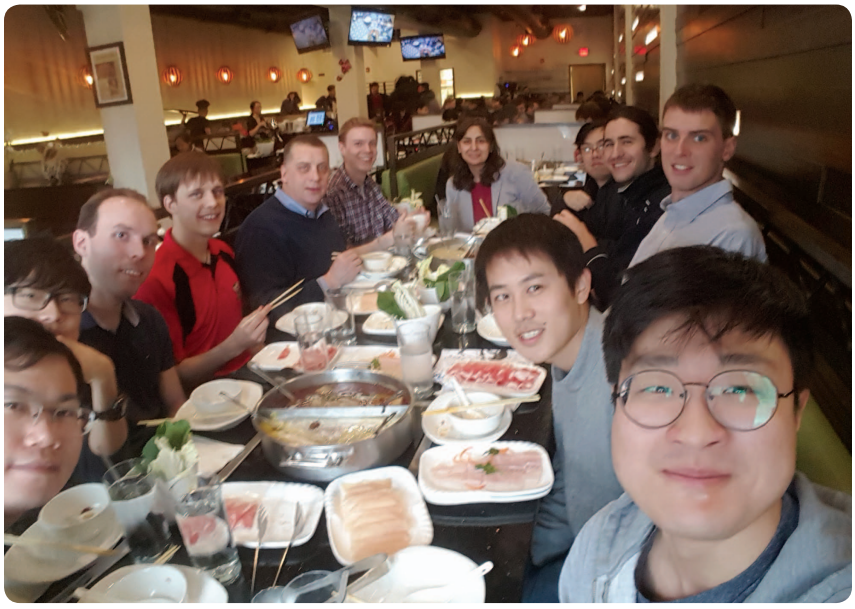
Ethically Aligned Design

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems recently released a report, "Ethically Aligned Design" (EAD) [1], [2] that provides some interesting insights into the recent discussions about the ethics of artificial intelligence (AI) [3], [4].

Reference [1] provides an updated version of the original 2016 report [5]. The primary purpose is to solicit feedback that can be incorporated into the final report that should appear in 2019. While the deadline for comments predates the publication of this column, the longer-term goals of inspiring the creation of standards (called IEEE P7000) and facilitating the emergence of national policies that align with the overall principles should still be a strong motivation to get involved in the effort.

A core principle of the committee is that the full benefits of autonomous and intelligent systems (A/IS) technologies will only be attained if they are aligned with "our defined values and ethical principles" [1]. The authors make the point of expanding the discussion beyond the ethics of AI to include robotics, machine learning, and intelligent systems engineering. The report raises several issues that are important to the IEEE Control Systems Society (CSS) community, so I strongly recommend that you read it over. In the process, I think you will find that it provides a good motivation for research that addressed key technical challenges in the field.

Digital Object Identifier 10.1109/MCS.2018.2810458
Date of publication: 18 May 2018



Members of Jonathan How's research group enjoying a hotpot Christmas lunch (from left): Macheng Shen, Yulun Tian, Jesko Mueller, Brandon Draper, Jonathan How, Björn Lütjens, Golnaz Habibi, Trong Nghia Hoang, Gabriel Bousquet, Michael Everett, Shenghao Jiang, and Dong Ki Kim.

Reference [1] (and the much shorter [2]) outlines several objectives of the effort, such as maintaining personal data rights, developing legal frameworks for accountability—how to allocate liability if A/IS do harm, and for systems that learn and "self-improve," ensuring that they are still able to create audit trails that support the decision-making process if further assessment is required. The report also highlights five general principles:

- ensuring that A/IS do not infringe on human rights
- developing metrics to help prioritize human well-being
- developing methods to assure that designers and operators are responsible and accountable
- creating methods to ensure that A/IS are transparent
- avoiding misuse of A/IS.

Transparency is an important concept for validation and certification of A/IS as it provides access to the underlying algorithms and data analysis techniques. As the developers of many of the control, AI, and machine learning tools being developed for systems such as self-driving cars, the CSS community could take more of a leadership role in this area.

