Since this requires direct involvement, hierarchical steering becomes less effective, even less possible (reason for replacing management concepts of direct instruction for partitioned process parts by such giving goals for more integral fields). This is the basis for requiring cooperative work. Complex

processes cannot be adequately steered and controlled from above. Also, steering software needs failure control and optimization, needs experiences.

A culture of cooperative reflection is to be based on this way to deal with the new technics, with AI etc. It must be developed from such needs and discussions about them as anchored in the changed industrial processes.

- These lines and corresponding tasks are to be extended to the whole society, which implies a re-orientation of processes towards democratic procedures. The permanent design of all processes in their increased interrelations requires this. Failure and optimization considerations include awareness of risks, thus a safety culture. Of course, different, even conflicting views are important, as for cooperative design and control processes in general. Decisions about goals and processes are to be based on broad participation in design of the society instead of determination by economical and technical forces, implemented systems and executing experts.

For this, the technical changes yield a basis due to increased technical possibilities as well as due to the general options of high productivity and automation, namely the reduction of necessary working time. This yields options to involve all people more in designing the social processes, in contrast to the threat of loss of work. General involvement is also necessary to manage in a democratic way the increasing risks as well as the combined chances due to the new technics with increasing capabilities to intervene deeply in all aspects of life.

Finally, the interrelations imply the need to jointly define: where shall the society go, what to be produced, what is really wanted? E.g., more time for human relations, for participation in deciding about such questions, for engagement in cultural interests, in learning about nature and human affairs, more human work in care for humans and nature instead of leaving more and more to machines, developing machines for this, - or vice versa? What do we want to be replaced by technics, what do we want to keep in human activity? Technics replacing everything, means for new fascinations, always new enjoyment, or support to relax work to be able to develop own interests, human relations, etc? How do we want to live?

Such questions are increasingly raised by the development itself which pushes questions how to design work and society due to the design needs and options it produces, **needs** coming from the increasing interrelations and interdependencies and also the deeply intervening, invasive character of this development into nature and humans, into all fields, **options** due to the rich field of possibilities. These are not primarily ethical questions, but questions how we want to live, to live together. …
Decisions on such questions cannot come from ethical considerations, especially not as delegated to experts, philosophers, engineers etc. This is the task of the whole society and one has rather to think how society, democratic society, can be enabled to perform this task in a democratic way. We cannot leave this to capitalist rules, powers of institutions or experts (often involved in such rules and powers). Ethics must be put down to real life, i.e. to the questions really posed by it.

Solution perspectives in these directions concerning both, production and society, are being developed in the frame of InkriT (http://www.inkrit.de/neuinkrit/index.php/en/, e.g. PAQ - Project of Automation and Qualification - already in 1970s, and the vision of 4in1-Perspective http://www.friggahaug.inkrit.de/documents/4in1_englisch_fin.pdf), more basically, and, more in realization, in the frame of FST (forum for socially sustainable design of technology: http://sustainabilitymaker.org/partner/forum-soziale-technikgestaltung/, see also: http://www.blog-zukunft-der-arbeit.de/, linked to the German Federation of Trade Unions, and the initiative of a social network on social cohesion in digital world (e.g.: https://kunstgebaeude.org/intro/en/bauhaus-forum-civil-society/ http://www.ev-akademie-boll.de/hu/tagung/210518.html).

The vision of 4in1 aims at participation of everybody in all main fields - production, reproductive work, culture and politics - which are essential for designing the society. Struggling for this can start with requiring cooperative work in technically changed (by AI, IoT, etc) production processes as indicated above, requiring for permanent control and optimization sufficiently large working groups, thus - against replacement of humans - really humans in the center of controlling and designing the processes … Struggling for this may also aim directly at the decisions about development of society in view of the deeply intervening changes.

Even rules and committees considered to control processes will not avoid failures of technologies and abuse if social control, a culture of awareness does not exist or is not effective. Thus, establishing such a culture is the major task.

2)First considerations on EAD-V2 as well as on Well-being subject

(MB, Febr. 2018)
My impression is similar to that outlined in my comments on EAD-V1:
Based on the general aim to promote **benefit from A/IS** (Autonomous and Intelligent Systems) to humanity and nature, to mitigate risks and negative impacts, main considerations concern implementation of **ethical rules** or in somewhat reduced way (due to problems to determine ethics) **norms** agreed in a society for a certain context (ead-v2, p. 37) into A/IS. By this, A/IS shall act in human interest.

However, this appears to become both rather general and mostly technologically determined, even more than in V1, now by trying to specify things concerning changing norms e.g. due to changing conditions. Then, the need for norm updating is transformed into a need that A/IS must be enabled as a learning system "to be capable of identifying and adding new norms to its baseline system" (p. 39). I.e., essentially the system is to be programmed to do this itself. Only as correction, measures of communication for performance appear, to be implemented as support in A/IS, as "asking for guidance … when uncertainty … exceeds a critical threshold". Other correction options are considered to be based on communication with humans, again to be initially implemented, and the system may document and inform about norm changes.

In my view, this is a much too passive way to deal with such systems in mainly two aspects:

1) The systems appear already to be established, as self-driving cars or robots for healthcare and care of old or ill people, not considering any more what is really needed or wished in order to improve human living conditions.

2) Handling these systems by transferring control to them and/or to experts of norms, technics and programming, delegates human influence and democratic

means to design the process, delegates finally democracy. The remaining major questions appear to be, whether human control can be established in the initial construction and programming and afterwards by control and correction procedures which usually get technologically determined.

Instead, the emphasis should be to build teams, communities, society in a way to be able to control and design the processes, the development permanently. This is the major task, not the definition of the norms to be implemented in A/IS and the technical realization with adequate updating options. The task is to build a society which is able to manage the increasing technological options with their increasing intervention capabilities into human life, humans and nature, in a human perspective to be developed.

E.g., concerning healthcare, major questions are how teams should be organized (number of employees, qualification, organization of work, cooperative structure, support etc), how their concept of work is (relations in group and to their clients etc), their social engagement, and concerning which tasks they consider A/IS tools could be a support. Only then and oriented at decisions about this context, related to this, it makes sense in a human perspective to consider ethics and norms for robots and A/IS, to consider adequate programming and updating. This must then always be oriented at possibly changing evaluations of the groups and must always be under control of them, not determined by self-changing of machines and software. A societal, overall perspective to be developed beyond single considerations of groups must be based on societal discussions including and based on the working groups, the involved people and the whole society. Initiatives to promote such discussions and clarification processes especially now, being confronted to introduction of A/IS, should be initiated by IEEE in cooperation with other organizations.

This is similar in other fields, e.g. the always cited example of self-driving cars where basic decisions on how to deal with transport systems, globally and locally, are required and interests of people have to be discussed not only based on given systems of private cars and consumer views induced by car producers and their interest in development.

An overall view beyond given orientations can be provided by **well-being** considerations criticizing and transcending views based on given capitalist, market and resulting growth economy. A step in this direction is to question GDP as typical measure for this economical orientation and asking for different measures of well-being.

However, again it remains highly problematic to reduce this questioning and considering about well-being, good life etc. in an again technocratic way to a quantification by even single – although combined - indexes. This can also not provide a real perspective to evaluate the chances which A/IS may yield for well-being and human development. Again, it is necessary to come into a deep and broad discussion about what could be helpful and what not – chances and risks (not only concerning safety problems, big data etc., but concerning human perspectives at all).

How problematic index considerations are, can directly be seen from results of different approaches. These differ strongly, e.g. BCG report 2015 with the SEDA index yields Poland as top performer – absolutely in contrast to what we consider in EU context - , the Happy Planet Index (HPI) of 2016 shows Costa Rica on top followed by Mexico and Colombia … Even if there would be some truth in this (e.g. due to a major effect of division by ecological footprint in case of HPI), I doubt that such results will have any wished influence to support considerations on well-being, on orientation towards conditions of better life, on evaluations about reasonable use of A/IS for this.

Rather, I consider a discussion about qualitative aspects questioning GDP and promoting other views and orientation as important. Looking at the positive effects of catastrophic events as hurricane Katrina in GDP certainly illustrates best the failure of GDP (see Talberth et al: GPI indicator 2006) even as measure of economic welfare. Indicators of satisfaction, life expectancy, various inequalities between people and ecological footprint as applied in the HPI or in a more detailed hierarchy of criteria in SEDA may help to develop a different view (SEDA: Economics addressing Income, Stability, Employment; Investments addressing Health, Education, Infrastructure; Sustainability addressing Social Inclusion - Income equality, Civil society, Governance - and Environment, each with further specifications, see BCG report 2015, p.16).

By this, considerations about an alternative orientation of economics and use of technics may be promoted in general. But even this will not be sufficient to really develop alternatives. This can only be done by considering real changes of conditions of work (cooperative structures with joint evaluations about processes, corrections etc), of care (see above), of transport organization, of energy and resources supply, their use and management, of consumption and/or active participation in planning and design of society. Only in this way, human orientation for use of A/IS can be developed.

Again, IEEE should attempt to initiate such concrete discussion processes in relevant institutions as well as in the whole society. Due to networking, interrelations, integration as tendencies in new technics the basis for joint societal design is developed in principle.
If in such processes alternative indexes are used to put something against GDP, it should be further checked, how this could influence political and economical decisions.

3)Comments and Proposals for ead-v2

The above considerations are the basis for proposals of modifications here.
The additions are given in red, comments green, deletions are not reproduced.
Ad Executive Summary, proposals, p.6 ff:

I. Purpose

… effects on social well-being. In general, relief from necessary work and liberation to creative work can be taken as promises of the new technologies, while increasing risks exist in their high penetration potential into human nature, society and environment, nature in general. Because of their nature, ~~the full~~ benefit of these … Keeping human control and design must be the overall perspective.

II. Goals

…

• Well-being: Prioritize goals of well-being

III. Objectives

Comment: Well-being as overall orientation should be first, then "Personal Data Rights…"; "Well-being …" replaced by:

Well-being to be Promoted by new technics (A/IS)

In order to use the new technologies for the benefit of humans and society, for humanity in general, permanent human control and design is essential. This cannot be realized by just leaving these technologies to given economic structures, mechanisms and goals, especially market-defined competition, profit and economic growth goals, the latter expressed by GDP. A shift to other thinking about goals of society is necessary, as expressed by goals of well-being, good life and happiness, pursued by initiatives highlighting these goals by alternative indices of well-being than GDP. Such goals orient at non-destructive effects (destructive ones positively accounted in GDP metrics), at sustainability in society and nature, equity and participation in design of society, at good work and living and at happiness. The technologies themselves provide means and requirements for such use.

Legal Frameworks for Accountability

at the end:

Beyond these regulations, it is additionally required to develop a culture of joint, societal awareness, control and design (defining purposes of A/IS use) in order to keep human aims and control. Otherwise, it will be difficult, in view of permanent development of extended A/IS, to even detect deviations from decisions and operations which should not be delegated (see automation of decision-making by self-improving algorithms – following section).

Transparency and Individual Rights

…

- …rigorous testing. Processes for this are to be developed.

at the end:

As stated in the above section, more is finally required than transparency, rigorous testing and third-party verification, namely involvement of the whole society.

Policies for Education, Awareness and Intervention

…

• Develop workforce expertise in related technologies as well as working processes allowing and asking for intervention and participation in design and control for workforces.

…

• Develop and organize participation of society in decisions about technologies.

IV. Foundations

Classical Ethics

…

Well-being Concepts and Metrics - section replaced by:

A basis for thinking about well-being concepts results from the development of technologies itself, the production potential and the general wealth in industrialized countries as well as the inherent contradictions in this development. Under capitalistic conditions, this development produced, besides the high productivity and wealth, high and even increasing inequalities, options of creative and good life together with limitations and bad conditions for large parts of humanity, even under richness loss of sense and orientation, with compensation by consumption offers, yielding problems for individuals and society. Questions about sense of orientation at economic growth and about its driving forces resulted from this as well as from increasing risks with reinforced growth visible by technical disasters and increasing environmental damages.

Based on these experiences, e.g. thinkers as Martha Nussbaum and Amartya Sen formulated alternatives. Shifting from quantitative to qualitative economic growth came into discussion, promoted also by the Club of Rome already in the beginning of 1970es, questioning depletion of resources and environmental burdens. In order to put something against the GDP metrics, various approaches to replace this by a different metrics are being developed, accounting negatively for destructive effects and positively for qualities of life. The new technologies are also addressed in these considerations and approaches.

Common metrics of success include profit, occupational safety, and fiscal health. While important, these metrics fail to encompass the full spectrum of wellbeing for individuals or society. Psychological, social, and environmental factors matter. Wellbeing considerations and metrics aim to capture such factors and thus to better evaluate benefits arising from technological progress as well as negative

consequences that could diminish human wellbeing. New routes to societal and technological innovation could thus be provided.

Embedding Values into Autonomous Systems

To add at end:

The tasks considered here address special attempts to design autonomous systems. This does not replace needs to decide on the use and kind of use of such systems. Permanent human control must also be guaranteed for which structures and a culture of designing and intervening are to be developed in the working processes and in society.

Methodologies to Guide Ethical Research and Design

To add at end:

These methodologies have also to address the development of structures in society which are necessary to use the systems for human well-being, as indicated in the previous sections.

V. Future Technology Concerns

Mixed Reality

To add at end:

Mixed reality becomes an issue since activities with computers and information techniques become an essential, even dominant part of our world. This is basically linked to the increasing importance of steering work, of system character of technological development. Dominance of intellectual work results while handwork and even large parts of intellectual work (not only repetitive ones) are more and more replaced by intelligent machines, with remaining need of steering, design, programming and control. Problems arising from this development are to be solved by corresponding structural changes in society yielding real liberation, emancipation of the whole society to steering and designing, by this creating identities, identification with the societal processes, not as separated identity problem of individuals.

Ad General Principles, p. 20/21, proposal to add at end:

In an extended perspective, it is to be considered that technical solutions-by-design cannot cover the field of problems opened by the use of A/IS, nor can this be reduced to ethical concerns about actions of A/IS. A/IS yield chances and risks for

human perspectives which are to be explored and elaborated in societal discussions. These have to decide about products and systems wished, the extent and the way of using them, the specific support considered for human perspective, for well-being of human life and society, already before considering specific ethical implementations for specific technics. Solving ethical concerns and gaining orientation towards benefits for humanity cannot start with given technical systems.

Thus, principles to be considered for use and design of A/IS start in the present document with human rights which are to be kept or even deepened and supported by A/IS, then the orientation and priority at well-being (better inverse sequence). The following principles concerning accountability, transparency and awareness of misuse can be considered as means in favor of human rights and well-being. They also need more than technical implementations in A/IS, namely structures of orientation, decision and control in the society, a respective culture of cooperative work and design. Without establishing such a culture, attempts to guarantee ethical principles by implementation in technical systems can finally not be successful.

In the frame of the present Global Initiative, launched by a technical community, the required change processes in society to adequately address A/IS cannot be in the center. In view of this, technological aspects and solution thinking may necessarily be over-emphasized and the solution attempts may partly run into limited optimization lines which miss overall views. However, the needs to widen the view are in general recognized which also means that the need is in principle seen to embed the initiative as a collective action in a wider perspective considering processes in society. This is e.g. expressed by the new section and standard work group P7010 on well-being directly addressing and orienting at such a wider perspective.

Principle 1 – Human Rights, p.22:
Proposal to add a sentence before Candidate Recommendations:

As outlined above, this assessment must be embedded in a culture of awareness to be developed and established in general and with institutions controlling and balancing each other, as e.g. described by a paper of Prof. Theofanous on risk management in the nuclear reactor field and in general.

Ref.:
The 10th International Topical Meeting on Nuclear Reactor Thermal Hydraulics (NURETH-10)
Seoul, Korea, October 5-9, 2003
Keynote Lecture
RISK, SEVERE ACCIDENTS, AND THERMOHYDRAULICS
T.G. Theofanous
Center for RiskStudies and Safety, University of California, Santa Barbara, USA
Tel: (805) 893-4900 — E-mail: theo@theo.ucsb.edu

Principle 2 – Well-being, p. 24 – 26, Comment:
In general, I have the impression that the formulations go too directly to state that quantitative measurements of well-being are the alternative to GDP and thus have the tendency to reduce the task of formulating alternatives (finally vs. mechanisms of development given by capitalist profit, competition and growth orientation connected in technocratic way with technical development) towards another technocratic perspective in which qualitative alternatives and decisions about concepts cannot be discussed adequately. A sense of this is formulated in the second to last paragraph under "Background" on p. 25: "Nonetheless, quantitative indicators of individual well-being should be introduced with caution, as they may provoke in users an automatic urge for numerical optimization ".This concern is not really solved by the last statement in this paragraph.

p.24 (in red my proposals of modification):

Issue:
Traditional metrics of prosperity as GDP do not capture human well-being as a whole and the effect of A/IS technologies on human well-being.

Background
Proposed replacement:
…That question is: What is the societal success for "ethical AI" once released to the world? Or formulated in a wider perspective: How can A/IS yield benefit for humanity and contribute to human well-being?

At first, this means to define these categories and the orientation implied and this means to determine how we want to live, what we consider as good life. It is essentially a task to determine qualities. Correspondingly, well-being, for the purpose of *The IEEE Global Initiative*, is defined as encompassing human satisfaction with life and the conditions of life as well as an appropriate balance between positive and negative affect. Although it is there already connected to a measuring task, this definition is based in principle on the Organization for Economic Co-Operation and Development's (OECD) … life alongside other social and economic dimensions."

Rest of section replaced:
Since common considerations on well-being largely and traditionally relate to consumption levels and especially their increase in Western industrial countries after 2nd world war, measuring by GDP was introduced and became familiar. Economic crises, destruction of living options due to work stresses, loss of orientation, dominance of consumption orientation, failure of social contexts, destruction of and threats to environment, technical catastrophes etc., just in view of increasing options by technics and production, revealed the obvious failure and misconception of GDP to measure well-being. This becomes obvious by its positive accounts for destructions due to needs of repair work and for plundering of natural resources without account for negative consequences, by not considering natural and social sustainability (concerning the latter e.g. inequalities of income and wealth) and also by not considering positive contributions to well-being from unpaid home and care works.

It is now widely agreed that GDP is at best incomplete, and at worst misleading, as a metric of true prosperity for society at large and A/IS technologies (as noted in *The Oxford Handbook of Well-Being and Public Policy*).

Thus, searching alternatives, discussions about the basis of well-being came up, other measures came into view. Orientation towards other qualities was expressed e.g. by addressing sustainability of society and nature, harmonization by social balancing, post-growth or de-growth. The above guidelines of OECD - released 2013 – were considered to put something against the standard of GDP, to replace it and thus give a different perspective. It is further expected that respective activities

of national statistical agencies have "the potential to revolutionize our understanding of subjective well-being".

