Creating the Human Standard for Ethical Autonomous and Intelligent Systems (A/IS)

John C. Havens (IEEE; john.havens.us@ieee.org)

DOI: 10.1145/3203247.3203255

Introduction

Like most people, I first encountered Artificial Intelligence through movies - The Terminator, Blade Runner, 2001. As a rule, the future in these stories was always dystopian which I found irritating. If humanity was able to create such amazing technology, wouldn't they also have created ethical codes or standards to keep things from going awry? Granted, a film about a code of ethics isn't as sexy as killer robots, but picturing utopian futures powered or assisted by AI seemed like something not enough people were doing in my estimation.

I've been writing about technology since around 2011 for publications like The Guardian, Slate, and HuffPo. But it was an ongoing series on Artificial Intelligence I wrote for Mashable that led me to my current work with IEEE. In 2014 I wrote an article called "Coming to Terms With Humanity?s Inevitable Union With Machines[1]" as a way to genuinely confront fears I was facing about the nature not of killer robots but algorithms that might make choices for me to the point where I'd lose myself. Not being an engineer or programmer by training, I realize now this was an uninformed perspective, but it's one I believe the general public often shares when not fully understanding how AI functions under the hood. The article led to my writing my book, "Heartificial Intelligence: Embracing Our Humanity to Maximize Machines[2]", which I spoke about at the SXSW conference in Austin as a guest of IEEE.

When I spoke I had already done a great deal of research to identify any existing Codes of Ethics for AI (this was in 2014). Everyone I interviewed kept quoting Asimov's Laws of Robotics which as a newbie to AI I found to be quite alarming. Didn't people realize these came from his short story, "Runaround" from 1942[3]? While I appreciated the nature of the story was to demonstrate the conundrum of trying to have one simple set of laws apply to any robot / system (eg, "do no harm" doesn't make sense if you're creating a surgical robot) I also didn't understand why nobody had created a more formal and updated set of Principles.

Fortunately during my talk at SXSW there were two people from IEEE staff leadership in the audience who agreed with my assessment that there was a need to identify a set of global principles for AI. They recommended I present my ideas to IEEE on how they could create these Principles which I did a few months after SXSW.

The Council and The Chairs

When I presented my ideas in 2015 for members of IEEE's Management Council, it was the first time I met Konstantinos Karachalios. He's the Managing Director for the IEEE Standards Association and also sits on IEEE's Management Council. Konstantinos is the person who helped shape my initial ideas into what has now become The IEEE Global Initiative that also inspired the P7000 Standards Working Group series.

The Chair is Raja Chatila. After the initial core structure for The Initiative was in place, Konstantinos approached Raja who was at that time completing his tenure as President of the IEEEE Robotics and Automation Society to talk to him about The Initiative. Thankfully Raja was interested and began further shaping and developing the structure and makeup of The Initiative.

Our Vice-Chair is Kay Firth-Butterfield. I first

[1] https://mashable.com/2014/04/11/digital-humanity/

[2] https://www.amazon.com/Heartificial-Intelligence-Embracing-Humanity-Maximize/dp/0399171711

[3] https://en.wikipedia.org/wiki/Three_Laws_of_Robotics

met Kay while she was at Lucid AI[4], serving as Chief Officer of their Ethics Advisory Panel. Beyond being a barrister by trade and a gifted speaker, Kay is one of the most connected and respected people in the AI Ethics world. She's now serving as the Head of Artificial Intelligence and Machine Learning at the World Economic Forum.

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (A/IS) was launched in April of 2016 to move beyond the paranoia and the uncritical admiration regarding autonomous and intelligent technologies and to illustrate that aligning technology development and use with ethical values will help advance innovation while diminishing fear in the process.

The goal of The IEEE Global Initiative is to incorporate ethical aspects of human well-being that may not automatically be considered in the current design and manufacture of A/IS technologies and to reframe the notion of success so human progress can include the intentional prioritization of individual, community, and societal ethical values.

The IEEE Global Initiative has two primary outputs. First, the creation and iteration of a body of work known as “Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems[5]”. Second, the identification and recommendation of ideas for Standards Projects focused on prioritizing ethical considerations in A/IS.

Version 1 of Ethically Aligned Design (EAD) was released in December of 2016 as a Creative Commons document so any organization could utilize it as an immediate and pragmatic resource. Launched as a Request for Input (RFI) to solicit response from the public in a globally consensus-building manner, the document received over two hundred pages of feedback at the time of the RFI's deadline.

Ethically Aligned Design, Version 2 features five new sections in addition to updated iterations of the original eight sections of EADv1. The IEEE Global Initiative has now increased from 100 AI/Ethics experts to more than 250 individuals including new members from China, Japan, South Korea, India, and Brazil and EADv2 now contains over 120 key Issues and Candidate Recommendations. Version 2 was also launched as a Request for Input. (You can download Ethically Aligned Design, Version 2 at this link: http://standards.ieee.org/develop/indconn/ec/auto_sys_form.html)

The Mission of The IEEE Global Initiative is to ensure every stakeholder involved in the design and development 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. By identifying A/IS oriented Principles and creating Standards directly relating to the challenges brought about by the widespread use of A/IS, The Initiative hopes to complement and evolve how engineers create technology in the algorithmic age.

