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

Becoming a Leader in Global Ethics

Creating a Collaborative, Inclusive Path for Establishing Ethical Principles for Artificial Intelligence and Autonomous Systems

By Sara R. Mattingly-Jordan

Organizations aspiring to such a grand mission as creating global standards for artificial intelligence and autonomous systems must accept that standards development is a provisional, deliberative, exercise that must be infused with the spirit of generosity, curiosity, critique, and collaboration.

Between January and May 2017, leadership of The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems welcomed input from around the globe on their founding document – *Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems (AI/AS)* (hereafter referred to as *EAD* or *EADv1)*. Broken into multiple sections, EADv1 features over eighty Issues and Candidate Recommendations designed to provide academics, technologists and policy makers with specific, directional principles to help prioritize ethical considerations at the forefront of AI/AS design. The document also serves as impetus for Committees of The IEEE Global Initiative to submit Standards ideas to the IEEE Standards Association based on their efforts. In this way *Ethically Aligned Design* is helping to both frame key global principles and build pragmatic soft governance that can deeply and positively influence the AI/AS landscape.

Instead of approaching the process of iterating EAD through a lens of technocratic expertise alone, The IEEE Global Initiative offered their ambitious work for global, multidisciplinary review. Infused with a spirit of provisionalism and democracy, The IEEE Global Initiative threw open all elements of the document – from principle selection to grammatical inflection – to conscientious revisions by any and all interested parties.

Within this five-month period, the authors of *Ethically Aligned Design* received comments from over 35 individuals or organizations, from 10 nations, who volunteered their time and knowledge to advance the cause of improving the *Ethically Aligned Design* version 1 document. These individuals offered comments and resources to build a more well-rounded and global vision for The IEEE Global Initiative. Whether through imparting their disciplinary, national, or ethical perspectives to the committee, these commentators submitted over 65,000 words of feedback, including providing over ninety references for further examination by the committees of The IEEE Global Initiative.

As a summary and response to the feedback could not do each individual's contribution full justice, you may read all comments here.

As we discuss below, based upon feedback from the contributors, subsequent versions of *EAD* will work to embrace global ethics through more developed attention to polylingual, collaborative, interdisciplinary, and non-Western principled approaches.

Polylingual Inclusion

In which natural language will an artificial intelligence think? Should a global organization like IEEE draft principles leading to standards in the idiom in which artificial intelligence will think?

Producing global document invariably requires that the drafting committee make a choice of language for the production of the draft. *EAD* reviewers from European nations brought to our attention the limits of relying on English-language resources for discussions of key ethical principles. They advocated, instead, a clearer explanation of the resources used, particularly those in non-English journals, and requested that attention to other languages be given in the construction of a polylingual glossary of key terms.

The basis for recommendations of a polylingual perspective for the building blocks of subsequent versions of *EAD* is social constructivist in nature. The respondents made their case in two ways. The first is that inclusion of different languages into the intellectual background of the drafting discussions will invite inclusion of alternative constructions of ethical norms and social reality. Within the comparative humanities and social sciences disciplines, it is convention that full comprehension of cultures requires comprehension of its symbolic system, including its language. The gracious reviewers who pointed out that resources in other languages should be incorporated into the intellectual background of the *EAD* project gently prodded the drafting committee to consider these conventions in their endeavors. (Others were more direct to point out that additional expertise from humanities and social sciences experts should be incorporated for this reason and others.)

The second way in which this case was made is that inclusion of polylingual resources into developmental discussions will, by proxy, include the communities speaking various languages into a setting discussing the setting of potential standards. Inclusion of communities not obviously represented in the list of existing

Committees is a point taken up next.

Collaborative Inclusion

Who will craft the ethical principles and make recommendations for standards that are imparted into artificial intelligence systems? While some reviewers pointed out a diffuse concern for representativeness of the drafting committees, reviewers from East Asia brought to the table two perspectives that broaden the base of participation in for The IEEE Global Initiative: 1) organizational collaboration and 2) integrative pedagogy.

IEEE and other organizations with interests in the AI/AS arenas are globally connected through chapters and affiliation networks of related professionals. As pointed out chiefly by our Asian collaborators, many members of global chapters are eager to impart technical and ethical knowledge to the *EAD* authors. However, an existing barrier to inclusion is translation of key documents into Asian languages, specifically Chinese and Japanese. Through the collaborative partnerships with IEEE chapters, many members are offering to organize translation projects and to coordinate dissemination of those efforts. (To date, the Asian members of The IEEE Global Initiative have translated the Executive Summary of EADv1 into Chinese, Japanese and Korean, and a Portuguese version is also being discussed).