Several recommendations are provided for each of these principles in the executive summary and then expanded on in the later sections of the report. For example, section 4 of the discussion of "Law" is on "Transparency,

Accountability, and Verifiability in A/IS" and provides ten detailed candidate recommendations [1, p. 159]. Of course, classical ethics are a key part of these recommendations, but I was also interested to see a discussion of social and moral norms in the section "Embedding Values into Autonomous Systems." While it is often hard to even write down what is meant by social norms, the report makes the point that these norms will depend on the community in which the system will be deployed and that they might change over time. My group has recently identified these as key issues when developing collision-avoidance algorithms with pedestrians—what is considered as a socially acceptable way to pass a pedestrian will depend on where the A/IS agent is operating, including which country and, often, which city as well.

Perhaps not surprisingly, there is a significant discussion about the design of autonomous systems as weapons [1, pp. 113–130]. The report discusses ten issues and provides candidate recommendations for each. The overall recommendations are that these systems be created so that designers of the technologies understand the implications of their work and the autonomous functions are predictable to the human operators. This includes ensuring that adaptive and learning systems can explain their reasoning and decisions to the operators in a transparent and understandable way. As before, the core technical challenges of predictability and transparency for A/IS represent key ways that members of the CSS field could get involved.

Given the large breadth of the effort, the committee (apparently composed of "several hundred participants") is subdivided into numerous working groups [6]: Transparency of Autonomous Systems (IEEE P7001), Algorithmic Bias Considerations (IEEE P7003), Standard for Ethically Driven Nudging for Robotic, Intelligent, and Automation Systems (IEEE P7008), and Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems (IEEE P7009).

As these working group titles suggest, these groups are considering topics that are well aligned with the research pursued by researchers within the CSS. The biographies indicate that these efforts are being led by members from the IEEE Vehicular Technology, Robotics and Automation, and the Computer Societies. However, there appears to be limited leadership and involvement from the CSS community, which is a bit surprising.

Other societies have started to develop conferences/workshops to discuss these issues, such as the Association for the Advancement of Artificial Intelligence (AAAI)/Association for Computing Machinery (ACM) Conference on AI, Ethics, and Society in February 2018 [7] and the Future of Life Institute in Boston, Massashusetts, that recently hosted a workshop, BENEFICIAL AI 2017 [8]. Similar workshops could also prove to be beneficial for the CSS community.

EAD is an important aspect of the future of many aspects of control and intelligent systems, so I recommend that you read the reports, think about how to apply the recommendations to your work, get involved in discussions of these working groups, and get involved in the efforts to create the future standards and policies.

REFERENCES

[1] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2017). *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems, version 2*. Accessed on: Feb. 6, 2018. [Online]. Available: http://standards.ieee.org/develop/indconn/ec/ead_v2.pdf

[2] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2017). *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems, Overview: Version 2*. Accessed on: Feb. 6, 2018. [Online]. Available: http://standards.ieee.org/develop/indconn/ec/ead_brochure_v2.pdf

[3] N. Scharping. (July 18, 2017). Artificial intelligence experts respond to Elon Musk's dire warning for U.S. governors. *Discover Mag.* [Online]. Available: http://blogs.discovermagazine.com/d-brief/2017/07/18/artificial-intelligence-elon-musk/

[4] Open AI. (2018). *Wikipedia*. [Online]. Available: https://en.wikipedia.org/wiki/OpenAI

[5] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2016). *Ethically aligned design: A vision for prioritizing well-being with artificial intelligence and autonomous systems, version 1*. [Online]. Available: http://standards.ieee.org/develop/indconn/ec/ead_v1.pdf

[6] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2018, Feb). IEEE Standards. [Online]. Available: at http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html

[7] AAAI/ACM Conference on AI, Ethics, and Society. [Online]. Available: http://www.aies-conference.com/steering-committee/

[8] Beneficial AI. (2017). [Online]. Available: https://futureoflife.org/bai-2017/

Jonathan P. How

Early Control Systems

Since their earliest period steam boilers have provided opportunities of application for a variety of feedback devices. The oldest of these is the safety valve first described in 1681 by Denis Papin as a pressure regulator for pressure cookers but used by him as a steam boiler safety valve as early as 1707. It soon became an indispensable accessory of steam engines. Despite its simplicity — it consists of a weight-loaded valve in the boiler wall which is forced open by excessive steam pressure until the surplus steam is released — it represents a closed loop with negative feedback.

—Otto Mayr, *Feedback Mechanisms in the Historical Collections of the National Museum of History and Technology*, Smithsonian Institution Press, Washington, D.C., 1971, p. 70.