The IEEE P7000^{TM} series of Approved Standardization Projects

Along with creating and evolving Ethically Aligned Design, members of The IEEE Global Initiative are encouraged to recommend Standards Projects to IEEE based on their work. Below are titles and descriptions for each of these approved IEEE Standards Projects, and more information is available via the links included:

The IEEE P7000^{TM} series of standards projects under development represent a unique addition to the collection of over 1300 global IEEE standards and projects. Whereas more traditional standards have a focus on technology interoperability, safety and trade facilitation, the P7000 series address specific issues at the intersection of technological and ethical considerations. Like their technical standards counterparts, the P7000 series empower innovation across borders and enable societal benefit.

There are currently thirteen approved Standards in the Series, incorporating key issues

[4] https://mashable.com/2015/10/03/ethics-artificial-intelligence/

[5] http://standards.ieee.org/news/2016/ethically_aligned_design.html

within the Autonomous/Intelligent and ethical realm including transparency, data access and control, algorithmic bias, robotic nudging, wellbeing, and more:

- IEEE P7000^{TM} - Model Process for Addressing Ethical Concerns During System Design
- IEEE P7001^{TM} - Transparency of Autonomous Systems
- IEEE P7002^{TM} - Data Privacy Process
- IEEE P7003^{TM} - Algorithmic Bias Considerations
- IEEE P7004^{TM} - Standard on Child and Student Data Governance
- IEEE P7005^{TM} - Standard on Employer Data Governance
- IEEE P7006^{TM} - Standard on Personal Data AI Agent Working Group
- IEEE P7007^{TM} - Ontological Standard for Ethically driven Robotics and Automation Systems
- IEEE P7008^{TM} - Standard for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems
- IEEE P7009^{TM} - Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems
- IEEE P7010^{TM} - Wellbeing Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems
- IEEE P7011^{TM} - Standard for the Process of Identifying and Rating the Trustworthiness of News Sources
- IEEE P7012^{TM} - Standard for Machine Readable Personal Privacy Terms

For further information, please see: https://ethicsinaction.ieee.org

The Future

We are deeply fortunate to have a fantastic Executive Committee made up of representatives from UNESCO, The Partnership on AI, private sector industry and many more. Sven Koenig of SIGAI (ACM's Special Interest Group on Artificial Intelligence) is also member of our Executive Committee. Along with benefitting from Sven's deep expertise in AI, it has been fantastic to see ACM's efforts in the AI space, including their groundbreaking "Statement on Algorithmic Transparency and Accountability[6]".

It is with these thought leaders that we?ve only recently completed our plans for how ww'll complete the final version of Ethically Aligned Design. When looking at the document, you'll note that each of the thirteen committees has listed "Issues" and "Candidate Recommendations." Initially we were using the term, "concerns" instead of "issues" but Francesca Rossi (who's on our Executive Committee) made the excellent point that we didn't want an entire paper comprised of only "concerns." (We made that change before publishing version 1 of Ethically Aligned Design).

The idea of "Candidate" Recommendations (if memory serves) came from Richard Mallah of FLI. Rather than have EADv1 make it seem like we had finalized our thoughts on any particular subject, this process let us release EADv1 and EADv2 in a public Request For Input process. We received over two hundred pages of feedback for EADv1 (which you can see here: http://standards.ieee.org/develop/indconn/ec/rfi_responses_document.pdf) and are currently getting feedback for version 2. This is a unique process for IEEE which mirrors their consensus-building processes found in their Standards creation and other processes. For us, we wanted to make sure to not infer that a group of largely Western A/IS experts could define ethics in one fell swoop.

A lot of the feedback we received for Version 1 was people outside of the US and the EU letting us know it was important to include non-Western ethical ideas in future versions. We agreed, and we ended up inviting the people providing feebback along with a number of other global thought leaders from China, Japan, South Korea, Brazil, India, Mexico, Thailand and Africa to our work. A number of people from those countries even translated the introduction of EADv1 into their own languages (which you can see here: http://standards.ieee.org/develop/indconn/ec/ead_v1.html).

[6] https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf

For our final push as we move forward, we'll be working to add more members from regions and societal constituencies we may have missed or from which we need more representatives. We are excited to have recently added a High School Committee made up of about twenty students from the great organization AI4ALL, plus we?ll be working with the IEEE Young Professionals to increase diversity of age and orientation as well as regionalization.

All Members from this point on, along with updating content for Ethically Aligned Design, will be asked to vote at various points to finalize our "Candidate Recommendations." We'll also be refining our list of General Principles to use as the criteria to help Committees decide what "Issues" align with those Principles so our final document will be a cohesive whole united by our overall philosophy of "Advancing Technology for Humanity" (that's IEEE's tagline) and Prioritizing Human Wellbeing with Autonomous and Intelligent Systems (the subtitle of Ethically Aligned Design).

Our goal is to publish the final version of EAD around Q2 of 2019. We'll also be releasing a number of white papers focusing on Committee content over the next few months, along with videos from members and a few big surprises planned for when we launch the final version.

So, stay tuned, and consider joining our ranks as a Member of The Initiative or in one of the IEEE P7000 Working Groups. We would greatly welcome any ACM Members to help us shape the future of A/IS ethics principles and standards.

John C. Havens is the Executive Director of The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html To get involved or learn more, please email John.