However, attempts to concretize became various and multifaceted and already by this questionable. In addition, they yield strongly different results in evaluating and comparing the status and development of nations. This does not justify GDP as measure and orientation of prosperity, good life and well-being and it also does not disqualify the search for and especially the discussion on alternatives. While the possibility of quantifying considerations on the quality of life is questioned, justification for the attempts lies in the in-depth considerations and weighing attempts about good life and well-being promoted hereby and contrasting the GDP measure. The various measuring attempts are to be considered especially as means to provide alternative thinking about societal developments.

An orientation beyond GDP becomes especially important with respect to the new technologies as A/IS. Due to their system character of combining everything and aiming at system decisions taking into account various interdependencies as well as due to their potential and already realized intervention depth into nature, humans and society (in combination with new genetic and bio-technologies), decisions about the use and design of these technologies with awareness of the influence on all aspects of life become increasingly important, decisions about where to go as humans and society.

These decisions cannot be made in an easy and general way since they imply chances and risks and need weighing, which becomes increasingly difficult due to high and increasing complexity. This again questions quantitative indexes as decision basis, also for the well-being benefit of A/IS, but just not considerations about various aspects and especially not reflecting holistic aspects of society versus the impact of any one technology. A/IS undoubtedly hold positive promise for society. But beyond the critical importance of designing and manufacturing these technologies in an ethically driven and responsible manner is the seminal question of determining options and criteria for their use, kind of use and benefit for human goals.

A/IS technologies can be narrowly conceived from an ethical standpoint; be legal, profitable, and safe in their usage; and yet not positively contribute to human well-

being. This means technologies created with the best intentions, but without considering well-being criteria, can still have dramatic negative consequences on people's mental health, emotions, sense of themselves, their autonomy, their ability to achieve their goals, and other dimensions of well-being.

Considering the use of quantitative indexes for individual and societal well-being in contrast to GDP, this should at least be done with caution, as they may provoke in users an automatic urge for numerical optimization. While this tendency is theoretically unavoidable, efforts should be invested in guaranteeing that it will not flatten the diversity of human experience and result in effects which are in contrast to those expected.

Rather than just considering indexes with respect to decisions about use and kind of use of technologies as A/IS, the underlying discussions about various aspects influencing and determining good life are important. Quantitatively, dynamic models about interdependencies may rather be helpful as support for decisions, as the original Forrester/Meadows model, now in much more developed state used in decision systems provided by A/IS. Thus, A/IS themselves can help to determine options of their use for better life.

However, this needs knowledge and awareness about the interactions considered in the models, about the applied approximations, in order to evaluate results (not essentially about the programming features, thus reduced need to require transparency in this respect). Further, it needs awareness and agreement about human goals in society as well as about possible contributions and risks of technologies as A/IS. Although broad discussions about well-being, well-being elements and indexes in contrast to present GDP orientation can support this, emphasis must finally be put on specific areas, as organization of health care, traffic, energy system, production goals etc.

Transformation of society towards a different orientation than GDP, even to no growth goals, needs finally participation of the whole society. This is the same with well-being and ethical criteria to be applied to A/IS. Decisions how to use A/IS are decisions about how to live and therefore even emphasize this need. Thus, this cannot be reduced to a technical task of implementing ethics and well-being criteria into given technical products and systems, into given A/IS as care robots (to

replace human care tasks and reduce care teams), self-driving cars (just to replace individual cars as overall traffic system), decision systems as Watson of IBM (without organizing human work processes to keep control) etc.

The task of reorganizing society and their goals in order to keep human control and support human development can also not just be delegated to technical experts. The task of experts is to make the consequences of decisions visible. This is also a task to be maintained during operation of A/IS and needs as well organization of participation of employees applying the systems.

Even technically it is not feasible to replace human control by aiming at perfect automated systems, only to be developed and controlled by few experts. Avoiding errors or misuse, guaranteeing safety, needs a culture of knowledge, reflection, discussion and awareness based on qualified working teams and involved society. – see also Theofanous paper, above, p. 11)

In conclusion, it is widely agreed …

Candidate Recommendation

A/IS should prioritize human well-being as an outcome in all system designs and applications. Well-being considerations in the frame of attempts to replace GDP as measure and orientation can be taken as reference. However, decisions are not to be restricted to final design of given products and systems. Considerations and discussions in working groups and society are to be promoted to decide about use and kind of use of AI/S in specific areas and contexts.

A general vision to be pursued is participation of all people at work and in society by enabling their access to major fields of design of society, productive, reproductive, cultural and political (see 4in1 perspective of Frigga Haug: http://www.friggahaug.inkrit.de/documents/4in1_englisch_fin.pdf). System aspects, becoming dominant with new technics, esp. A/IS, as well as increasing intervention potential of technics in all fields, with high chances and high risks, require this participation if decisions are not to be solely given to experts, thus undermining democracy. Also, technical control of complex systems needs collaborative control, not only control by a limited elite of experts.

Principle 4 – Transparency

Add at end of Candidate Recommendation (p.30):

Again, as described in the introductory remarks, the means to provide transparency will not be sufficient to avoid risks and even not to yield transparency. Only permanent control and involvement in design and decisions about use within a respective culture including institutions in feedback (checks and balances) can yield this.

Principle 5 – A/IS Technology Misuse …(p.31)

Background
end of first section, p.31:
… associated with the misuse of A/IS as well as a need to establish a respective culture supported by controlling institutions..
…
Candidate Recommendations
…
4. Establishing and supporting a culture of awareness, in addition based on controlling institutions and participation of initiatives of citizens, as indicated under 3.

Embedding Values into Autonomous Intelligent Systems (p.33)

At end of p.34 (above References):
A different position about the problems envisaged here and especially the way of solution picks up the difficulties of definition of general values outlined above and further of the way to implement them in A/IS and of making A/IS really work according to such values. The above conclusion to embed explicit norms as more realistic goal remains weak, due to the outlined need to specify for specific contexts, for a specific community and even specific purposes. This means to leave the problem open for such specific conditions. Further, expressing norms by obligations and prohibitions for the A/IS means to restrict capabilities of learning and deciding, including changing or updating of norms, which are just envisaged with the development of such systems. Therefore, this approach is prone to be undermined in real applications (see e.g. self-driving cars).

Thus, in an alternative position, the solution is not searched essentially by embedding values or norms in machines (a goal to be pursued additionally), but in creating societal solutions, in preparing the society to manage and control A/IS. This kind of approach is also required for adequately defining values and norms. It means to initiate clarification and decision processes about specific developments and applications of A/IS, discussions and struggles about the goals as well as the measures and to establish institutions and collective working teams permanently controlling and directing developments. This should be considered as a primary goal and includes to decide also about use and kind of use of A/IS in specific cases. In a democratic way, this should not be reduced to a task for experts.

An emphasis on the quality of working and managing structures to handle A/IS, more than on qualifying A/IS themselves, is also required in order to support proper functioning of the complex technics. The system character of technics itself requires from humans a thinking in systems, capabilities of judgement about proper functioning as well as overall goals. Building of empowered and well-trained teams and a culture of cooperation becomes a task of establishing standards. Finally, human responsibility can only be created and maintained in such processes, rather than by establishing an ethics or norms in advance in A/IS.

The need of permanent human intervention is partly expressed in following sections but not clearly outlined as a major task, e.g. p.41: "A/IS must have learning capacities to … incorporate user input". Rather, an emphasis lies on ways how an AI system could modify norms itself by learning "about its context and human norms" (p.43). Systematic risk analysis and mitigation means are considered on p.46/47, but not their dependence on a safety culture. P.50 at least states that the evaluation process must continue throughout the life cycle and claims some control or evaluation of "a system's norm-conforming behavior". Finally, questions are at least raised on p.53 (above candidate recommendations) as to whom the A/IS are accountable…

Well-being, p.240 (my proposal in red)

Prioritizing … While this is an encouraging trend, key questions remain.

As already outlined under General Principles and Principle 2 in General Principles, solving ethical concerns and gaining orientation towards benefits for humanity cannot start with given technical systems. In view of chances and risks, their use in specific fields and the kind of use has to be decided before. This requires decisions in processes of clarifying and discussing within involved working groups and the whole society. Decisions are to be found on products and systems wished, the extent and the way of using them, the specific support considered for human perspective, for well-being of human life and society, already before considering specific ethical implementations for specific technics. Questions arise how such decision processes including final decisions about design questions can be organized and come to conclusions. If an orientation at human benefit, well-being of individuals, society and nature is chosen, this means to decide along such criteria. Finally, people involved and the whole society have to decide.

Attempts have been undertaken to find objectified criteria for human benefit and well-being, others than common questionable criteria, and to use them for decisions and as measure of success. It has been shown that common metrics of success which include profit, gross domestic product (GDP),… planet and population.

The detailed approaches to introduce alternative measures have increased the understanding of goals and ways to be followed in the societal decisions on re-construction of working processes and of the whole society and thus also give in principle a sound basis for decisions on A/IS. However, broadening of respective views in society, and specifically within the technical community dealing with A/IS, is still strongly required. While the needs and possibilities of quantifying metrics to put other indices against GDP remains a subject of discussion, already due to the various approaches with differing results and further the basic question whether the qualitative goals can in principle be captured quantitatively, the detailed and profound elaborations provide openings and support a vision beyond present praxis and orientation with their obvious deficits, even destructive effects.

Those engaged with well-being metrics consider that for A/IS to demonstrably advance the well-being of humanity, there need to be concise and useful indicators to measure those advancements. They also realize that there is not yet a common understanding of what well-being indicators are, or which ones are available.

Concluding that therefore technologists will use best-practice metrics available even if, unbeknownst to them, said metrics are inappropriate or, worse, potentially harmful, it is considered to avoid resulting unintended negative consequences and to increase value for users and society by clear guidance on what well-being is and how it should be measured.

Therefore, the present document identifies examples of existing well-being metrics that capture such factors, allowing the benefits of A/IS to be more comprehensively evaluated.

[comment: I cannot really see this in the following text, which remains rather general, also the tables. Indicating a procedure is not sufficient. Construction of indices and outcome should be shown for **relevant examples** to enable assessment and discussions.]
While these indicators vary in their scope and use… measurement and efficiency of well-being indicators.

Finally, even with successful metrics, i.e. adequate well-being measuring and basic agreement in society, there remains the need to consider ways to realize alternatives to present praxis, to really push changes against resistances based on economic power and common arguments dealing with economic mechanisms, e.g. needs of economic growth, competitiveness, risk of loss of employment. Joining the present technical developments, this means the necessity to show that the change in orientation also yields better system solutions, e.g. concerning traffic and care systems, better social and economic solutions avoiding destructive perspectives. Concerning adequate changes of working processes, in accordance with system character of new technics, this addresses the general need of cooperative work to enable good handling of complexities and keep human control.
[As a basis for supporting this, theoretical and empirical studies about automation processes in the 70es/80es by Frigga Haug et al can be cited, which give still a valid basis. These basic analyses consider the inherent contradictions in the processes under capitalistic regime and thus yield orientation on how to develop alternative paths, also as solution paths, taking into account chances and risks. Considerations about other kinds of living are an essential part of these analyses]

Concerning design of society, the system character of technics and the transfer of large amounts of work to machines demands in democratic perspective a vision of general participation as indicated already under Principle 2 (4-in-1perspective).

Comments on following parts of Well-being (p. 241ff):

Concerning the following sections, I cannot make proposals since they are totally oriented at well-being metrics. As discussed and modified (as proposal) in the above pages, I cannot agree to this pointed position. In my view this really results in a positivistic perspective (in the 1960s there was a strong discussion of Adorno and Habermas against Popper known as "Positivismusstreit" – capitalist reality seems to have finally won with Popper), reducing everything to measuring and thus failing qualities although these are attempted to be introduced just by the metrics. This supports the reduction to quantifying as seemingly essential expression of science and inherent tendency in capitalism.to capture everything under market conditions as exchange of wares in terms of measuring values of goods (Tauschwert) in contrast to their utility for living. Thus, I am afraid that this emphasis in measuring well-being rather supports capitalist way than the intended development of alternatives.

In my above attempts, I tried to introduce balances and to keep the question open about contributions to develop alternatives in society from qualitative well-being considerations and quantitative approaches. This cannot be continued for the following sections. Perhaps, one could in the introductory sentences on p.241 at least modify saying something like:

belief that A/IS should prioritize human well-being as an outcome in all system designs, using the best available and widely accepted well-being concepts and metrics as their reference point.

Just the emphasis on "science" and its general (in my view wrong – and I say this as physicist!) identification with quantitative measuring empiricism acts against other thinking. And this can e.g. also be seen in the example with wheelchairs on p.247:

A crucial distinction between well-being metrics and potential interventions in their use is that a well-being metric does not dictate an intervention, but points the way for developing an intervention that will push a metric in a positive direction. For example, a team seeking to increase the well-being of people using wheelchairs found that when provided the opportunity to use a smart wheelchair, some users were delighted with the opportunity for more mobility, while others felt it would decrease their opportunities for social contact and lead to an overall decrease in their wellbeing. The point being that even increased well-being due to a smart wheelchair does not mean that this wheelchair should automatically be adopted. Well-being is only one value in the mix for adoption, where other values to consider would be human rights, respect, privacy, justice, freedom, culture, etc.

Here, due to the discovered ambiguity in use of smart wheelchairs (technical support versus reduction of human care and social contact), not to be resolved by metrics because both aspects have in principle equal validity (qualitative validity!), thus resolved by considering a difference between well-being metrics and intervention (i.e. consequences), finally leaving the decision open due to declaring well-being as not the only value to be considered.

Finally, it remains out of view that the ambiguity may disappear, if social contact and care is guaranteed anyway due to corresponding establishment of teams, working conditions with sufficient personnel, qualification and education, due to a culture of human relations. Measuring results depend on conditions which have to be determined by qualitative argumentation!

Setting “science” as main argument for the need and perspective to change is really problematic. We have to consider and elaborate our interests and visions of human living. Science may help us in this, but especially in analyzing realistic options, possible ways to go, obstacles etc. We have to analyze options inherent in the historical and now especially in the technical development as well as the contradictory elements. Only based on this we can come to a realistic perspective, to a concrete utopia (in the sense of Ernst Bloch).

P. 257/258:

Here, it is addressed that just most developed A/IS may increase manipulative potentials of the technics itself and just by this also of operators: "... significant risk

that unscrupulous operators will abuse the technology for unethical commercial, or outright criminal purposes. The widespread manipulation of humans by A/IS ... is by definition a reduction in human well-being" (p. 258). There is also specific address to deep learning methods (p. 257/258) which make the behavior of the systems unsure, difficult to be understood, intransparent and even manipulative. However, in the conclusions from this, I miss the requirements to organize working and community structures implying permanent control, i.e. establishment of a respective culture (see above). This should then be the primary goal, not implementing ethics and well-being orientation in the systems which by self-learning can anyway yield contrary features. The goal must be that humans and the society keep permanent control, can in permanent reflection decide on goals and ways to go. The candidate recommendations on p.258 are too weak for this.

Donovan Anderson BSc (Hons) MBCS AFA MIPA
Project Administrative Assistant - Responsible Ethical Learning with Robots
Centre for Computing and Social Responsibility

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems

Additional paragraph to add to document on *Ethically Aligned Design (version 2)* page **201**

This emerging and exciting new concept aims to also push the boundaries to incorporate relevant stakeholders whose influence in responsible research is on a global stage. While this concept initially focuses on the workplace setting, success will only be achieved through the active involvement from private companies of industry, those who are at the forefront in autonomous and intelligent system design. Responsible research and innovation (RRI) will be achieved through careful research and innovation governance that will ensure research purpose, process and outcomes are acceptable, sustainable and even desirable. It will be incumbent on RRI 'experts' to engage at a level where private companies will feel empowered and embrace this concept seeing it both as practical to implement and action.

Additional paragraph to add to document on *Ethically Aligned Design (version 2)* page **202** under:

Further Resources

Stahl, Bernd Carsten, 1968- ; Obach, Michael ; Yaghmaei, Emad ; Ikonen, Veikko ; Chatfield, Kate ; Brem, Alexander. **"The Responsible Research and Innovation (RRI) Maturity Model: Linking Theory and Practice. "** Sustainability, 06/2017, Volume 9, Issue 6

I have included via e-mail my comments on the *Ethically Aligned Design, v.2* document.

---

p.27 – Principle 3 - Accountability:

I would recommend that in the list of candidate recommendations a point is also included that designers and developers of A/IS should be responsible for implementing oversight mechanisms to monitor the performance of the A / IS to validate that at all times it operates in line with expected system performance and operational parameters.

---

p.68 – Section 4 – Lack of Transparency:

In the last sentence in Candidate Recommendations, you may want to also include 'comparability' as a criteria for the documentation. This would be useful for users to compare across different A / AS systems. This would also be useful for pre-trained models and/or open source systems that are developed by a third-party (see below for further discussion).

---

p.68 – Section 4 – Lack of Transparency:

An area that has not been explicitly addressed in this section is transparency into the underlying dataset(s) that are used to train the A /AS and/or later used in its operation. Many AI developers and operators utilize data sets that have been initially created and compiled by a third-party. As the data used to train and operate an A / AS can have a significant impact on its operation it's important that the creators of the dataset provide information on the dataset. This could include the creator's name, the context under which the data was collected, the nature of the data, the source of the data, the collection methods used, known missing or incorrect data, known biases, etc.

In the last several months, there have been two research teams that published papers / approaches on creating a profile sheet for datasets. On March 28, 2018 the Gebru et al. MIT-affiliated research team (arXiv:1803.09010v1) published a paper on *Datasheets for Datasets*, which recommends a standardized datasheet be mandated for inclusion with all datasets. Their paper includes a preliminary list of questions that could be included in the Datasheet. A second MIT research team in April 2018 published on line an abstract and prototype of their Data Nutritional Label which populates the label's qualitative and quantitative metrics using interrogative techniques against the dataset, similar to a food label.

The use of a datasheet or nutrition label could also be utilized for pre-trained models, APIs and other open source technology that is available for use. This would provide transparency to the secondary party on the appropriateness of the technology to their use, and any limitations that they should be aware.

---

p. 81 – General Principles:

The 2nd candidate recommendation which indicates that Technologists should minimize outcomes from the system that hinges on 'virtuousness of the operators' may be difficult to achieve due to the complexity of A / AS use cases and as in some cases the code is later transferred to and can be modified by the operator. It would be useful to include some additional practices that could be used including codifying restrictions in the terms of use / contractual language, maintaining ownership over the code, and monitoring operator use.