Many reviewers enthusiastically endorsed the use of the *EAD* document as a tool for engagement of students in the development of student's beliefs about ethics in AI/AS. Reviewers from the Asia-Pacific region endorsed the use of the *EAD* documents, whether the full version or overview versions in their undergraduate and graduate courses. Some particularly engaged reviewers enumerated the links between extant resources for teaching topics identified within the *EAD* document. Adoption of *EAD* into an interdisciplinary classroom was also a feature of Latin American feedback.

Interdisciplinary Inclusion

Perceptions of the design of artificial intelligence as being a matter reserved for only experts in the computer sciences, electrical engineering disciplines, and ethicists of science and technology were directly challenged by our reviewers. Reviewers from Latin America, specifically Mexico and Brazil, expanded the disciplinary horizons of The IEEE Global Initiative by bringing important variations on these disciplines to the table. For example, eight undergraduate term papers, penned by students from a mechanical engineering background taking a course in bioethics, were submitted by students from Mexico.

Emphasizing the importance of student participation and inclusion of EAD as a basis for fruitful classroom discussions and assignments, these students offered the *EAD* working groups new avenues of thinking about AI from the perspective of religion and human-health centered value systems.

Scholars from India offered that related disciplines, such as the study of cybernetics, should be included in the group of perspectives evaluated for principles and best practices for governance of ethical AI/AS. Pointing to a wealth of scholarship available in journals housed in India, these reviewers were also careful to point out that disciplinary language and publication conventions may be a barrier to effective incorporation of *EAD* into professional and teaching venues. To increase the level of public knowledge of The IEEE Global Initiative and of *Ethically Aligned Design*, these reviewers suggested a more aggressive approach to translation and dissemination in multidisciplinary venues.

Inclusion of Non-Western Principles

Assuming an AI could be designed to be ethical, which ethical principles would take priority? Ethical artificial intelligence as envisioned by the authors of *Ethically Aligned Design* will be adaptive, supportive of human endeavors, and inclusive of the many ways of being that make up the mosaic of humankind. To that end, many of the reviewers of this document added helpful insights for elaborating and expanding upon the various ethical systems and principles that could be included in the next draft of *Ethically Aligned Design*. While some of the recommendations offered were diffuse – to better include the Global South, for instance – the intersection of the various recommendations provided clues for a start to the expansion of the ethical systems. In particular, respondents recommended that East Asian (e.g., Confucian) and South Asian (e.g., Vedic) values be more actively considered in the elaboration of value systems in the next iteration from The IEEE Global Initiative.

In the remainder of this document, a number of principles from non-western systems are brought out and offered as potentially compatible principles for exploration in the subsequent iterations from The IEEE Global Initiative and *Ethically Aligned Design.* In brief those value systems include Classical Chinese values, particularly those of Confucians, Mohists, and Taoists, and faiths that have origins or strong relationships to the Vedic tradition, such as Jainism and Buddhism.[1]

[1] *No claim is made that this is a comprehensive, scholarly, evaluation of comparative ethical principles for governance of artificial*

Chinese Traditions

Confucian Principles
Classical Confucian values permeate the cultures of contemporary China, Hong Kong, Japan, South Korea, and Taiwan. Mixed with other unique ethno-cultural traditions and texts, the *Analects* and *Great Learning* outline principles and practices for a life well-lived, including life lived with new and emerging technologies and traditions. While the vastness of the traditions can scarcely be summarized effectively here, five values are emphasized for their role as augmenting and enlivening principles and practices selected for discussion in *EAD v1.*

Ren/ Jen/ Benevolence

Crafting guidance and recommending standards to ensure the benevolence of artificial intelligence and autonomous systems is a cardinal goal for the leaders and participants in The IEEE Global Initiative. Fostering a culture with benevolence (expressed as *ren* or *jen)* at its center was also the goal of the Classical Confucian authors. Benevolence is held to be the central virtue for governments to uphold when guiding people towards a good life.

The value of *ren* can be taken up by The IEEE Global Initiative in two ways: as a goal to strive for when recommending standards and as a principle towards which to guide the designers of artificial intelligence as they create new technologies that will inevitably affect human lives as much, if not more, than ordinary government.

Li/ Rites and Social Order

Multiple observers to the documents pointed out that it is imperative that the values of distinct communities be preserved in the context of creating AI/AS. The importance of community rites for maintaining social cohesion and social order for the Confucian system of values cannot be overstated. Li or the rites held a paramount place of importance in all levels of social order, from family to government.

As the authors of *Ethically Aligned Design* grapple with the exciting challenge of guiding professionals in the artificial intelligence space towards creating community-sensitive, ethical AI/AS, the value of *li* can guide their efforts in two ways. First, this community, to include the Committee chairs, Committee members, and

intelligence systems. The author of this document is happy to share such scholarly evaluations with interested observers. Please contact Dr. Sara Jordan at srjordan at vt dot edu.

thoughtful public commentators on the first version, are participating in a set of contemporary rites of deliberation. Protecting and respecting the values that guide this community practice is one expression of the value of *li*.