---

p. 108 – Personal Data and Individual Access Control:

The Candidate Recommendation on "privacy offsets" as a business alternative, could also mention the creation of public or private central data stores, or online digital identify safes which could hold an individual's data and negotiate data

exchanges based on a user's preferences. The central source could also be used to anonymize the data or utilize homomorphic encryption, where feasible.

Thank you,

Cathy R. Cobey
Partner, EY
cathy.r.cobey@ca.ey.com

Comments and feedback regarding Version 2 of Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (A/IS)

Prepared by:
Marek Havrda, Ph.D.
Olga Afanasjeva
GoodAI

Executive Summary - Overview

We propose to change the wording regarding Well-being:

"Well-being: Prioritise well-being through implementing well-being metrics in their design and use"

Comment: The prioritization of the metrics should not be the end-goal in itself, but should help in prioritising human well-being.

Methodologies to Guide Ethical Research and Design

We propose the following candidate Recommendation to be added (p 70-71):

Emphasize limits and edge cases: It is crucial that developers communicate transparently to customers and end users about the limits of the system (not only explain what it can do), and also give explicit examples where the system could fail. It might not be so straightforward to test the system on all possible scenarios before its deployment, but technologists should at least strive to extensively test it and describe possible extreme scenarios. Involvement of behavioral scientists and user experience designers in the process might help identify potential erroneous user scenarios (both unintentional, coming from natural assumptions about the product, and intentional) that could cause harm, and help inform users about them. An example of a situation to be avoided is Tesla's name choice for its driving assistance system, Autopilot: the name could have resulted in users relying too much on the system's capabilities with potentially harmful consequences.

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

We would suggest to include the issue of AI Race in Section 2 — General Principles

Issue:
Research and development of artificial intelligence is making encouraging progress. AI is being recognized as a strategic priority by a range of actors, including businesses representatives, private research groups, companies, and governments. This progress may lead to an AI race, where stakeholders compete to be the first to develop and deploy AI that would give them a strategic advantage. Such a transformative AI can be either AGI (a system that can perform a broad set of intellectual tasks while continually improving itself), or sufficiently advanced specialized AIs. It is important to address the potential pitfalls of competition towards transformative AI, where:

- Key stakeholders, including the developers, may ignore or underestimate safety procedures, or agreements, in favor of faster utilization
- The fruits of the technology won't be shared by the majority of people to benefit humanity, but only by a selected few

Candidate recommendation:
Initiate discussion among the teams developing AGI and other relevant actors which would address motivation mechanisms of various types to support cooperation and/or mutual oversight in order to reduce related risks, including existential risks to humanity, and to ensure AI is used to benefit the majority of humanity and not just a select few.

Personal Data and Individual Access Control

Indeed the A/IS presents a new opportunity to offer individuals/end users a "real choice" with respect to personal information. However, this may require new regulation and institutional settings. Therefore we suggest to add a candidate recommendation:

"Investigate the need for new regulation in order to move from "either/or" to "Yes and" relationship."

Economics and Humanitarian Issues

Issue: The complexities of employment are being neglected regarding A/IS.
We fully share the call to go beyond the sole numbers of jobs and suggest to explicitly add to the analysis the need to research distributional effects on various groups (by education, age, gender, sector, etc.) and regions (geographical distribution). Such an analysis should in turn enable better designed and target mitigation strategies.

Affective Computing

System Manipulation/Nudging/Deception

We fully agreed with the need to take into account ethical considerations when deliberate nudging is performed by A/IS. The relevant candidate recommendations should be supplemented by the need to monitor and study unintended consequences of nudging. For example, a system may nudge the user to the content which maximizes the engagement with given services. Yet this maximisation at the same time can lead to the erosion of cohesion across society, this seems to be the case of some existing social networks. These unintended impacts of nudging need to be monitored potential through the use of various well-being metrics. In general, new types of negative externalities such as erosion of society need to be studied.

Policy

Objective: Provide effective regulation of A/IS to ensure public safety and responsibility while fostering a robust AI industry.

Apart from UBI other possible means to ensure distribution of resources should be researched. It is important to investigate novel forms of public-private partnerships in this regards. It may become the joint responsibility of governments together with companies to derive new models to share AI co-generated wealth, either directly in the form of various grants or in the form of capital gains, e.g. creating new ownership structures favourable to individual investors such as cooperatives, or

through other special investment vehicles (which might generate income similar to UBI). Moreover, it is important to focus research on the tax base which may potentially erode due to developments of the labor markets. Considering tax reform there may be a need to find novel ways to automatically adjust the tax base in order to generate the needed revenue as traditional sources of taxes, relying on taxing labour, may be at least temporarily reduced. For example, a higher rate of VAT for less human labour intensive (i.e. more AI intensive) production would generate both more tax revenue, as well as encourage companies to keep humans in employment in order to create a time buffer for structural adjustment. Furthermore, a higher income tax on those profiting most from AI technology could help speed up the implementation of a new welfare system including UBI-type schemes.

Classical Ethics in A/IS

Section 2 — Classical Ethics From Globally Diverse Traditions
It seems that when it comes to ethics many traditions (not only Buddhism) see "relationships" as extremely important. The related candidate recommendation may therefore include an explicit call for research on how to identify main commonalities of "relationship" approaches from different cultures and how to operationalised them for A/IS to complement classical methodologies of deontological and teleological ethics.

Mixed Reality in Information and Communications Technology Committee

Issue:
A/IS, artificial consciousness, and augmented/mixed reality has the potential to create a parallel set of social norms.

Experimental research proves that even explicit opt-in (or opt-out in other situations) may not be a sufficient tool to help individuals make informed choices in a way that they would make choosers better off, as judged by themselves. Therefore, candidate recommendations should include research into various mechanisms of how to maintain autonomous decision making. This is also relevant to the section on System Manipulation/Nudging/Deception.

Well-being

1. General comments. Indeed, well-being metrics could be used for both defining objectives and help monitor for unexpected (negative) unintended consequences. However, we need to bear in mind certain limitations of the use of any metrics for such a purpose. In particular, Goodhard Law "When a measure becomes a target, it ceases to be a good measure." In other words, when a feature of the economy or society is selected as an indicator of a larger system, then it ceases to function as a good indicator because people (and A/IS) can start gaming the system by finding ways to achieve good scores on the indicator without achieving the aims which the measure was supposed to promote. The relevant candidate recommendation should stress the need to establish procedures against such a situation which may include continuous monitoring that each metric used remains to actually capture the qualities it is supposed to measure and promote.

2. Section 3 — Adaptation of Well-being Metrics for A/IS. The candidate recommendation could be complemented by stressing the need to address the trade-offs among various metrics. This should also include various "distributional effects", i.e. how is a given variable such as income or health distributed among various groups inside a given society.

3. Section 2 Well-being impact assessment (p 251). It is not clear how certain aspects based on Maslow's Hierarchy of Needs such as Self-Actualization can be applied to environment. More importantly, there needs to be more considerations on actual (not only a single expected) use of A/IS. This could be carried out in form of various scenarios of future usage of the A/IS which is being assessed. Any impact assessment should include the impacts along the whole (expected) life cycle, an analogy with LCA (life cycle assessment) used in environmental assessments could help to illustrate such an approach.

4. Human Rights and Well-being (p 259 - 206). The candidate recommendation could also reflect the need to develop and operationalize metrics which would capture the quality of Human Rights.

General comments

It seems that better internal integration of recommendations would significantly improve the value of the document, in particular integrating Well-being metrics and Classical Ethics with Embedding Values into A/IS would be beneficial. In addition, it may be useful to clearly indicate who is the primary addressee of each recommendation (e.g. government).

Submission: Babita Ramlal, Ontario Ministry of the Attorney General, Innovation Office

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. *Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems,* Version 2, IEEE, 2017.

1. *On page 8, 193-195, 203, I recommend that the team reviews other ethical traditions, other than classical western ethics (Hobbes, Locke, Rousseau, Kant, Bentham, Mills, Aristotle etc.) Indigenous Peoples, India (Hinduism, Buddhism, Jainism, Sikhism), and various traditions from China (Confucianism, Mohism, Daoism) have their own history of ethics and virtues. Some expertise and discussion from this area can inform a more universal set ethical principles.*

   Regarding this subject, I recommend reviewing this online application for Making an Ethical Decision: A practical tool for thinking through tough choices[41], from Markkula Center for Applied Ethics and the value systems of Hinduism, Confucianism, Buddhism, and Indigenous peoples.

2. *On page 23, I recommend that the United Nations Declaration on the Rights of Indigenous Peoples[42] be included as a resource. Technology has always been used to subjugate and exterminate Indigenous Peoples.*

3. *On page 27, I recommend the inclusion of the development of Life Cycle Continuum for AI/S use through Maturity Models for Organizations with AIS implementations. Accountability is already built into most governance models for projects and organizations, which also have core values and codes of ethics. However, it is the lack of enforcement and not having consequences for actions that will prove to be an issue. Accountability will have to depend on a combination of policy instruments (from international law, domestics laws, regulations, standards) to professional ethics,*

[41] https://www.scu.edu/ethics-app/

[42] https://www.un.org/development/desa/indigenouspeoples/declaration-on-the-rights-of-indigenous-peoples.html

organizational ethics boards, internal and external audits (with certifications) etc.

4. *On page 29, a comment on Principle 4, Transparency. An improvement in transparency usually includes an offset of some degree of Privacy. The concept of providing evidence in court is based on the memory and recall of observations by human witness who provide expert witness testimony. It has been proven that this is a flawed system, because human recollection is not reliable, and also biased. This is one area in which AI/S may have the advantage as it can be a record of it decisions and actions for review.*

5. *On page 36, section 1 – Identifying Norms for AIS. Universal rights and the constraint on norms are mentioned but what about adaptability in cultural programming for AIS that cross borders. Delineating community in which the AIS operates is not enough. Norms and values are fluid even within a country.*

6. *On Page 50, Section 3 – Evaluating the Implementation of AI/S, it is possible, guidelines for Quality Assurance and User Testing of AI/S will need to be elaborated in more detail and will include using universal methodologies such as independent observation of user testing, audits by groups such as ISACA and State Auditors. In a competitive market, producers of AI/S may be tempted to cut corners on QA and User Testing to get product to market and criteria for product recall when there is an error.*

7. *On page 58, we can elaborate to add that the lack of sustained interdisciplinary collaboration is being offset by the increase in informal information sharing groups through online media sites and subject matter groups and government supported innovation programs, hackathons, incubators and competitions to that sponsor and support technological innovations.*

8. *On page 68, the Candidate Recommendation should include the development of enterprise architecture standards (conceptual, logical and physical groups of artefacts) for AI/S.*

9. *On page 71, the first Candidate Recommendation placing the onus on engineers for safety and well-being needs to be elaborated upon. In organizations technical staff are partially responsible for design choices and the execution of that design. The program/product owner or project sponsor from the governance team holds the accountability. Therefore, a RASCI should outline roles, responsibilities and accountabilities, including the impact of the AI/S created on human/animal well-being.*

10. *On page 78, we should include something on the updating of safety standards and laws such as the ones under OSHA, to include AI/S innovations. A safety mindset is good but safety should be part of the promise of AI/S not a possible by-product, just as privacy should be built in using Privacy by Design principles.*

11. *On page 81, in the Candidate Recommendation, we should include 3rd party review by an independent party, for AI/S systems that meet certain criteria (e.g. a threat to jobs of many industries, potential for misuse by organized crime etc.)*

12. *On page 84, digital personas are being used in a benign way by government to create user-centric service design of public services, however, something more must be said about the covert use by private entities such as Uber, who track phones of their passengers and sell the data to an analytics firm that uses to create profiles of customers for a bank.*

13. *On page 98, I recommend that a committee should work on a methodology or framework to develop PIAs for AI/S as the existing frameworks may not be suitable. Also, conformity assessment should be by an independent 3rd party similar to the process used by ISO and UL.*

14. *On page 164, 166-167, Systems across Cultures. Engineers should consider the need for cross-border movement of affective systems. Therefore, the need for flexibility in cultural programming and mechanisms to enable and disable such 'add-ons' should be considered.*

15. *On page 182, since the second draft of the standard, many countries have created guidelines or principles related to AI/S approaches. The Objectives listed on this page should be reviewed in light of these developments.*

16. *Regarding Candidate Recommendation on page 196, it is inevitable that advanced affective AI/S will be subject to anthropomorphism (as are animal companions today). This is inevitable, and there is little than we can put in place, apart from education and guidelines.*

17. *On page 180, this section should be elaborated upon. The candidate recommendations needs to address the concept of synthetic emotions further and thresholds for personhood, as more and more organic beings are being granted personhood. A continuum of criteria and behaviours need to be developed for what constitutes personhood in AI/S.*

18. *On page 198, Classical and Ethical Education should become a part of mandatory education for all who work on, govern or regulate AI/S. However, Ethics in high school curricula can lay the foundation for a different generation of users of AI/S who can better navigate their use and relationships with evolving systems.*

I wish to thank the writers for allowing responses to an already thoughtful, thorough and timely report. The following remark are by no means critical but only an effort to emphasize certain aspects. My remark pertain mostly to the chapter that represent the work of the Committee for Classical Ethics in Autonomous and Intelligent Systems (pp: 196) and to the general ethical approach in other parts of the report.

The writers courageously embark on predictive/ speculative ethical analysis of artifacts that are still technically immature and will be implemented in the future. The exact morals and behavioral contracts of that future are (to a certain degree) still unknown and will probably be shaped by those same artifacts. To evaluate possible risks to the rights and wellbeing of future stakeholders the authors follow the most basic and classical philosophy. However there is possibly a need to doubt those principles and search for others that could be more supportive to a new field of technological artifacts especially where risks and safety issues are less imminent. For example, many aspects of human nature that were not necessarily jeopardized by former technologies, could be curbed by A/ISs. Some of those traits are related to the wondering and wandering aspects of human nature. Adjectives like "serendipity, eclectic, intuition, accidental, blunder, and "getting lost"" come to mind as they represent aspects of human mode of thinking and behaviors and many times are the initiating steps of creativity or at least un-predicted experiences. These sides of human natural behaviors were not as evident in classical ethical discussions (such as in bio-medical ethics). However they could possibly become endangered when certain applications of some A/IS technologies emerge.

I believe that innovative revolutionary technologies in general, and A/IS in particular, deserve innovative "ground breaking" moral evaluation. This does not nullify traditional methodologies (scenarios!) or all the current ethical principles. However, certain issues as described above are beyond their scope.

Sincerely,
Dr. Ilana kepten
ORT Braude Engineering College
Carmiel, Israel

Pradyot Sahu
Director, 3innovate, India
https://www.linkedin.com/in/pradyot-sahu/

1- RISE OF A/IS PLATFORMS

Recently, there are new A/IS Platforms. The platform providers are the front-runners of A/IS such as Google, Microsoft and Amazon enabling A/IS implementations in each mobile, web or stand-alone app and applications. As practically every A/IS system that use an A/IS Platform to implement will be difficult to evaluate and monitor just like it is difficult to monitor each android mobile app.

A) It will be easier to evaluate and monitor a platform and make the platform builders liable to the problems they create intentionally and unintentionally.

B) Platform builders must use platform controls to enable control of A/IS applications against any potential misuse of A/IS technology

C) A/IS Platforms related principles must be discussed, created and included in the General Principles of IEEE EAD document.

2- EVALUATION AND COMPLIANCE

Without evaluation and compliance, there is no way to make A/IS system builders and platform providers accountable.

A) There may be national level organizations such as **United States Department of Artificial Intelligence**, **USDAI** like United States Department of Health.

B) and international organizations in the form of United Nations Regulatory Agency such as **International Artificial Intelligence Organization, IAIO** to regulate A/IS.

C) EU General Data Protection Regulation, GDPR and Health Insurance Portability and Accountability Act of 1996, HIPAA style evaluation and compliance are a must for the A/IS system and platform builders.

Best regards

Pradyot Sahu

Director, 3innovate, India

https://www.linkedin.com/in/pradyot-sahu/

IEEE – Ethically Aligned Design, v.2
Written Comments of the Electronic Frontier Foundation

Jamie Williams, J.D.; Jeremy Gillula, Ph.D.; Lena Gunn
Electronic Frontier Foundation
815 Eddy Street
San Francisco, CA 94109
United States of America
Telephone: +1 (415) 436-9333
Email: jamie@eff.org

1. The Electronic Frontier Foundation (EFF) submits the following comments in response to the second version of IEEE's Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (EAD Version 2). EFF is a member-supported, nonprofit, public interest organization composed of activists, lawyers, and technologists, all dedicated to protecting privacy, civil liberties, transparency, and innovation in the digital age. Founded in 1990, EFF represents tens of thousands of dues-paying members, including consumers, hobbyists, computer programmers, entrepreneurs, students, teachers, and researchers. EFF and its members are united in their commitment to ensuring that new technologies are not used to undermine civil liberties. EFF is particularly concerned with the implications of autonomous and intelligence systems (A/IS) on privacy, government surveillance, and transparency, and offers these comments to enhance protection for individual rights.

2. For the purposes of our comments, we use a fairly broad definition of intelligent systems, which includes everything from simple machine-learning systems to advanced deep-learning techniques, and recommend that IEEE do the same for purposes of EAD Version 2. While others may focus their remarks on more advanced systems, we believe it is important to acknowledge that even simple AI systems in use today (which some may no longer even classify as AI) are already having a dramatic impact on society. In addition, given the human tendency toward "automation bias"—to view machine-generated outcomes as inherently trustworthy—even simple machine-learning systems intended to *assist* human decision making may in effect *replace* human decision making. And because less-advanced systems will serve as the

foundation for continued progress, EAD Version 2's recommendations will have the most impact on the design and development of advanced systems of the future if organizations begin adopting those ethical design principles today.

3. **Human Rights**. We agree that the ultimate goal of A/IS should be to promote human wellbeing and that A/IS should be designed and operated in a way that respects fundamental human rights. Unfortunately, as examples from around the globe show, that is not always the reality on the ground. To ensure that A/IS best honor human rights, and to prevent systems from amplifying existing inequalities and negatively impacting vulnerable communities, EFF recommends the following:

   a. On p. 23, adding the following recommendation: In the very first stages of developing A/IS, organizations and data scientists should engage the various stakeholders, including both civil society and the community that will be impacted, to ensure that the broader impact of the A/IS is considered, to assess whether the A/IS will reduce or reinforce existing inequities, and to help avoid past mistakes or oversights and create more societally beneficial systems. This requires taking a close look at who the A/IS could impact and what that impact would look like, and identifying the inequalities that already exist in the current system.

   b. On p. 23, adding the following resources: (i) AI Now Institute's report on algorithmic impact assessments; (ii) an April 2018 interview with Cathy O'Neil, discussing using an "ethical matrix" to help lay out all competing implications, motivations and considerations of a system; (iii) a March 2018 *Economist* article recommending that policymakers "apply the lessons of the horseless carriage to the driverless car"; (iv) a May 2018 EFF blog post by Jamie Williams and Lena Gunn, "Math Can't Solve Everything: Questions We Need To Be Asking Before Deciding an Algorithm is the Answer," and, as examples of algorithms that raise ethical questions, (v) a March 2018 *Wired* article about the discriminatory impact of an algorithm used to make custody decisions in the UK, and (vi) a September 2017 *Guardian* article about an algorithm that guessed whether a person was gay based on a photograph.

4. **Surveillance.** A/IS pose a serious privacy risk. Already, the deployment of machine learning techniques has enabled efficient large-scale surveillance both

by intelligence agencies and commercial actors. In the past, large-scale surveillance of a population was limited by the human resources available to sift through the data collected. Only societies like East Germany, that were willing to recruit one informant per 6.5 citizens, could possibly watch and pay attention to all of their citizens' actions. But the combination of already-deployed surveillance technologies and machine learning for analyzing the data will mean that exhaustive surveillance is becoming possible without the need for such enormous commitments of money and labor. The potential of machine learning to enable such effective large-scale surveillance has reduced the price tag of authoritarianism, and poses a novel threat to free and open societies. For this reason, EFF believes that machine learning algorithms operated or controlled by or on behalf of any governmental entity should get access to a person's data only with their consent and control, or a properly issued warrant. Given the history of governments—particularly that of the United States—to create secret surveillance law via confidential legal memos, court opinions, and agreements with foreign nations, in addition to other methods that leave no avenue for public scrutiny, strict compliance with a warrant requirement is necessary to combat illegal surveillance efforts. With these concerns in mind, EFF recommends:

a. On p. 99, updating the "background" section to include: Another source of the problem is pervasive government secrecy surrounding mass surveillance, including not only the scope of government surveillance but also the legal justifications, leading to the rapid erosion of any meaningful checks on governmental power.

b. On p. 100, updating the recommendation that each data acquisition come on a case-by-case basis to include a warrant requirement. Delete the phrase "unless the ongoing access has become law"; given governmental abuse of secret law to justify rampant dragnet surveillance that violates the human rights of individuals across the globe, this could be used to justify secret dragnet surveillance and/or to circumvent legal privacy protections.

c. On p. 111, adding the following resource: Elizabeth Goitein's report for the Brennan Center for Justice, "The New Era of Secret Law," which

documents the pervasive secrecy surrounding surveillance law in the United States.

5. **Transparency.** We are happy to see transparency included as a core principle. As algorithms and A/IS are increasingly relied upon to assist or replace human decision making about access to resources or other issues with the potential to negatively impact people's lives, there is a great risk that the public will lose access to—and thus the ability to influence or challenge—the policies and practices governing their lives. Transparency, explainability, and documentation is critical whenever human decision making with the potential to significantly impact human life is delegated to a machine—be it within the context of a government agency, judicial system, or private organization. And true transparency necessitates not only an explanation of a given model, but also explanation and documentation of the processes behind the model's development and use, and making those systems available to researchers for detailed analysis.

   We believe EAD Version 2's recommendations can be enhanced to foster robust and meaningful transparency. EFF recommends:

   a. On pp. 29-30, (i) expanding the concept of transparency to include not only transparency of a model's operation (e.g., why the model made the decision it did), but also transparency regarding the processes behind the model's development and use (e.g., what happened in the design process); and (ii) update the second category of stakeholders to explicitly include independent researchers, auditors, and journalists, and add a recommendation that A/IS—particularly those with the potential for amplifying existing societal inequalities and/or negatively impacting vulnerable community—be subject not only to validation and certification, but investigation by independent research scientists, auditors, and journalists.

   b. On p. 30, (i) adding the following recommendation: Impact assessments should be mandated—by either industry standards or law—to document decision-making processes in a system's development/design and use. Documentation could show, for example, that an engineering team tested a model with and without the social media data and found that using the data reduced the disproportionate impact of the model, or that a

team considered adding additional features necessary for a more accurate and fairer model but, after discovering that such features were exceedingly difficult or costly to measure, the company decided to instead use social media data, which increased accuracy and fairness under the practical constraints faced by the company; and (b) adding the following resources (from which the recommendation and examples in this paragraph are drawn): Andrew Selbst's and Solon Barocas's 2018 paper, "The Intuitive Appeal of Explainable Machines."

c. On p. 30, also adding the following resource: Sandra Wachter, Brent Mittelstadt, and Chris Russell's 2018 paper on counterfaction explanation research.

d. On p. 68, (i) recommending that documentation be required not only about a system's performance, limitation, risks, and data flows, but also about a system's development and use (*e.g.*, choices or tradeoffs made concerning a model); and (ii) adding the Selbst/Barocas paper listed above to the list of further resources.

e. On pp. 159-160, recommending mandated documentation of a system's development and use, including the fairness measure employed, and preparation of impact assessment. Governments and industry groups should consider establishing standards that require such documentation and impact assessments, in addition to logs and auditing trails.

f. On p. 160, recommending that A/IS—particularly those with the potential for amplifying existing societal inequalities and/or negatively impacting vulnerable community—be subject not only to validation and certification, but investigation by independent research scientists or auditors.

g. On p. 160, remove recommendation 10, that "[T]he general public should be informed when articles/press releases related to political figures or issues are posted by an A/IS, such as a bot." This sort of mandate, while appealing, could dramatically reduce the ability of individuals to speak anonymously and therefore raises significant free speech concerns.

We also believe that EAD's recommendations concerning transparency and individuals rights implicated by governmental use of A/IS could be strengthened. EFF recommends:

h. On p. 152, expanding the second recommendation. Governments should also not employ A/IS if:
(i) they do not have documentation of the decision-making processes behind the system's development and use (including documentation from any third parties vendors, developers, data scientists, etc); (ii) they have not conducted an impact assessment that considers concerns of the community impacted and all relevant stakeholders; and (iii) they have failed to provide a meaningful opportunity for public review and comment. Governments should only employ an A/IS if they have worked with third party developers/engineers at all stages of the system's development to ensure that the concerns of impacted communities are considered, and to ensure that government employees interacting with the systems are properly trained on the system's limitations. Governments should also not employ A/IS if, after conducting an impact assessment, the risks of causing harm or amplifying social inequalities cannot be mitigated by safeguards. Impact assessments should be conducting in an open and transparent manner, and impact reports/statements should be made public. Records relating to impact assessments and to an A/IS's development and use should be subject to release pursuant to any applicable freedom of information or open records laws.

i. On p. 153, expanding the fourth recommendation to include not only an opportunity for individuals negatively impacted by an A/IS to make a case for "extenuating circumstances," but also to seek human review/reconsideration of any A/IS's decision—without the need for any additional facts, data, or extenuating circumstances. The decision of an A/IS should not be presumed accurate, fair, or just absent extenuating circumstances. Review should be reasonably expeditious, considering the circumstances, and individuals should have a right to inspect the facts and law supporting the decisions, the system's logic and algorithm (wherever possible), audit trails, and any data about the individual utilized by the system in reaching its result. Individuals should have fair notice of their

rights to seek review and a meaningful opportunity to challenge the A/IS's decision before the decision is acted upon.

j. On p. 153, adding a recommendation that any proposal to employ A/IS by local governments or municipalities be approved by the city council or other applicable elected governing body, following a review process that includes: public notice and comment on a use policy and impact report; a public vote, after a public hearing, regarding whether the benefits of the proposal outweigh the costs and whether there are sufficient safeguards to mitigate the potential harms; and annual impact reports to the elected governing body.

k. On p. 153, adding a recommendation that any companies developing A/IS for governmental or law enforcement use develop contractual terms that prohibit the use of their products in unethical or unlawful ways and that allow companies to withdraw their systems upon learning of unethical or unlawful uses. To limit unethical or unlawful uses of their systems, companies should also refuse to sell a particular A/IS or feature to a governmental agency unless the agency adopts transparent and enforceable safeguards that are supported by impacted communities, and unless the technology has been approved by the applicable elected governing body. Companies developing A/IS for governmental or law enforcement use should build public transparency and accountability directly in its design decisions.

6. **Safety-Critical Systems and Data Sharing**. As A/IS with the potential to put humans at risk become more pervasive, it is critical for companies to be transparent when things go wrong—and not just with accident investigators. In testing and deploying these systems in the wild, companies are deciding what risk it will impose on the rest of society. The public has the right to understand that risk, what companies are doing to mitigate it, and whether they've been subject to any unnecessary risk. They also have the right to demand that companies take reasonable steps to help make the technology safer for everyone—including by sharing incident data. Acting in isolation, companies have few if any incentives to share data. But if sharing is the rule, technologies will be collectively safer, and the public will be much better off. EFF suggests the following:

a. On p. 30, adding a recommendation that, for safety-critical systems, companies should develop incident-response protocols that include sharing data about accidents or other safety incidents. That data needs to be shared between companies—so that they can analyze what went wrong, learn from each other's mistakes, and all get safer faster—in addition to government regulators, academic research labs, and ideally the public. The exact scope of data that should be shared, and who it should be shared with, should be determined on a technology-by-technology basis, carefully balancing the specific privacy implications at play.

b. On p. 30, adding the following resource: (i) a March 2018 EFF blog post by Jamie Williams and Peter Eckersley, "Some Easy Things We Could Do to Make All Autonomous Cars Safer"; and (ii) an April 2018 Los Angeles Times article, "As the Number of Driverless Cars Increase, So Does the Need for Car Maker Transparency."

7. **Participation and Fairness Norms.** Just as norms and values vary across communities, there are many measures of fairness. For instance, does the system treat like groups similarly, or disparately? Is the system optimizing for fairness, for public safety, for equal treatment, or for the most efficient allocation of resources? Is there an opportunity for those adversely impacted to seek meaningful and expeditious review, before any undue harm is felt? Organizations should be transparent about which fairness measure its system is using. The community that will be impacted by an A/IS should have an opportunity to participate in and influence decisions about which fairness measures the system will use. EFF recommends the following:

a. On p. 37, expanding the existing recommendation to explicitly refer to fairness. The impacted community's fairness norms should be considered in assessing the fairness measure(s) the A/IS will use, and the community should have an opportunity to participate in and influence decisions about the fairness measure(s) employed. This is particularly important in cases involving A/IS with the potential to negatively impact people lives or exacerbate existing inequalities.

b. On p. 37, adding the following resource: Shiraa Mitchell 2018 paper on qualitative fairness, "Mirror Mirror."

8. **Autonomous Weapons Systems**. As EAD Version 2 recognizes, using AI systems in military situations is incredibly risky, where even seemingly small problems can result in fatalities, escalation of conflicts, or wider instability. Autonomous Weapons Systems (AWS) can often be difficult to control and may fail in surprising ways. In military situations, failure of AI could be grave, subtle, and hard to address, and the boundaries of what is and is not dangerous can be difficult to identify. Society has not yet agreed upon necessary rules and standards for transparency, risk, and accountability for non-military uses of AI, much less for military uses. Given the hefty ethical stakes, companies working with military agencies must be extremely cautious—particularly where the application involves potential harm to humans or could contribute to arms races or geopolitical instability. To bolster EAD Version, EFF recommends the following:

   a. On p. 121, adding the following recommendation: Contractors working with military agencies should consider the consequences of the AWS and demand accountability and standards of behavior from the military agencies that seek their expertise; they should not assume that the contracting military agency has fully assessed the risks, or that they don't have a responsibility to do so independently.

   b. On p. 121, adding the following recommendation: At a minimum, any company, or any worker, considering whether to work with the military on a project with potentially dangerous or risky applications should ask:

      - Is it possible to create strong and binding international institutions or agreements that define acceptable military uses and limitations in the use of A/IS/AWS?
      - Is there a robust process for studying and mitigating the safety and geopolitical stability problems that could result from the deployment of military A/IS/AWS? Does this process apply before work commences, along the development pathway and after deployment? Could it incorporate sufficient expertise to address subtle and complex technical problems? Would those leading the process have sufficient independence and authority to ensure that it can check companies' and military agencies' decisions?

- Are the contracting agencies willing to commit to not using the AWS offensively, or to ensuring that any defensive AWS are carefully engineered to avoid risks of accidental harm or conflict escalation? Are present testing and formal verification methods adequate for that task?
- Can there be transparent, accountable oversight from an independently constituted ethics board or similar entity with both the power to veto aspects of the program and the power to bring public transparency to issues where necessary or appropriate?

c. On p. 121, adding the following resource: an April 2018 EFF blog post by Peter Eckersley and Cindy Cohn, "Google Should Not Help the U.S. Military Build Unaccountable AI Systems."

d. On p. 128, adding the following recommendations: (i) police and security systems should not be permitted to deploy AWS that enables unlawful surveillance, and must be transparent about their use of any AWS; and (ii) any use of an AWS by a local government or municipality must be approved by the city council or other applicable elected governing body, following a review process that includes: public notice and comment on a use policy and impact report; a public vote, after a public hearing, regarding whether the benefits of the proposal outweigh the costs and whether there are sufficient safeguards to mitigate the potential harms; and annual impact reports to the elected governing body.

e. On p. 128, adding the following recommendation: Companies developing AWS that will be used by domestic police forces should develop contractual terms that prohibit the use of their products in unethical or unlawful ways and that allow companies to withdraw their systems upon learning of unethical or unlawful uses. To limit unethical or unlawful uses of their systems, companies should refuse to work on an AWS project with domestic law enforcement agencies unless they adopt transparent and enforceable safeguards that are supported by impacted communities, and unless the AWS has been approved by the applicable elected governing body.

9. **Copyright, Fair Use, and User Control.** Today, the pervasive use of dangerous digital rights management (DRM) technologies threatens users'

security and privacy, distorts markets, confiscates public rights, undermines innovation and fair use, and impedes basic user rights and expectations. Companies routinely misuse laws like the Digital Millennium Copyright Act (DMCA)—which was originally meant to deter illegal copying of software and generally prohibits unlocking "access controls" like DRM—to chill competition, free speech, and fair use. We must take care not to build this failed approach into A/IS. EFF recommends:

a. On p. 236, removing the recommendation for research into "new forms of creative copyright to be embedded within physical and virtual environments." While such a recommendation sounds appealing, it could impede innovation, security, and basic user rights and expectations.

Feedback for *Ethically Aligned Design, Version 2* of The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems

Author: Emerson Rocha <rocha@ieee.org>
Organization: Etica.AI

This feedback proposes a creation of a new special working group concerned with internationalization of the future final version of Ethically Aligned Design and it's intrinsic related extra documents.

The internationalization problem

> "*To speak a language is to take on a world, **a culture***" — Frantz Fanon
>
> "*We, the English educated Indians, often unconsciously make the terrible mistake of thinking that the microscopic **minority of the English-speaking** Indians **is the whole** of India.*" — Mahatma Gandhi
>
> "*Only those who do not consider political questions deeply can **ignore language questions** in South Africa*" — Neville Alexander
>
> *"No pedagogy which is truly liberating can remain distant from the oppressed by treating them as unfortunates and by **presenting for their emulation models** from among the oppressors. The oppressed must be their own example in the struggle for their redemption*" — Paulo Freire, 1970

We use English as working language to draft this international document and not another language because of key events happened in the world history. A lot of these events on last centuries were not based on, what we call on the EAD document, *informed consent*, just to avoid be more explicit. But ok. Now, the question that we all — and this include even me and other non-native English speakers — is: **are we using English just as working language, or a language**?

To make it very clear in advance: this feedback do not ask to change the working language from English to any other from the The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, or any other group who work with promotion whith ethics on A/IS because English has advantages as a working language at this

point in history. This also is not even about translations, because **the core issue here is about internationalization**[43].

My endorsement of Chinese committee comments

I endorse the public comments made by Jia He, IEEE Global Initiative China Committee member on the EADv1 http://standards.ieee.org/develop/indconn/ec/rfi_responses_document.pdf related to the need of promoting awareness of A/IS on developing countries and the need of make those people work together. But I would go further: without optimization of the way the documents are translated, the full translation of not just the summary, but the full 260+ pages document, the document will not get translated to languages that it would really need more.

When it is understood just as "a translation issue", without think about the big picture, the obvious universal response is try to use money resources as a brute force mechanism to mitigate less reliable translations: we pay specialists. This strategy is non-scalable[44] and optimized for an economic and political system of the last centuries[45].

The real problem is on countries that do not have even members on working groups like the IEEE Global Initiative. And the way we think about improve internationalization of A/IS documents and other very important required international documents should be able to work on scenarios where no central organization already exist.

The problem on translations beyond the 6 officials for Human Rights documents

To understand how deep is the problem with translations related to ethics on A/IS we must consider that very important documents, defined internationally in United

[43] "Internationalization is the process of designing a software application so that it can be adapted to various languages and regions without engineering changes." https://en.wikipedia.org/wiki/Internationalization_and_localization

[44] An ideal scenario means a scalable system, in the sense of able to accommodate large amount of users (translators) working together (e.g. a language or a dialect of a language)., See also: https://en.wikipedia.org/wiki/Scalability

[45] This disproportionately impact countries with different spoken languages. One example is the case of language on Africa https://en.wikipedia.org/wiki/Education_in_Africa#Language

Nations by over a hundred nations, can have no translation at all even on the official language of these nations. And this issue must be addressed, **as ethics on AI/S is required to honor human ethics**.

Beyond the "Universal Declaration of Human Rights" document (with over 500 translations and a guiness world record as the most translated document) the only translations we can take for granted, even for the very important documents, are the UN 6 official working languages: Arabic, Chinese, English, French, Russian and Spanish. So, what happens if a language is not one of those?

The experience of try to find a reliable translation seventh most spoken in the world (Portuguese) is already unpleasant, at best. Very few documents have translations on the same places of the 6 UN languages, and the others depends of external sites.

Some human rights documents could have multiple translations (like one from Organization of American States have differences from the document at Brazilian Chamber of Deputies) and I believe that is even possible some laws around the world for those countries outside the 6 UN languages may have been influenced by opinions based on inaccurate translations.

If the seventh most spoken language in the world already have these type of problems with documents from five decades ago, what happens to the other hundreds of languages?

Creation of a working group specific about internationalization

It is really important to have an exclusive working group to deal only with internationalization (not about translation, but make easier to future translations) of the document from the final version of the Ethically Aligned Design. At the current point, we are using English as "working language" with the same meaning of the United Nations 6 official languages.

We have problems, from the file used to distribute the document (a pdf) to implicit usage of reference documents on the body of the text that assumes the viewer read a document that will never be translated (e.g. the documents that requires payment).

Angeles Manjarrés, Departamento de Inteligencia Artificial. Universidad Nacional de Educación a Distancia (UNED) de España

- Simon Pickin, Departamento de Sistemas Informáticos y Computación. Universidad Complutense de Madrid, España.
- Miguel A. Artaso, Departamento de Inteligencia Artificial. Universidad Nacional de Educación a Distancia (UNED) de España

We think it is important that the following points be handled (apologies for the haste and the length of the text):

Economic and Humanitarian Issues

According to the UN "Sustainable development" has been defined as "development that meets the needs of the present without compromising the ability of future generations to meet their own needs:

- Sustainable development calls for concerted efforts towards building an inclusive, sustainable and resilient future for people and planet.
- For sustainable development to be achieved, it is crucial to harmonize three core elements: economic growth, social inclusion and environmental protection. These elements are interconnected and all are crucial for the well-being of individuals and societies.
- Eradicating poverty in all its forms and dimensions is an indispensable requirement for sustainable development. To this end, there must be promotion of sustainable, inclusive and equitable economic growth, creating greater opportunities for all, reducing inequalities, raising basic standards of living, fostering equitable social development and inclusion, and promoting integrated and sustainable management of natural resources and ecosystems."

According to the report World Economic Situation and Prospects 2018, weak growth in per capita income poses challenges to sustainable development targets in several regions. Further setbacks or negligible growth in per capita gross domestic product (GDP) are anticipated in Central, Southern and West Africa, Western Asia, and Latin America and the Caribbean. These regions combined are home to 275 million people living in extreme poverty. But this data augurs a future of very adverse

social and environmental conditions not just for the inhabitants of the LMIC but for all humanity.

"The road to dignity" document of the UN 2030 Agenda for Sustainable Development project puts great hope in science, technology and innovation. One of the targets of the 2030 Agenda is "The spread of information and communications technology and global interconnectedness" arguing that it "has great potential to accelerate human progress, to bridge the digital divide and to develop knowledge societies, as does scientific and technological innovation across areas as diverse as medicine and energy."

An increasing number of ethical investment schemes oriented towards the common good are being offered by the financial sector. In collaboration with the IBRD (part of the World Bank), Banque SYZ is offering the first bonds in Switzerland that directly link private investors to the Sustainable Development Goals (SDGs). As stated by the IBRD: "Returns are linked to the performance of companies advancing global development priorities set out in the SDGs, including gender equality, health, and sustainable infrastructure. The World Bank will use the proceeds to support the financing of projects that advance its goals of eliminating extreme poverty and boosting shared prosperity, and that are aligned with the SDGs. The return on investment in the bonds is directly linked to the stock performance of companies included in the Solactive Sustainable Development Goals World MV Index. The index includes 30 companies that, based on the methodology developed by Vigeo Eiris' Equitics, dedicate at least one fifth of their activities to sustainable products, or are recognized leaders in their industries on socially and environmentally sustainable issues."

A/IS can play an important role in the solution of the deep social problems of our civilization (food crises, financial and economic crises, refugee crises, crises of democratic and human rights, climate-change crisis, armed conflicts…), contributing to the transformation of society away from an environmentally unsustainable system that is a generator of inequalities, exploitation, growing poverty and exclusion. It is now well-recognized that these crises, in particular, climate change, will have a significantly more devastating impact on the poor and developing nations than on the wealthy and developed nations.

Issue:
A/IS should contribute to achieving the UN Sustainable Development Goals.

Background

after the first paragraph:
As indicated in the chapter "Classical Ethics in A/IS" of this document, ethical frameworks and spiritual traditions of all times put stress on solidarity, inclusiveness and alleviating suffering. The ultimate goal of putting A/IS at the service of humanity, from a holistic, economic-prosperity perspective, should be at the core of A/IS ethics.

after the third paragraph:
It should not be forgotten that the basic research that makes technological innovation possible is largely financed and promoted using public funds (often taken from defense agency budgets), i.e. with contributions from all citizens, so that the benefits should also fall equally on all society.

Candidate Recommendations

- For innovation to benefit all humanity requires bringing together investment and knowledge from the public and private sectors. The universities can play an important role in bringing together distinct actors for this dialogue.

- The idiosyncrasies of A/IS in developing countries is an important line of research which merits more attention in the developed world. Emphasis needs to be placed on the research and development of applications in sustainability domains (management of natural resources, climate change, renewable energy), universal access to basic services (healthcare, transportation, etc), and promotion of democratic values, of human dignity, and of cultural diversity. This line of research would benefit from interdisciplinary teams involving specialists in A/IS as well as development experts.

- Regarding what type of indicators should be used, the chapter "Well-being" points to the need to use a wider set of indicators than has been usual to date to assess impact, since traditional metrics of prosperity do not take into account the full effect of AI technologies on human well-being. In

technological projects in LMICs, particular attention has long been paid to the quantification of impact through a wide variety of indicators, in particular, important indicators that are not mentioned in the aforementioned chapter, such as life expectancy, infant mortality, the possibility of leading a dignified life (through access to economic resources), access to knowledge (through access to education), percentage of the population living below the poverty line or in extreme poverty, etc. Given the diversity of realities covered by LMICs, in each geographical area, there will be a set of relevant specific indicators based on development priorities. It is also of interest to take into account indicators for different social groups, considering differentiating factors such as gender, ethnicity, etc. and their possible multiplicative effect. As already stated, the systems developed should themselves include modules for measuring impact, for example, collecting data of interaction with users via tracking methodologies, for further analysis also through AI and computational sustainability methods (while respecting privacy).

Further Resources

Anand, S. and A. K. Sen. (1996). "Sustainable Human Development: Concepts and Priorities". ODS Discussion Paper Series.

United Nations Development Programme (1990). "The Human Development Programme". 1990

United Nations Development Program (2017). Human Development Report 2016

United Nations (2014). "The Road to Dignity by 2030". Synthesis Report of the Secretary-General on the Post-2015 Agenda. United Nations, New York, 2014.

United Nations (2015). "Transforming Our World: the 2030 Agenda for Sustainable Development". United Nations Tech. rep.

United Nations (2018). "World Economic Situation and Prospects 2018", p. 205.

International Telecommunications Union (2017). "ICT facts and figures 2017". ITU

Issue:
It is unclear how developing nations can best implement A/IS via existing resources.

Background

At the beginning

Technological innovation in LMICs comes up against many obstacles:

- Patents, royalties, etc.
- Lack of the infrastructure and knowledge required to adapt technolgies to resolve the problems addressed by the SDG (Sustainable Development Goals).
- The difficulty of taking technological solutions to where they are needed.
- The lack of organisational and business models to adapt the technolgies to the specific needs of different regions.
- The lack of active participation of the target population.

At the end

In academic literature, many articles report on experiences of using A/IS en LMIC. From the 90s onwards, it was considered that expert systems held great promise for supporting policy making, for capacity building, and for providing universal access to basic services, some of these experiences integrating social, economic, political, cultural, religious and ethical dimensions in the engineering process and in the knowledge-elicitation methodologies. Expert systems (ESs) help to solve problems even in scenarios with uncertainty and/or when the system modelling may be very difficult or not even feasible. They allow hand-held field systems to be linked with real-time information sources, thereby enabling less well-trained people to perform technical tasks assisted by an ES. More than twenty-five years later, with the high penetration of mobile devices in LMICs and the wide diversification of ESs, these observations about the potential of ESs as an ICT for HSD in LMICs are even more pertinent.

ESs have the potential to make a significant contribution in key sectors such as agriculture, water resource management, health, education… since the problems to be solved in these sectors have characteristics that make them well-suited to the use of ESs or contain aspects which ESs have traditionally been used to address: the need for expert knowledge of a range of disciplines, the advisability of

combining traditional and modern techniques, the need to adapt solutions to cultural and environmental particularities, complexity and uncertainty, etc.

For example, in the literature there are articles reporting on experiences of using A/IS in agriculture, focusing on farming planning with a component of biodiversity and in training of the population; on water resource management, the primary goal of the applications being the sustainable management of the water cycle by the communities involved, accompanied by access to, and training to use, ICT systems for prevention of weather-related natural disasters; in the health sector supporting a universal health-care system with a focus on the development of policies, the generation and maintenance of health databases, and epidemiology research.

Another area of A/IS that has shown great possibility in the development context is Big Data, as is reflected in the 2012 UN white paper "Big Data for Development: Opportunities and Challenges".

Difficulties in developing ESs are scarcity of expert knowledge which, if present, may be scattered or informal, since there are unlikely to be many experts and they are likely to be poorly-trained. On the other hand, ESs are well-suited to LMICs in the sense that they do not normally require a powerful hardware or software base and because they can take advantage of the surprisingly widespread deployment of mobile telephony in LMICs

As far as applications of big data are concerned, there is a dearth of management and information resources, data repositories, etc., a problem that is worsened by weak infrastructures and poor communication systems. Having disperse communities in these countries makes it more difficult to implement solutions, to access the target population and, when this is required, to train the future users of the systems.

A review of the literature shows that applications that had been designed and implemented as part of a large, institutional cooperation-for-development project satisfy the criteria for good practice in development interventions to an acceptable degree, while in small-scale initiatives, research interests tend to take precedence over those of HSD. However, even in those projects explicitly concerned with HSD, the objectives usually betray a rather narrow perspective. Although they define overall goals related to HSD, in general, these concern only productivity and environmental sustainability, not fairness or empowerment. Moreover, impact

indicators are not generally established and studies of synergies with other HSD objectives are lacking. Lastly, criteria for usability and accessibility in the particular circumstances of each development context are rarely addressed.

Although the use of ES is diverse in LMICs, many promising application fields remain unexplored. One reason for this is a lack of NGDO involvement; most applications are promoted by international cooperation institutions, government entities and by private entities in collaboration with research centers. Another problem is that even though the recipients are mainly national institutions and specialized technical departments, they were not queried about the possible applications of interest prior to the project's initiation.

Finally, projects are frequently developed focused on a particular institution rather than on a certain group of people with particular needs. And while projects for institutional strengthening may be useful, it is important that the benefits are not limited to providing services to the institution, with a negligible incidence on society at large.

With respect to the process of eliciting and conceptualizing knowledge, we highlight the difficulty in extracting knowledge from geographically-dispersed experts of different cultural levels, the lack of access to specialized scientific literature, and the poor data records available in the projects reviewed. It is noteworthy that in the applications, rather than local domain experts being the primary source of knowledge, use of scientific, First-World knowledge is common.

Candidate Recommendations

- LMICs demand a specific methodological and technical approach, inspired by the "appropriate ICTs" concept and on the criteria for good practice in development interventions, as yet not covered by the most holistic development A/IS methodologies.
- Existing methodological and technical tools that would facilitate the control of the factors behind the success of A/I S deployments in LMIC should be considered, such as the “Logical Framework Approach”, an internationally widely-accepted methodological approach to development projects. Its specialization to AI/S projects needs further study. It is widely recognized that development actions must be comprehensive, consideration which, when applied to Expert Systems in LMIC contexts, implies that multiple fields of

expertise may be involved. The integral perspective of the LFA is of great importance in the development context since it implies determining not only needs, but also the causes of the problems leading to these needs, the strengths and weaknesses of the potential solutions, the synergies with other projects, etc.

- Reuse and standards-based engineering is also relevant in this context. Applications can benefit from a wide range of reuse approaches such as AI problem-solving methods and tasks, as well as ontologies, object-oriented analysis and design patterns, frameworks, components, open architectures… The use of standards favours the interoperability, accessibility and usability of applications.
- Regarding user-centred and participation-based engineering, the techniques for context analysis and modeling in context-aware and culture-aware system development, user-modeling, etc., may have a decisive role in the LMIC context. The close involvement of the recipient community improves the quality of the system by increasing the likelihood of accurately capturing the right requirements, the importance of this involvement increasing with system complexity. Moreover, greater understanding of the system by the users leads to greater acceptance of the system, acceptance being a vulnerability of Expert Systems in general which may be exacerbated in LMIC contexts, and thereby results in more effective use. This also shows the importance of the knowledge engineers having multidisciplinary and multicultural abilities.
- Reproducible research techniques are also relevant. Reproducibility should be one of the main principles of any innovative project. Even for context-aware and culture-aware applications, the techniques themselves should be reproducible. Reproducible research is desirable for many reasons. Firstly, it facilitates checking the accuracy of statements made in research reports by allowing other researchers to easily replicate experiments. Of importance in a development context is the fact that it broadens the impact and helps in the continuity of projects. Reproducibility does not preclude the portability, scalability and customizability of the solutions proposed.
- It has long been claimed that openness, where this refers not only to open-source but to openness in general, is a desirable characteristic for software

projects in LMICs (while, at the same time, recognising that, for a variety of reasons, the potential of open-source software in LMICs to date has not been fully realized).

- Mobile applications have shown great potential for helping to close the digital divide, due to the wide dissemination of increasingly-affordable internet-enabled phones in LMIC contexts.
- Finally, we add two aspects already highlighted in other chapters of this document: user-centred and participation-based engineering. As far as applications in LMIC are concerned, emphasis should be placed on ethnographic approaches and contextual design. The concept of context-aware system, of particular importance in the field of ubiquitous computing, refers to the ability to detect, and adapt to, context-dependent aspects such as location, time, noise-level, lighting, network availability… In culture-aware AI systems, culture-related information is modelled and used to design human-machine interfaces or to intervene somehow in the tasks carried out by the system, two of the most well-known types of culture-aware systems being culture-aware intelligent tutoring systems and cross-cultural decision support systems. Note that part of cultural-awareness is simply trying to avoid imposing the cultural assumptions of the development team or of the technology used. Note also that culture-awareness must strive to not to reinforce any discriminatory practices, for which reason, human rights must figure in the indicators used to measure success.
- Take advantage of all the energy and creative potential for innovation: entrepreneurship, not according to the myth of the individual entrepreneur but through the creation of an infrastructure, a connecting fabric, in which the university can play a crucial role in the dialogue between the different actors.
- In A/IS research projects in LMICS, care must be taken that research interests do not take precedence of those of HSD. The projects should have a wide perspective and define general goals related to HSD, concerning not only productivity and environmental sustainability, but also fairness or empowerment. Impact indicators need to be established and studies of synergies with other HSD objectives need to be studied. Criteria for usability and accessibility in the particular circumstances of each development context

have to be addressed. Applications focused on a particular institution rather than on a certain group of people with particular needs should be avoided. While projects for institutional strengthening may be useful, it is important that the benefits are not limited to providing services to the institution, with a negligible incidence on society at large.

- Promising application fields that remain unexplored should be explored
- International cooperation institutions government entities or private entities, in collaboration with research centers. It is advisable to consult national and international institutions about target areas for HSD, and partnerships with NGDO should be encouraged, given these organisations' proven professionalism
- With respect to the process of eliciting and conceptualizing knowledge, local domain experts have to be the primary source of knowledge instead of scientific, First-World knowledge
- Effort needs to be made in data gathering.

Further Resources

Schumacher, E. F. (1973). "Small Is Beautiful. A Study of Economics as if People Mattered". Blond & Briggs.

United States Agency for International Development (1970). "The Logical Framework Approach". Final Report, Contract csd-2510. Tech. rep. United States Agency for International Development.

World Association of Non-Governmental Organizations (2002). "Code of Ethics and Conduct for NGOs". Tech. rep. WANGO.

Bostrom, N. (2017). "Strategic implications of openness in AI development". In: Global Policy 8.2, pp. 135–148.

Reijswoud, V. van and E. Mulo (2012). "Evaluating the Potential of Free and Open Source Software in the Developing World". In: Int. J. Open Source Softw. Process.

Touray, K. S. (2004). "Constraints against the adoption and use of FOSS in developing countries". linux.com.

Donoho, D. L. (2010). "An Invitation to Reproducible Computational Research". Biostatistics 11(3).

Fomel, S. and J. F. Claerbout (2009). "Reproducible research". Computing in Science & Engineering 11(1)

Peng, R. D. (2011). "Reproducible Research in Computational Science". In: Science (New York, Ny) 334.6060, p. 1226.

Heimgärtner, R. (2018). "Culturally-Aware HCI Systems". In Advances in Culturally-Aware Intelligent Systems and in Cross-Cultural Psychological Studies. Springer, Cham.

Nye, B. D. (2015). "Intelligent tutoring systems by and for the developing world: A review of trends and approaches for educational technology in a global context". International Journal of Artificial Intelligence in Education, 25(2), 177-203.

Eagle, N. and Horvitz, E. (2010). AAAI Spring Symposium on Artificial Intelligence for Development. Tech. rep. Association for the Advancement of Artificial Intelligence and Stanford University.

Frias-Martinez, V., J. Quinn, and L. Subramanian (2014). "Computational Sustainability and Artificial Intelligence in the Developing World". AI Magazine Special Issue on Computational Sustainability.

Manjarrés Riesco, A., M. A. Artaso, M. Belandia, and S. Pickin (2014). "Still in the First Steps of a Barefoot Knowledge Engineering". Computer Applications & Research (WSCAR), 2014 World Symposium on. IEEE, pp. 1–6.

Mann, C. and S. Ruth (1992). "Expert systems in developing countries: practice and promise". Westview special studies in social, political, and economic development. Westview Press.

Giupponi, C., & Sgobbi, A. (2013). "Decision support systems for water resources management in developing countries: Learning from experiences in Africa". Water, 5(2).

Papathanasiou, J., Ploskas, N., & Linden, I. (Eds.) (2016). "Real-world Decision Support Systems: Case Studies" Integrated Series In Information Systems, Vol. 37. Springer.

Hilbert, M. (2016). "Big data for development: A review of promises and challenges". Development Policy Review, 34(1).

United Nations Global Pulse (2012) "Big Data for Development: Opportunities and Challenges"

United Nations, Sustainable Development Goal indicators. https://unstats.un.org/sdgs

"World Bank Offers First Sustainable Growth Bonds for Private Investors in Switzerland". http://treasury.worldbank.org/cmd/htm/World-Bank-Offers-First-Sustainable-Growth-Bonds-for-Private-Investors-in-Switzerland.html

Global Impact Investing Network (GIIN), https://thegiin.org/

Issue:
The complexities of employment are being neglected regarding A/IS.

Candidate Recommendations

Promote policies of employment with social value which cannot be supplanted by A/IS technology.

Issue:
Technological change is happening too fast for existing methods of (re)training the workforce.

Candidate Recommendations

To integrate with the third recommendation:

- Governments and universities should promote and give support and visibility to private social-innovation, social-enterprise and social-entrepreneurship initiatives in the area of A/IS, focused on sustainable development. This would promote pilot projects which, if successful could scale up. These initiatives could come from companies, NGOs or other social organisations.

Further Resources

"Social enterprises", European Commission. http://ec.europa.eu/growth/sectors/social-economy/enterprises_en

Ebrahim, A., Battilana, J., Mair, J. (2014). "The Governance of Social Enterprises: Mission Drift and Accountability Challenges in Hybrid Organizations". Research in Organizational Behavior, 34.

Ebrahim, A., Rangan, V.K. (2014). "What Impact? A Framework for Measuring the Scale & Scope of Social Performance". California Management Review, 56 (3).

Kroeger, A., Weber, C. (2014). "Developing a Conceptual Framework for Comparing Social Value Creation". Academy of Management Review, 39 (4).

Nicholls, A. (2009). "'We Do Good Things, Don't We?': Blended Value Accounting in Social Entrepreneurship". Accounting, Organizations and Society, 34 (6-7).

Issue:
How best to incorporate the "global dimension of engineering" approach in undergraduate and postgraduate education in A/IS.

Candidate Recommendations

- Implement "International Service Learning" programs in universities. Service-Learning denotes a set of programs or activities of solidarity service carried out by the students that are oriented to meet the needs of a community, but are planned in an integrated manner with the contents of the curriculum, in order to optimize the learning. On the American continent, service learning has been incorporated in all the educational spaces and, in particular, the higher education in technological areas, though not yet in A/IS studies.

- It is the responsibility of the universities to educate public opinion regarding the challenges presented by A/IS and its potential. The universities need to speak to society of the desirable, inclusive, positive scenarios that A/IS can bring, that is, those of the UN Sustainable Development Goals, as opposed to the dystopias presented by some science-fiction.

Further Resources

Jacoby, B. (2014). "Service-Learning Essentials: Questions, Answers, and Lessons Learned". John Wiley & Sons.

Comments on the Version 2 of the Ethically Aligned Designed – A Vision for Prioritizing Human Well- being with Autonomous and Intelligent Systems

Tomas Jucha; Investment Protection and AI Policy Advisor assisting Deputy Prime Minister's Office of the Slovak Republic for Investment and Informatization on AI Policy Matters; comments provided in personal capacity; pp 188 – 191 of the Draft EAD - version 2 document

Dear colleagues from The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems ("The IEEE Global Initiative"),

Many thanks for the excellent work being done and your efforts in this regard. It's my honor as a Member of the IEEE Global Initiative to provide brief comments to the perfectly, timely and comprehensively drafted *Version 2 of the Ethically Aligned Designed – A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems* ("Draft EAD - version 2 document"). I duly understand and respect the guidelines concerning submission of the relevant comments therefore try to be brief and efficient as possible. Given my area of focus, I would address 2 matters which concerns Policy making concerning Autonomous and Intelligent Systems (A/IS).

Add objective: Provide effective regulation of A/IS to ensure public safety and responsibility while fostering a robust AI industry, pp 188-189

Firstly, I would recommend to add the following adjectives (in bold) "**prompt, dynamic and internationally enforceable**" to the title of the chapter on the page 188 which states as follows "*Provide effective, **prompt, dynamic and internationally enforceable** regulation of A/IS to ensure public safety and responsibility while fostering a robust AI industry*". While it may be that words "effective" and "ensure" in the original title capture all or some of the added adjectives and their meaning (however, this is not clear from the detailed explanations provided in the follow up text on pages 188 and 189), given the importance they have I would recommend to add them explicitly. The reason for adding "**prompt and dynamic**" lies in the specific nature of the A/IS, which emerge and develop in the unprecedented exponential way. In order for the policy makers to be able to set forth the appropriate level playing field, the respective regulation needs to be dynamic and flexible enough to comprehensively, proactively

and timely govern the rights and relations of all stakeholders in the society affected by such A/IS. The traditional means of policy and regulation making require reconsideration in order to keep the pace with this this fast changing and increased impact having technologies. Creation of so called uniform regulatory sandboxes as called upon in the Visegrad 4 countries' thoughts on the Artificial Intelligence may be one of the solutions. The effective regulation is e.g. being pointed out by Microsoft in its current publication The future computed – Artificial Intelligence and its role in society, as well discussed during the World Economic Forum 2018. The speakers at latter event recognized, that we are facing more questions than answers on topic how to regulate AI development, and emerging technologies in general, where the initiatives that would monitor what's happening on the frontiers of AI or initiative that would be a platform for nations to exchange information about implementing and regulating AI have been discussed (an as rightly emphasized in the Draft EAD - version 2 document).

The addition of adjective "**internationally enforceable**" regulation is of particular and separate importance given the transnational nature of A/IS and the increased major investments and attention on A/IS research and development worldwide. Since the A/IS designed in one country can be easily transferred, deployed and used in another country, in addition to regulations on local and national level, adoption of the internationally recognized regulation and principles would be strongly advisable accompanying at the same time by the effective enforcement mechanisms to ensure that such regulation and principles of Ethically aligned design of A/IS are observed by all the relevant parties worldwide. Otherwise it would be practically impossible to "ensure public safety and responsibility" as called upon by one of the main objectives in the Police chapter in the Draft EAD - version 2 document on page 188.

Given the above, concrete adjustment to the Candidate Recommendations on page 189, second bullet point may be added by the text (in bold) *"To foster a safe international community of A/IS users, policymakers should take similar work being carried out around the world into consideration. Due to the transnational nature of A/IS, globally synchronized policies can have a greater impact on public safety and technological innovation* ***and therefore it is strongly advisable to adopt internationally recognized and enforceable policies, regulations and principles of Ethically aligned design of I/AS****."*

Add objective: Facilitate public understanding of the rewards and risks of A/IS, pp 190 - 191

Secondly and lastly, I would recommend to explicitly mention the importance of Technology assessment centers (such as TA-SWISS) that deal with the possible consequences (opportunities and risks) of using new technologies on social coexistence and on the environment. The chapter covers mainly the questions of understanding of the "social and ethical implications of the A/IS", and just briefly mentioned the "the best practices for using and developing A/IS". On one side, there is a great availability of new research and assessments of technology use and its impact on the individual's life, health and the society development and values in general, e.g. on spending a proper amount of time before screen devices, using social networks on a daily bases while replacing personal contact with the relatives, excessive reliance on the mobile devices while performing daily tasks, etc., however, one the other side, lack of the awareness and insufficient active role of the industry, nongovernmental organizations and governments per se that do not harness the benefits of such reported results and its recommendations for optimal technology use. The work of technology assessment centers and recommendations for the optimal use of technology can significantly help the individuals to live and experience a balanced and more fulfilling lives.

Therefore, I would recommend to stress more the role of the technology assessment centers and the importance of building awareness on the proper and responsible technological use in this chapter, e.g. by adding the text (in bold) on p. 190 *"If society approaches these technologies primarily with fear and suspicion, societal resistance may result, impeding important work on ensuring the safety and reliability of A/IS technologies. On the other hand, if society is informed of the positive contributions,* ***optimal use*** *and the opportunities A/IS create, then the technologies emerging from the field could profoundly transform society for the better in the coming decades"* and on p. 191 *"Empower and enable independent* ***technology assessment centers****, journalists and media outlets to report on A/IS* ***and its use****, both by providing access to technical expertise and funding for independent journalism.".*

I hereby declare that the comments above are kindly provided in my personal capacity and do not represent any policy, statement or position of any organization,

institution or group representing any interest. I also hereby agree that all of the information provided in this document can be published for the purposes as suggested in the Submission Guidelines for Ethically Aligned Design, Version 2 dated December 12, 2017.

Bratislava, May 7, 2018

Dr Ozlem Ulgen* – Feedback submission for IEEE Report *Ethically Aligned Design (V2)*

* Senior Lecturer in Law, School of Law, Birmingham City University, UK, and Visiting Fellow at Wolfson College, University of Cambridge.

I would recommend to:

pp. 7, 152 and 155 - consider/include whether the public have a right to know if they are interacting with a human or a bot (e.g. issues of informed consent; privacy; attributing harmful/wrongful behaviour).

p. 33 - re-phrase or explain "levels of trust": "trust" has specific philosophical, legal, and engineering connotations. It is attributable to humans, and relates to claims and actions people make. Machines, robots, and algorithms lack the ability to make claims and so cannot be attributed with trust. They cannot determine whether something is trustworthy or not. Software engineers may refer to "trusting" the data, but this relates to the data's authenticity and veracity to ensure software performance. In the context of lethal autonomous weapons perhaps the meaning and effect of "trust" is reflected in the term "functional reliability"; that there is confidence in the technology's predictability, reliability, and compliance with international humanitarian law so that machines, robots, or algorithms perform tasks for the set purpose without error or minimal error that is acceptable.

p.113 - include "capable of causing physical harm": "capable" is probably better than simply "design" because it could be argued that the weapon was not designed to cause physical harm and yet it actually achieves that effect. See: Ulgen O., "Definition and Regulation of LAWS" (UN GGE LAWS Report, 5 April 2018) 1-24 at: https://www.unog.ch/__80256ee600585943.nsf/(httpPages)/7c335e71dfcb29d1c1258243003e8724?OpenDocument&ExpandSection=6#_Section6

p.113 - distinguish between "automated" and "autonomous": automated functions of a weapons system suggests pre-programmed actions involving human input, whereas autonomous suggests a capability separate and independent from human input. It is important to distinguish between the two in order to determine the degree of weapons autonomy and the extent of human involvement. See: Ulgen O., "Definition and Regulation of LAWS" (UN GGE LAWS Report, 5 April 2018) 1-24 at: https://www.unog.ch/__80256ee600585943.nsf/(httpPages)/7c335e71dfcb29d1c1258243003e8724?OpenDocument&ExpandSection=6#_Section6; Ulgen O., "'World Community

Interest' approach to interim measures on 'robot weapons': revisiting the Nuclear Test Cases" (2016) 14 *New Zealand Yearbook of International Law* 3-34 https://brill.com/view/title/34877

pp. 113, **117 and 121**- use "human control" or "human involvement" as less ambiguous and contested than "meaningful human control". This could then be linked to the degree of weapons autonomy within the critical functions to determine what sort of human control/involvement is necessary or appropriate. See ibid.

p. 113 and 117 - specify the principles, requirements, and obligations under international humanitarian law applicable to AWS. See: Ulgen O., "Pre-deployment common law duty of care and Article 36 obligations in relation to autonomous weapons: interface between domestic law and international humanitarian law?" (2016/17) 55(1) *The Military Law and the Law of War Review* 1-29 http://www.open-access.bcu.ac.uk/5581/; Ulgen O., "Human Dignity in an Age of Autonomous Weapons: Are We in Danger of Losing an 'Elementary Consideration of Humanity'?" (2016) 8(9) *ESIL Conference Paper Series* 1-19 European Society of International Law (ESIL) 2016 Annual Conference (Riga). Published on January 31 2017, ESIL SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2912002 , also in I. Ziemele and G. Ulrich (eds.), *How International Law Works in Times of Crisis* (OUP 2018); Ulgen O., "Definition and Regulation of LAWS" (UN GGE LAWS Report, 5 April 2018) 1-24 at: https://www.unog.ch/__80256ee600585943.nsf/(httpPages)/7c335e71dfcb29d1c1258243003e8724?OpenDocument&ExpandSection=6#_Section6

p.117 - specify which law(s) are relevant to "evaluate the conformity of a system to the law".

Dear EAD team,

following discussion at Ethically Designed (EAD) Workspace about the topic "Immutable purpose of A/IS applications" Robert Donaldson raised few days ago, we propose to input the Committee regarding the general principles driving Ethically Aligned Design :

Input to consider assessment is about "Purpose".
It may be a new Principle or to adapt some of the existing Principles. That needs to be assessed.

In the very short timeframe we are unable to provide the Committee with a complete mature description as the other Principles are worded. Nevertheless following John Havens suggestion, we submit for further development. The topic "Immutable purpose of A/IS applications" at EAD workspace already provide a bit background in a synthetic way.

We'll be glad to develop accordingly the assessment by the Committee.

"Purpose" consists concretely with 3 statements:

1. A/IS dynamic systems be required to have a known purpose that defines it through requirements definition, build, testing, deployment and self-evolving configuration at all phases of operation with appropriate monitoring.

2. The concept of a business owner with direct accountability for the purpose of the A/IS be required for all A/IS development, operation and modification.

3. Designers and developers of A/IS have to develop the solution with built in precautions to secure inflight change accordingly ethic and safety (with human/humankind as priority to protect).

Kind regards,

Yannick Fourastier
Head of Industry 4.0 Solutions & Services
Bombardier Global, Germany

Bob Donaldson

Yannick Fourastier
Head of Industry 4.0 Solutions & Services
Bombardier Global, Germany

Principle 2:

Candidate Recommendation

A/IS should prioritize human well-being as a stated part of the business purpose and outcome in all system designs, using the best available, and widely accepted, well-being metrics as their reference point.

Principle 3:

Issue:

How can we assure that designers, manufacturers, owners, and operators of A/IS are responsible and accountable?

Background

The programming, output, and purpose of A/IS are often not discernible by the general public. Based on the cultural context, application, and use of A/IS, people and institutions need clarity around the purpose, manufacture and deployment of these systems to establish responsibility and accountability, and avoid potential harm. Additionally, manufacturers of these systems must be able to provide programmatic-level accountability proving why a system operates in certain ways to address legal issues of culpability, if necessary apportion culpability among several responsible designers, manufacturers, owners, and/or operators, to avoid confusion or fear within the general public. Note that accountability is enhanced with transparency, thus this principle is closely linked with Principle 4 — Transparency.

Candidate Recommendations

To best address issues of responsibility and accountability:

1. Legislatures/courts should clarify issues of responsibility, culpability, liability, and accountability for A/IS where possible during development and deployment (so that manufacturers and users understand their rights and obligations).

2. Designers and developers of A/IS should remain aware of, and take into account when relevant, the diversity of existing cultural norms among the groups of users of these A/IS.
3. Multi-stakeholder ecosystems should be developed to help create norms (which can mature to best practices and laws) where they do not exist because A/IS-oriented technology and their impacts are too new (including representatives of civil society, law enforcement, insurers, manufacturers, engineers, lawyers, etc.).
4. Systems for registration and record-keeping should be created so that it is always possible to find out who is legally responsible for a particular A/IS. Manufacturers/operators/owners of A/IS should register key, high-level parameters, including but not limited to:
    - Intended use including the scope of that use
    - Training data/training environment (if applicable)
    - Sensors/real world data sources
    - Algorithms
    - Process graphs
    - Model features (at various levels)
    - User interfaces
    - Actuators/outputs
    - Optimization goal/loss function/reward function
5. A/IS **dynamic systems** be required to have a **known purpose** that defines it through requirements definition, build, testing, deployment **and self-evolving configuration at all phases of operation with appropriate monitoring**.
6. The concept of a business owner with direct accountability for the purpose of the A/IS be required for all A/IS development, **operation and modification**. Where the business owner and the operator of the application are different, the accountability for the outcomes of the application shall require an appropriate legal instrument, to define and assign accountability for the ethical usage of the application, between the parties.

Principle 4:

Candidate Recommendation

1. Develop new standards* that describe measurable, testable levels of transparency, so that systems can be objectively assessed against their stated purpose and level of compliance determined. For designers, such standards will provide a guide for self-assessing transparency during development and suggest mechanisms for improving transparency. (The mechanisms by which

transparency is provided will vary significantly, for instance 1) for users of care or domestic robots, a why-did-you-do-that button which, when pressed, causes the robot to explain the action it just took, 2) for validation or certification agencies, the algorithms underlying the A/IS and how they have been verified, and 3) for accident investigators, secure storage of sensor and internal state data, comparable to a flight data recorder or black box.)

2. Designers and developers of A/IS should **have to** develop the solution with built in precautions **to secure the application against** inflight change **according to ethical and safety objectives (with human/humankind as priority to protect)**.

3. Designers and developers should provide a "what do you recommend functionality" for the user to request before output the answer, recommendation or action of the A/IS.

Principle 5:

Candidate Recommendations

Raise public awareness around the issues of potential A/IS technology misuse in an informed and measured way by:

1. Providing ethics education and security awareness that sensitizes society to the potential risks of misuse of A/IS (e.g., by providing "data privacy" warnings that some smart devices will collect their user's personal data).
2. Delivering this education in scalable and effective ways, beginning with those having the greatest credibility and impact that also minimize generalized (e.g., non-productive) fear about A/IS (e.g., via credible research institutions or think tanks via social media such as Facebook or YouTube).
3. Educating government, lawmakers, and enforcement agencies surrounding these issues so citizens work collaboratively with them to avoid fear or confusion (e.g., in the same way police officers have given public safety lectures in schools for years; in the near future they could provide workshops on safe A/IS).
4. Provide ethics (EAD) education and security awareness to IEEE societies as they can become evangelists for ethically aligned design within their disciplines, corporation's professional organizations and governments.

Emil P. Vlad P. Eng.

1. on p.9, first para. from "Well-being Promoted by Economic Effects " - Somewhere in this paragraph some essential concrete wellbeing attributes should be mentioned (at least as examples within brackets) like health, safety (cyber)security. Especially considering these are key core humankind wellbeing attributes that are also ethically valued across the globe; also considering these are the key desired non-functional attributes/properties currently regulated and engineered in existing (currently in use or development) automation/automated systems.

2. on p.10, first para. from "Policies for Education and Awareness" - Also physical not only cybersecurity should be mentioned in this list as equally important; also health and environmental protection too.

3. on p.11, first para. from "Well-being Metrics" - For consistency with my previous comment on the related Well-being paragraph from section III; I suggest to add (cyber)security to the list of key wellbeing metrics, which are already managed and known in the context of existing automation systems.

4. on p.14, first para. from "Educational materials" - This metaphoric "Evergreen in nature" expression, should be replaced by a less metaphorical one that is more precise/accurate and literal; considering this is a document from an international engineering organizations and for many members and even so more users English it's not maternal language so literary subtleties obfuscate the semantic clarity.

5. on p.23, first para. from "General Principles" - I suggest to add on this page or somewhere considered appropriate in this section as early statement reflecting maybe in more detail the principle that: because A/IS are a subset of automated systems: the autonomous, and highly intelligent ones; all applicable ethical and dependability rules and practices know from the development and use or exiting automated systems should be applicable by extrapolation and adaptation. This would be a legitimate and important for two reasons

- it is a legitimate application of the existing GAMAB system safety principle;
- it would encourage engineers to build upon (instead reinventing the wheel) exiting knowledge and practices accumulated over at least a couple of decades of automation by avoiding mistakes that were paid dearly with accidents"

6. on p.34, first para. from "Issue: How can we extend the benefits and minimize the risks of A/IS technology being misused?" - Safety is barely mentioned here on this page, however it is very important and should be added besides security. Security is intentional misuse but safety covers unintentional misuse.

7. on p.49, first para. from "Embedding Values into Autonomous Intelligent Systems, regarding Transparency as intelligibility" - Although the AI design/algorithm/solution may be totally transparent it might unintelligible to humans. Thus intelligibility might be very limited in some very advanced AIs, as it is already the case in deep neural networks. The smarter the AI becomes the more complex their design patterns and algorithms will intrinsically be. Practically is quite likely impossible to achieve because in some applications the reason AI is used is exactly to address extremely complex problems well beyond human ability to solve; these problems may very likely equally incomprehensible solutions to humans. As long as all affected humans are informed and participate in taking the decision, this is more of a risk/benefit trade off issue than an ethical issue.

8. on p.50, first para. regarding "fail-safe" -The fail-safe principle is widely used in many automation systems currently so it will be applicable to AI systems as well, but this should be reformulate by also adding operate-safe as some (e.g. airplane FMS) automated system cause accidents if they fail to operate. They have to actively operate safe in some back-up or degraded mode but complete failure (system stopped) is not safe option.

9. on p.53, first para. from "Evaluating the Implementation of A/IS" - I suggest to add in this section a sentence about the applicability of the already largely used PDCA organization management system practices (maybe reference some of the related well known ISO std. 9001, 18001, 14001, etc.) allowing for systematic assurance of desired goals and

attributes for products. These are applied to existing automation products so all that experience should be applicable and adapted to A/IS.

10. on p.58, second para. from "Methodologies to Guide Ethical Research and Design, " - Noble intent in theory but I think it is unrealistically difficult to be applied in practice in the current geopolitical world. As it doesn't seem reasonable to expect that rapid increase in automation and robotization will slow down and wait for the world to become one, I suggest to add or complement this section with some more practical guidelines (similar to existing configurable systems) and principles along the following lines:

    - A/IS should be developed and produced ethically neutral by engineers & suppliers which should strive to do that to the largest extent possible;
    - A/IS should be designed to be highly configurable in general but in special with respect with the ethical values;
    - A/IS should be delivered to users with comprehensive manuals and training including about the configuration of the desired ethical values to be loaded as parameters by the users."

11. on p.61, first para. from "Issue: The need to differentiate culturally distinctive value embedded in AI design." - As mentioned in a previous comment considering the variation of ethical norms across the globe, this could be practically done by "ethically" configurable A/IS.

12. on p.62, first para. from "Methodologies to Guide Ethical Research and Design" - A suggested guideline should be added here to address this issue stating: that functionally generic and parametrized A/IS are more neutral ethically thus transferring as much as possible the ethical responsibility of the technology (A/IS) to the user; the ideal to strive for should be for example, that of a knife which is technology (albeit old and rudimentary now, it was revolutionary when invented/discovered) is ethically neutral and it can be used for both good and bad depending entirely of the user's ethics.

13. on p.67, first para. from "Lack of ownership or responsibility from the tech community." - This should be removed or rephrased because legal is different than ethical, while I don't think that safety should be taken for

granted; as it's not a given and tremendous effort is required to instill it in organizations and products currently.

14. on p.71, first para. from "Issue: Poor documentation hinders ethical design." - That is true but it should be added this situation applicable to A/IS is no different than the current discrepancy in most complex systems and automation systems today: in theory there are regulation and guidelines for documenting the systems (sw, hw, etc.) but in practice it either lacks, or it's incomplete or inaccurate, especially for SW. This is due to the fast pace of technological change and prioritizing resources (schedule and cost) in enterprises.

15. on p.72, first para. from "Issue: Inconsistent or lacking oversight for algorithms. " - That is true, but practically what is the guideline for already existing cases (deep neural networks, evolutionary computing) where the A/IS algorithm is just inherently so complex that humans cannot comprehend it. This poses a crucial risk/benefit dilemma as one of the main reasons for developing A/IS is the belief that it will be capable to solve very complex problems that humans (even in large teams) can't.

16. on p.82, first para. from "Candidate Recommendation"- That is recommendable in theory, but in practice even in current complex and automated systems not always done; often not because of engineers, but because managers "myopia" justified by schedule and cost reasons. Because these are in turn driven by customer and market competitive pressure, I think the most effective way to counteract is through legal economical means/measures that would internalize (currently externalities from a market perspective) health, safety environment risk/accident costs. If these basic ethics attributes cannot be instilled in engineered systems how will more general ethics values be?

17. on p.106, fourth para. regarding "service may be degraded" - This should be generalized beyond “personal data” into a general automation (widely applied currently) principle, by adding here or in a more general section of this document a statement like: wherever possible (the control or search algorithm isn't beyond human capability) the A/IS should have a manual

bypass degraded mode functionality allowing users to override it, whenever necessary in exceptional circumstances.

18. on p.199, fourth para. regarding "free will" - However, based on recent scientific free will seems to be more constrained (deterministically or statistically) laws of the universe than, construed in classical philosophy and psychology.

Based on my direct and indirect (reading from other experts) experience, AI is just an extremely advanced form of automation (i.e. autonomous automation). So many of the dilemmas, dilemmas and other paradoxes encountered in AI are just maybe more sophisticate variants of existing ones from the general industrial and automation fields. Unfortunately some of these have still not found optimal solutions, so these seemingly rushed development of AI (maybe due to an accelerated capitalism growth principle) may just aggravate the situation and create some new dilemma variants. I think humans have a pretty good understanding of risk and it's management, but there is a fundamental problem resulting from the fact that generally those taking the risks (decision factors) are few at the top while those bearing the consequences of risks taken are many more towards the bottom of the hierarchical social pyramid.

Ethics depend on culture and usually are developed to serve the social system (implicitly the ruling elite) of each respective civilization so they are relative although not entirely arbitrary. We are evolving towards a planetary global citizen ethic but humankind has not achieved that yet. I hope we will but there are still risks and dangers ahead.

Scientists discover the laws of nature then engineers apply those by inventing useful technologies which can then be used for some purpose by the human owner/user within the political limits imposed through laws and ethics of the country. A rational sane human would use any tool in its own/group interest. But interests could be often conflicting so good for someone may be bad for someone else.

Robots that have their controller based on classic control theory or AI (specialized not generalized one) are just advanced automation tool. Automation is an engineering technology which like any technology is essentially ethically neutral.

Assuming the robot control and decision algorithms as parametric (as it could and should be) then the owner is capable and responsible for configuring the robot with as desired within the legal and technology limits. Robots that have their controller based on classic control theory or AI are just advanced automation tool.
Automation is an engineering technology which like any technology is in essence ethically neutral. The application of the technology is characterized by the ethics of its owner/user/buyer (it depends). So if the robot is a retail household robot then the end user is the owner, then there is no medical purpose builtin so the robot should do whatever the user orders as long as the robot safety operates. If the robot was manufactured to be a robotic nurse replacement and it is owned by a healthcare organization then it is their responsibility to configure it to behave as a nurse which means besides operating safely, it should also take care of the health of the patient, so shouldn't give him junk food or whatever would negatively affect the patient health.

If the ageing of the population continues there isn't a better solution, as the society won't be able to bear otherwise the burden of some many pensioners that aren't productive, yet need company, help and healthcare.
As long as the robots are made reasonably dependable (safe, secure, reliable, etc.) as other current critical medical technology is, it should be an acceptable practical and intelligent solution to the problem. Like with any new or especially revolutionary technology there will take some time for people to get accustomed with it but after that it should become the new normal.

I think fully autonomous robots if dependably made could and should do the job autonomously. The acceptable level of dependability should be (as they are for current critical technology) determined by some legal regulatory standard form. Of course the current legal, regulatory and standardization is too slow to evolve so it will have to improve or the robotization should slow down otherwise.

Humans developed technology some of which can be used for weapons and had wars for thousands of years. Automation and AI are just the most advanced

technology being developed and like any other technology before it is being used for both helping and harming humans. Robotic weapons are developed because of a ("arms race" coevolution cybernetics principle) positive feedback loop between two teleological systems. I hope that as with nuclear weapons the development of LAWS will lead to a new equilibrium level (stalemate).

Even when talking about the generalized AI (or singularity) I hope that it will help us solve some existential very complex planetary problems (most of our own making) as it doesn't seem to me that humankind is capable to solve them timely.

I think that the risk of humankind extinction is otherwise high anyway: if not because of the AI, then because of other artificial catastrophes caused by humans. In besides agreeing that some serious effort has to made to make AI dependable and beneficial to humans, there might be a silver lining in the worry that AI will destroy humankind: it may save it from itself by finding complex solutions to the very complex problems caused from which some are apparently beyond humans ability to solve with existing technology not enabled by AI.

I'd be interested to further discuss and contribute to this initiative and topic.

Submission on the document: *Ethically aligned design: A vision for prioritising human well-being and autonomous and intelligent systems (A/IS)*. Version 2. IEEE.

Erica Southgate, PhD, Associate Professor of Education, University of Newcastle, Australia. Research profile - https://www.newcastle.edu.au/profile/erica-southgate

1. **There is a lack of definitional clarity on the concepts of ethics and ethical decision-making; inadequate background on the philosophical and cultural traditions/models on which the idea of 'ethics' is based; and a false assumption about knowledge 'input' for A/IS ethical decision-making that has a Western bias.**

There are many philosophical and cultural ethical traditions (for example, in the West there are theories about virtue ethics, consequentialist ethics, applied ethics etc.), and each will consider different aspects of an ethical conundrum (a different line of reasoning) to determine what is right or wrong and defend the determination. Sometimes different lines of reasoning will come to the same conclusion about right and wrong, but sometimes they will not. Often the document reads as if society is conceived in mechanistic or functionalist ways. This 'macro' conception of society is based on an assumption that there is a set of shared values, norms and identities, each of which play a role in making sure the 'social body' functions correctly. Functionalist conceptions of the social world are outdated and discredited, having been replaced in social and philosophical thought with more nuanced 'conflict theories' of how values and norms are continually contested. Thus, identifying norms and values for ethical decision-making is a fraught process.

The document switches between normative ethics (e.g. Executive summary (p.8), 'embedding values' and 'norms' into A/IS) and an acknowledgment that the ethical traditions/logics of different cultures (this can refer to national cultures, trans-national cultural groups, or subcultures/minority groups within a society etc) may be different from those of the perceived mainstream or majority e.g. in the section Systems across cultures (p.164-167). This later issuer is broader than the influence of affective computing, which has the potential to erase to erase cultural difference (an act of immense actual and symbolic violence). This lack of clarity on what the IEEE mean by ethics, ethical reasoning and the tradition that meaning derives from creates confusion in the document and will lead to a lack of clarity around design for A/IS. It appears that IEEE favours an applied ethics in the human rights

tradition which is evident in arguably the most universal model of ethics, the medical ethics principles framework, as indicated in some of the language used in the report and ideas that it alludes to (e.g. 'values', 'norms', 'freedom', justice, beneficence, 'autonomy', respect' etc.). If this is the case then the historical roots of this ethics tradition and its principles need to be clearly espoused in the document (see here for a brief summary of the main principles (see http://web.stanford.edu/class/siw198q/websites/reprotech/New%20Ways%20of%20Making%20Babies/EthicVoc.htm).

The document correctly acknowledges that cultural and minority groups may adhere to their own ethical traditions. For example, many First Nations people have specific customs that rely on consultative/collectivist decision-making practices and traditional law that guides the sharing of knowledge. There is an implicit (Western) assumption in the document that all knowledge (including beliefs, values) can be discovered, known and publicly shared by/with all people, and that this can be initially or continually 'input-ed' into A/IS for ethical machine learning and decision-making. For many cultures, this assumption does not hold, as this example from Indigenous culture from the Australian Human Rights Commission demonstrates:

'The rights to Indigenous traditional knowledge are generally owned collectively by the Indigenous community (or language group, or tribal group), as distinct from the individual. It may be a section of the community or, in certain circumstances, a particular person sanctioned by the community that is able to speak for or make decisions in relation to a particular instance of traditional knowledge. It is more often unwritten and handed down orally from generation to generation, and it is transmitted and preserved in that way. Some of the knowledge is of a highly sacred and secret nature and therefore extremely sensitive and culturally significant and not readily publicly available, even to members of the particular group.' (https://www.humanrights.gov.au/sites/default/files/content/social_justice/nt_report/ntreport 08/pdf/chap7.pdf)

An initial or continual "discovery' and 'input' model for deep learning A/IS is not appropriate in cultures in which knowledge is collectively owned, regulated by enduring custom and tradition, and in many cases, sacred and secret to the community or to particular sub-groups within the community. Such cases add a layer of complexity to using principles from even widely used human rights/medical

ethics frameworks (for example see, https://aiatsis.gov.au/research/ethical-research/guidelines-ethical-research-australian-indigenous-studies or http://www.pre.ethics.gc.ca/eng/policy-politique/initiatives/tcps2-eptc2/chapter9-chapitre9/).

2. Now is the time to produce developmentally appropriate, curriculum-aligned strategies and materials for educating and empowering children and young people about A/IS.

In addition to the ethical education of software engineers, computer scientist and other technologists (p.144), attention must be given to the education of children and young people on A/IS inclusive of but beyond privacy issues. IEEE should consider, as a matter of urgency, working with education specialists to develop developmentally appropriate curriculum material for digital literacy which can assist teachers in educating children and young people about: what A/IS is; its features and functions; the way A/IS is woven into the everyday interactions with technology including the IoT; ethical issues and A/IS; the potential of A/IS for good; and (for older children) legal and regulatory frameworks to ensure informed interaction with A/IS including privacy issues. To educate children, is to begin to educate the community as children bring their knowledge of A/IS into homes, real and virtual places of recreation and communication, and a range of public forums.

3. More serious consideration of the consequences of unethical conduct is required: The case for professional registration

Professions that have a significant impact on people (teaching, medicine, psychology, social work, law) have long included a substantial component of ethical training in their degrees. Furthermore, these professions generally operate on a registration basis which is regulated both by nation legal frameworks and professions themselves. Given the serious and substantial impact of the work of technologists on today's society — especially computer scientists and software engineers who develop and unleash new technology — it is timely to consider reorganising the profession at both national and international levels so that there is a powerful means to educate on and regulate the ethical behaviour of its members. Unethical behaviour by teachers, doctors or lawyers can result in them being 'struck off' of the professional association, and in cases where there are concomitant legal frameworks, from practicing the profession itself. It is time for serious

accountability. Technologists should be governed and made accountable through a professional registration process and legal frameworks: without this, there will be weak consequences for unethical conduct regarding A/IS and other technologies.

4. Stop developing autonomous weapons systems (AWS) as no good will come from them: This is the ethical position to take.

There is no greater crime than taking a human life. Humans should be held responsible for a death that they cause. This is a fundamental legal and ethical principle. This section of the document clearly outlines the case for NOT having AWS and provides a series of well explained, evidence-based Issues regarding the devastating effects that are likely with AWS (Issues 4-11, pp. 120-130). This includes the argument that humans may not be able to ultimately control AWS or understand the logic of the machine decision-making, thus logically making the machine responsible for human death and destruction. Therefore, to develop such machines and systems is a profound abdication of ethical responsibility on the part of technologists. It goes against the vision of this document and long-standing moral and legal principle of holding humans accountable for the death and destruction that they cause. The organisation of the section on AWS begins with an assumption that these systems should necessarily exist and be subject to national and international law as the primary mode of regulation (Issues 1-3 pp.115 - 120). However this assumption is convincingly challenged with Issues 4-11 outlining the case for not forging ahead with this technology (pp. 120-130). This section should begin with Issues 4-11 – after which there is no case for developing AWS. Furthermore, it is naive at best and disingenuous at worst to suggest that 'codes of conduct' and 'reflective practice' will be key practices in minimising harm from AWS. Codes of conduct and reflective practice are individualised responses to what is essentially a 'make or break' issue for your profession, which should hold a strong, evidence-based and highly ethical position on this issue: no good will come from AWS and so therefore they should not be allowed to be produced.

Ioannis C.MATSAS, (University of Cyprus, CY); PhD Candidature in the field of Ethics in Transport Innovations (Aristoteles University, GR)

Generic comments

1. All across the document the words "should", "would", "could", "need to", "recommend", "must", are used. The above wording, taking into consideration the "Standardization rules" leads to permissible or forbitten practices. e.g. page 71, 1st "Candidate Recommendation", page 87, "Candidate Recommendation" and page 92, "Candidate Recommendation". Is this the intention of the authors? If not, then a uniform approach should be followed all over the document, taking into consideration the Disclaimer of the Document.
2. In the chapters there is a reference to "Background" or to "Background Analysis". Is this an intentional differentiation or a calibration between the WGs wording is needed?
3. The terms: "person, identity, user, affective system, community, individual, organization, interested part, being and moral patient", should be added to the eadv2 -Glossary.
4. The approach of Committees in EADv1 and its chaptering seem to be different from the one of new committees for EADv2.To be more precise, I would propose to reorganize the structure of the document and split it in 3 parts, together with relevant restructure of the chapters' content. An approach could be: Part 1- Stating the principles, including General Principles, Embedding values in A/IS, Methodologies to guide ethical research and design, Safety and Beneficence of AGI and ASI, Law and Policy. Part 2- Applications, including Personal Data and individual access control, Reframing Autonomous weapons systems and Mixed reality in ICT. Part 3: Results and metrics, including Economic /humanitarian issues, Classical Ethics in A/IS, Affective computing and Well-being.
5. As the content of the chapters prepared by the New Committees, and to be more precise their proposed Candidate Recommendations, do not offer, at present, significant added value, as the content of the work of the Committees of EADv1 does, except "Policy" matters (probably because of the new scientific areas they deal with), I would propose to incorporate the content of the work prepared by the new committees in the existing chapters of EADv1, until further development and deepening is made.

Specific Comments

Paragraph	Comment	Proposal
§2, Definition of "stakeholder"	Further to the "universities, organizations, governments, and corporations making these technologies a reality for society", consumers, intermediate and end users could be considered as stakeholders.	To add the consumers, intermediate and end users in the definition of stakeholders. Further to that, I propose the definition to be included in the attached eadv2 -Glossary. In addition a table with the roles of each stake-holder/role players in the A/IS arena, would be very useful.
§4, Reference to "human norms"	A definition of the term would be useful	To add the definition of the term "human norm" in the eadv2 Glossary.
§1, 2, "Background Analysis" & "Candidate recommendation"	Should the eadv2 make a reference to the opposite relation? This might exist in the near future.	To add the expectation that A/IS might have of humans.
"Background Analysis", 2nd phrase	"*People will have some unique expectations for humans*". Although clear in anthropology and sociology, the difference between people and humans is missing.	To add the definition of the term "people" and "humans", as well as their differences.
"Candidate Recommendation"	The independent, internationally coordinated body although assesses criteria when deployed and their evolution after deployment and interaction with other products, does not assess/supervise the operation of A/IS when in service.	To add: "…..other products, **as well as during servicing**."
"Background",§3 & 4	Review boards should have the capacity and authority to assess A/IS projects in all steps of their lifetime.	To add the task of assessing A/IS projects, against existing relevant standards and accepted practices.

Paragraph	Comment	Proposal
"Background Analysis"	The development of very capable A/IS, further to the complete transformation of the global political landscape, would also transform the global social landscape.	To add: "…..but the global political **and social** landscape."
"Section 1-Digital personas"	There should be a clear definition of the term "persona"	To add the term and definition of "persona/digital personas" in the eadv2-Glossary
"Candidate recommendation"	It is not clear if the Service provider will be chosen by the user or, in the opposite case, will the user be informed of who the service provider really is?	Need for clarification
"Candidate recommendation", 3rd bullet	What will be result of deleting data? Is there gonna be a rupture in the information chain and data of an individual?	Need for an alternative proposal.
"Candidate Recommendation", last bullet	"Moving all computational values to the periphery (on the person) seems to be the only way to combat all the risks articulated." This view drives to a unique solution.	If the intention of the author is the proposal of a single solution further justification should be presented.
"Candidate Recommendations", 3rd bullet	"Artificial intelligence ethics certification for responsible institutions".	The term "responsible" should be further explained.
"Background"	The sense that emerges from the way standards are connected to A/IS, is that standards are rather "technical" than multy sciences standards.	Depending on the author's intention, it is proposed, either to connect those standards to computer science and technical aspects fields or to refer to standards that cover all spectrum of sciences, both applied and theoretic.

Paragraph	Comment	Proposal
"Background"	The word "deception" refers to an illegal activity rather than to the action of a person/ agent to protect a human or an entity, in general.	To replace the term "deception" with the wording "divergence from the commonly accepted or standard procedure/action".
"Background", §2	A reference is made to "ethical and moral decisions".	In order to have a clear view of the author's intentions there should be a reference in the eadv2 Glossary, of the difference between "ethical" versus "moral".

Members of the High School Committee Respond to *Ethically Aligned Design* v2

In December of 2017, The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems ("The IEEE Global Initiative") collaborated with AI4ALL, a nonprofit training the next generation of diverse leaders in AI, to launch a committee for young people to share their perspective on the challenges and promise of emerging A/IS.

The initial participants of the committee are primarily made up of AI4ALL alumni-high school students who have participated in AI4ALL's summer AI education programs located in university AI labs around North America-who are interested in helping to shepherd AI development towards a positive and inclusive future. In addition to around 20 high school students, participants include outreach committee members from The IEEE Global Initiative and AI4ALL representatives.

Response below were created by members of the High School Committee (young women under the age of 18) as general feedback to *Ethically Aligned Design*, version 2. While the comments are general in nature (versus providing feedback to specific portions of EADv2), the logic is that these comments will provide key perspectives to Committees of The IEEE Global Initiative as they rework their sections leading to the next iteration of *Ethically Aligned Design*.

Click here to see a version of these comments printed in a paper created by AI4ALL.

Click here to read more information about the paper and the High School Committee.

Note: Names of students contributing their thoughts have been removed from this document to honor IEEE's policies regarding minors. Please click on the links above to learn about the contributors.

****

When thinking about AI in 10 years, what are you most excited about?

a. When it comes to AI, I am most excited about not just the further development of AI, but the **increasing accessibility of the benefits of AI**. For example, I saw this when working on with AI to map poverty in Uganda in order to help resource distribution efforts in impoverished areas during [my time at Stanford AI4ALL, AI4ALL's summer camp for high schoolers at Stanford University]. AI is on track to develop into a technology that can help the poor and rich, if used based on fair standards. Seeing such a forward-thinking concept become our reality regardless of who we are is incredibly exciting. Benevolent applications of AI, such as social welfare and medical aid, have already started to develop and will continue developing as the trends point towards an increasing amount of resources and focus shifting towards AI. AI has so many applications that will ultimately have the capacity to help virtually everyone, billions of people, so I'm excited to see it become more accessible and thus more impactful.

b. I believe that **AI provides opportunities and makes basic privileges more accessible** to those who do not have them. For example, AI can be used to provide medical support to those who cannot afford doctor appointments, and computer vision can be used to read text out loud to the blind. Other than the obvious humanitarian benefits, this will also increase productivity and the overall wellness of society. I am hopeful that AI education becomes more easily accessible and understandable, and that by experiencing AI in their everyday lives, people will become more curious and invested in the world around them. Perhaps most importantly, I am excited that AI may have a positive benefit for the environment. It is undeniable that the environment is getting lower and lower on the world's list of priorities. With AI, perhaps we can save the world before the environment becomes completely unbearable.

c. I am most excited about **AI's potential to preserve and create knowledge**. For example, the natural language technique of knowledge base population and query answering can fill in gaps in incomplete information databases and generate new knowledge. This technology might be applied to history and archaeology to find meanings of ancient script and languages, and it can also be used to preserve indigenous cultures. I am excited to see how AI can be applied to not just make humans more productive but also lead us to new intellectual depths that we would not be able to reach alone.

d. I am most excited about **AI's potential to help doctors and researchers to make more accessible medicine and healthcare**. Especially in rural and developing areas, the lack of doctors creates the urgent need for more accessible screenings. Since AI can process images and analyze large amounts of data quickly, it is an optimal tool for disease detection. With time, more training data becomes available, and AI diagnosis could become more "experienced", allowing for accurate automation of disease detection. I also look forward to see the impacts of AI in other fields in the future!

e. When it comes to artificial intelligence, there are so many problems to be solved, but I am most excited about **solving problems that would help patients in the medical industry**. Surgeries could be done in a faster and possibly more efficient way that would not require as much human involvement. Also, AI systems have the ability to diagnose a patient and improving these abilities will prove to positively impact patients by giving them an idea on what they should expect. While thinking about AI and medicine, it will be exciting to see how the creation of similar systems can be used to make the lives of people much better.

f. I am most excited about the potential of **AI to overcome global barriers by improving the accessibility of personalized healthcare and education around the world**. I am intrigued by the impact of personalized healthcare in developing countries, where access to efficient, accurate disease prediction and detection can be inadequate. AI also promises to transform education into a universal platform with individual learning plans that reward critical thinking in students of diverse backgrounds, languages, and skill levels. Through these endeavors, AI can help unite the world and advance the quality of life globally.

g. I think **AI has promising applications in omics**. For instance, AI algorithms could analyze genomic data and pinpoint disease-causing genes or mutations. The human metabolome is also vastly complex, and AI could help to extrapolate and elucidate cause-effect relationships between metabolomic biomarkers and certain health conditions.

h. I am hoping to see some **AI incorporated in the classroom**. Applications of AI in education such as Intelligent Tutoring Systems can greatly benefit both teachers and students in a multitude of ways. For example, learning will become

much more personalized in order to address each student's specific needs, and teachers will receive better feedback on their students' academic performances. In addition, students who are introduced to AI at an earlier age are more likely to be increasingly aware of its implications in the future.

i. I am excited to see how **AI will impact and aid in scientific discovery**. AI enables machines to sort, process, and analyze data hundreds of times faster and more accurately than a normal human being can. As a result, it can help scientists, especially in fields such as biology and astronomy in which there are millions of data points to be analyzed. As we understand more about our universe, we will also gain insight on how to improve technology to solve problems around us. Furthermore, advancements in branches of biology, namely biomedicine and genetics, will allow doctors to understand and treat diseases more efficiently.

j. I am most excited to see **AI help families in their everyday life**. Developments that make home appliances smart, personal bots, and self-driving cars are advancements that will help make people's lifestyles more efficient by adjusting to provide individualized care and attention.

k. **I'm excited to learn about philosophical and psychological behaviors of robots** in addition to human interactions with AI.

l. I am so excited to see (and hopefully work on) how **AI will be used for aerospace technologies**. We can use it for search and rescue, more intelligent military aircraft, cyber security, and more. I'm extremely excited about its applications to space travel; I imagine it will be significant for getting to Mars.

m. I am excited about the **overall change AI will bring to the world**. With the development of AI, people's daily lives are going to change, and how the world functions is going to change. I cannot predict how things will change, but I am sure that the change will be there and that it will be drastic. I am really excited to see the gradual development in AI and see as it grows and eventually "takes over the world," just not in the way movies show AI take over the world.

n. I am incredibly excited about **AI's potential in numerous fields**. In justice, AI that is programmed properly could vastly improve prison systems across

America and the rest of the world, as well as streamline court procedures in a way that is truly, unequivocally fair. I also think that AI has amazing possibilities in healthcare, as already described by various people, and believe that AI could also globally interconnect people across language barriers and cultures.

o. I am most excited about **the potential that AI has to transform the assistive technologies industry**. AI systems may soon be able to assist people with a range of disabilities - including physical and mental limitations in the elderly, learning disabilities in children, etc. Millions of families and individuals across the world cannot afford full time or even part time care, and so assistive technologies integrated with AI systems (natural language processing systems, perhaps) could be an incredible service to facilitate their lives.

When thinking about AI in 10 years, what are you most worried about?

a. I am most worried about the increasing prevalence of **artificial intelligence bias** within the training and implementation of algorithms, something which can subsequently inhibit our ability to provide a fair implementation of AI to everyone whether it be socially, medically, financially etc. Seeing things such as racism, sexism, or other human-held social perceptions appear in our algorithms has widespread ramifications that will reduce our ability to create effective/meaningful opportunities and impacts for a diverse range of people looking to use AI. The beauty of developing AI lies in the core concept that the intelligence we are creating is able to achieve certain capabilities at a much greater capacity than humans. Perpetuating biases and stereotypes constructed by human society undermines the value and potential for AI to create a meaningful change in the world. AI is something that the media has portrayed often at times as malicious and out of control, thus transparency in algorithmic development and fair standards and practices of AI are critical to overcoming some of our greatest worries.

b. While working on self-driving cars and other AI used for convenience purposes is important, I am mostly worried that the development of AI and the chance to improve technology will drive attention away from the most important issues. For example, a lot of workers are using computer vision or natural language processing to develop robots that can complete everyday tasks for humans. This

offers a great opportunity to develop these aspects of AI. However, **AI can be applied to more pressing global issues**, such as world starvation or the refugee crisis. As with many products, AI devices and technology development are targeted only towards those who can afford it, as there is little motivation to develop AI for the poor, oppressed, or disabled. Additionally, the low level of transparency within company's use of data is unacceptable. I read a recent article claiming that apps can track and use your location and microphone (using natural language processing) for gathering information for advertisements. Often, any disclaimers regarding a company's use of personal data is written in small font within other large blocks of text where it is unlikely for users to read. Disclaimers discussing the use of personal data should be put separately from long blocks of text and should stand out. Lack of awareness decreases the general public's trust in AI. Beyond this, I worry that AI may become a competitive field, where the sharing of innovation is discouraged among scientists.

c. I am most concerned with the potential implications of the **misuse of AI**. For example, natural language processing technologies are becoming more and more ubiquitous, and we will probably eventually become reliant on them for a variety of tasks (including searching the web, storing information, and getting important information). However, these technologies are not infallible, and these risks are all the more concerning when the technology is used on a massive scale for high-stakes situations, such as matching people to resources during natural disasters and communicating during dangerous military operations. If a system fails to deliver resources, delivers the wrong resource or message, or makes some other kind of potentially fatal error, who is to be held accountable, and what preventative and curative measures can be taken to mitigate these risks?

d. From Siri to self-driving cars, AI is increasingly prevalent in today's world, creating the need of eliminating **bias in intelligent systems**. However, my biggest worry is that they will not be eliminated. These biases resurface more frequently as data is collected on a larger scale. Some machine learning algorithms utilize the data retrieved from the internet to train and test algorithms; these algorithms may become biased if the data used shows an inclination towards one group of a particular cultural, political, or racial background. If a large amount of opinionated data is used to train the

algorithms, machine learning algorithms based on that data will also inevitably grow biased. Since AI is already so widespread in the modern world - and will continue to become more integrated into mundane life, bias in AI algorithms can potentially have a larger impact, creating preconceptions either in favor of or against certain populations.

e. My biggest worry about artificial intelligence is about the **job losses** that will affect our country. This issue has been debated about since the increase of developments in AI and people wonder what the implications of this technology are. Slowly, small jobs will not be useful and more people will be put out of their jobs because of the "take over" by AI systems. Self-driving cars could take over car services and jobs in hospitals like surgical procedures and diagnosing a patient with a disease or disorder. Though jobs will be created to create these AI systems, people with certain areas of expertise might not have a good job, which is problematic for our world overall. Though AI is supposed to be used for the good of all, people's lives could be negatively affected by having one system that could take the jobs of five.

f. A **one-dimensional approach to AI** worries me most. Such an approach can take a variety of forms, including the advancement of financial interests at the expense of environmental needs, misuse of AI to prioritize individual gain over humanitarian concerns, and algorithmic bias that causes an intelligent system to favor one social group over another, propagating division in society. In order to maximize AI's benefits to society, we must embrace a multidimensional approach—equally considering AI's impacts on numerous, diverse sectors. For this reason, I think that the IEEE Global Initiative's principle of "Prioritizing Well-being" as a holistic metric for an AI system's success is critical.

g. I am concerned about **AI transparency**. We witness self-driving cars on highways and train machine learning algorithms that could outperform doctors at certain tasks, yet these technologies arguably largely remain in their experimental stages. One reason is that AI is like a black box—we see the input and output but can't really explain how the machine arrives at a decision. The fatal Tesla Autopilot crash in 2016 demonstrates this point that even if a computer arrives at the right answer most of the time, a small proportion of errors is enough to make people wary of machine learning. In order to prevent

the same mistakes from occurring in the future and to more effectively create safe AI, we need explainable AI.

h. I am worried about **the way AI is expressed through the media**.

i. I am most worried about the **lack of awareness regarding artificial intelligence**. Arguably, AI can be viewed as a double-edged sword that has both favorable and unfavorable ramifications, with a heavier emphasis on the latter. But when people start to think that artificial intelligence is dangerous, this could potentially impede progress in the field. It is important to spread the message that AI is not about robots overtaking jobs, but rather its purpose is to assist and benefit humanity. Therefore, I strongly advocate for IEEE's Policies for Education and Awareness, because when it comes down to things like AI, ignorance is not bliss.

j. As AI and technology advance, it is necessary for increasing quantities of data to be collected and stored for research and experimental purposes, as data is essential to machine learning. An issue that arises from this relates to **privacy and access in regards to personal data**. As human beings, we expect certain aspects of our life to be kept private. AI, however, increases the potential to infringe on this basic human right. Today, data is already collected from users without their knowledge or consent, which is the case when it comes to pop-up ads and Facebook. Through analyzing a user's internet search history, companies are able to find products and services better suited for their customers. Though this application of personal data is rather harmless, it is likely that we will use personal data for more in the future. As a result, I think it is important for us to set a universal definition of what personal information is, as well as discuss the extent to which it can be collected and used by AIs.

k. I am most worried about **AI being too expensive to be implemented widely across cultures**, and how AI will be modified to fit in with the social dynamic of various countries. It is also worrying to think of misuse with developing AI and the future of information security.

l. I'm worried about AI not receiving credit for its actions and not being granted **rights to protection** when operating in public areas. Society is tentative about a heavy reliance on AI, yet there are times when robots are faced with the

challenge of dealing with bullying and abuse when they're fulfilling their duties. Moreover, the act of imposing harm is unethical and affects human virtue.

m. One of my biggest worries is **uncovering the black box in AI** when things go wrong, especially morally. If an AI causes harm we'll need to figure out the underlying cause. Consider the training of an "all-intelligent" AI; the intelligence was largely trained with human input, and actions that should have had negative consequences could've gone without them. Additionally, there are many dilemmas to be considered in things such as drone warfare. When looking for issues in artificial intelligence, it's not just as simple as debugging another program.

n. I think that it is important to **emphasize the reality of artificial intelligence to the general public**. Many media outlets have portrayed AI as an omnipresent, all-powerful, all-knowing technology that possesses true intelligence and has emotions on or near the level of humans. A good portion of the public seems to have a perception that in the very near future, there is a real possibility of "robots taking over all of our jobs, and robots taking over the world". This is simply not the case, and so I think that as a community, scientists in the general fields of AI and computer science need to inform the public that contrary to popular opinion, the field known as "artificial intelligence" is simply comprised of methods that are computer algorithms that process large amounts of data. Yes, Machine and Deep Learning methods have tremendous potential, but they are still only algorithms that people have coded--AI is nowhere close to annihilating our society! Additionally, I am worried about the **stereotypes and biases that AI might perpetuate** as it becomes more and more integrated into the fabric of our society. Programmers of these algorithms and AI systems are indirectly impacting the lives of millions of people, as their products decide matters such as who gets job interviews, who gets parole, the diagnoses of patients, whether a bank should loan money to a given person, etc. Humans are often too trusting of mathematical models and computer algorithms because they hope that these systems will remove human bias, but in fact, algorithmic bias is rampant, and often hidden. A more evolving issue as AI systems begin to more heavily use Deep Learning methods is that many AI systems are "black boxes" - the public, and even the programmer, doesn't know how the algorithm spots the patterns it does, and so even if one detects algorithmic bias, how can it be fixed?

o. I'm worried about the **general portrayal and misunderstanding of artificial intelligence**. The media show artificial intelligence as a threat to the world potentially as dangerous as a zombie apocalypse would be. The larger problem, though, is that people tend to believe this, likely because it makes a good story. They also may unconsciously take in ideas from the media and use this to formulate biases against AI. The general portrayal of AI and the bias against AI seriously hinders the development of AI. This could easily be solved with more AI education for the media and explaining how things aren't likely to turn out the way they are depicted on the big screen.

p. I am most concerned about the **potential of artificial intelligence to institutionalize biases throughout society** if programed with such biases. China will launch an AI-generated Social Credit System in 2020 that based on several factors gives every citizen a score of trustworthiness; to be ranked publicly, these scores, among other things, will determine eligibility for jobs, loans or even travel visas. That means AI could, if biased, permanently exclude entire demographics and socioeconomic groups from vital access to financial services, dramatically amplifying already-harmful trends in the status quo. Such exclusion has already been a reality; earlier this year, Amazon's AI-generated maps of areas eligible for same-day delivery discriminated against the areas redlined by the Home Owners' Loan Corporation in the 1930s. (The 1930s redlining of these communities, which include South Side in Chicago and the Bronx in New York, caused their initial fall into decline that has lasted well into modern-day America.) If developed with our biases, AI can permanently undercut values of social equality and rupture our understanding of our government as bound to a social contract like Locke's, ethically responsible for its interactions with its citizens. The result could be social polarization between different groups of society, and a disintegration of the concept of common well-being, which in turn can damage the foundations of humanism and the principles upon which democracies are built.