Second, carefully cultivating knowledge of the rites of various communities before launching new products or programs into their milieus is an imperative of cosmopolitan respect and of the ancient value of *li*.

Yi/ Righteousness/ Propriety

The purpose of the endeavor to craft principles and recommend standards for ethical design of artificial intelligence is not to add layers of bureaucracy or rules, as at least one respondent fears may emerge from this process. Instead, the purpose of the endeavor is to identify the principles, practices, and questions that might guide groups designing AI/AS toward creating ethical machines and programs. The purpose of the document is to instill a sense of propriety rather than profit, righteous behavior rather than rapacious behavior into the fast-paced and lucrative world of AI. Where the principle of yi is instilled into the designers of AI, it is the hope that the creation of those designers would also act from a principle of benevolence and respect.

Tao/ Way/ Virtuous path

As evinced by the subtitle to the *Ethically Aligned Design* document – “A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems” – the highest end to which artificial intelligence designers should aim is human wellbeing. A number of respondents to the invitation for public discussion worried over the definition of wellbeing captured in the document. Definitions of wellbeing under consideration include happiness and eudaimonia.

An alternative, Confucian value of Tao (not to be confused with the Taoist interpretation of the same principle), can be compared to virtuous wellbeing. What the Confucian principle of Tao brings to the group of relevant principles for AI/AS is that it is not only a philosophical principle or mandate for virtue, it is a virtue that unites various practices. As noted by one reviewer, the practice of the various principles by AI/ AS designers *and* by their creations is the goal of the efforts of The IEEE Global Initiative in this venue.

Tsang meng/ Rectification of names

The definition of various principles, prioritized lists of ethical and technical terms, and priority arrangement of the topics were matters of clarification requested by respondents to the public call for responses. Clarity in the meaning of terms, virtues, and establishment of a rank of principles was a tenet of Classical Confucian authors called the rectification of names.

While establishment of a hardened fast set of names and responsibilities may not be part of the accomplishments possible for the next round of drafting for EAD, the iterative process of coming closer to a clear definition of what it means to design an ethical AI is a goal of this public feedback process.

Mohist Principles
Chien Ai/ Universal love
Casual observers of Chinese ethical philosophy may not be familiar with the fine-grained distinctions between various schools of thought. One school that is not well known is Mohism, whose major exponent espoused a principle similar to ethical utilitarianism. Mo Tzu advocated for the virtue of universal love between individuals as the route away from acquisitiveness and mutually destructive tendencies that might be found in a deeply competitive environment.

The virtue of universal love is one that could guide subsequent development of the *EAD* in two ways. First, as elaborated upon in the *EAD* already, the goal is establishment of principles that could foster development of AI/AS inline with benefits for all humanity, not solely those in well-resourced states or organizations. Consistent reminders of universal love or commitment to a fuller, principled, utilitarianism could ensure that the ethical commitment does not stray into mere economic consequentialism. Second, the virtue of universal love can stand in as a call to action for the Committee chairs to continue their efforts to seek and incorporate broader public concerns through public deliberation in the next rounds.

Taoist Principles
Wu Wei/ Non doing/ No unnecessary action
Public and scientific discussions of the potential of AI were highlighted by many of the reviewers for this document. Fear about AI run amok or about the eclipse of human power and potential through the AI singularity were reviewed as reminders of the public fears about AI/AS. Some respondents advocated a strong precautionary approach – to allow only AI/ AS with extensive, risk-based, testing onto the market. Others offered a weaker precautionary approach – to carefully monitor those AI/AS that do come to market.

As it stands, the perspective of The IEEE Global Initiative is that the AI/AS revolution is coming and that coordinating principles and recommending standards to prevent ethical malfeasance is more likely to be successful than is attempting to forestall the emergence of an AI/AS driven society. The Taoist principle of non-doing or avoiding unnecessary action is a principle that reinforces a quiescent approach to governance.

Instead of recommending obstruction, restriction, or even encouragement, the quiescent approach of wu wei recommends that leaders and philosophers work within the context of what is emergent. A quietist approach does not prevent individuals from working against evil-doing, but rather encourages to reflect on the source and to identify correctives that match the source of the problem, not its effects. In the case of The IEEE Global Initiative, working within the principle of wu wei recommends continuing the exercise to develop collaborative principles, education, and empowerment first before standards and regulations.

Vedic Traditions

Public comments to the first public version of *Ethically Aligned Design* mentioned two great resources of non-Western values: Confucian and Vedic traditions. The voluminous Vedic tradition, spanning thousands of years and schools of philosophical thought, provides an untold wealth of possible augmenting and alternative principles for the authors of the subsequent versions of *EAD* to explore.

Jainism

Ahimsa/ Harmlessness/ Nonviolence

A significant and often cited public concern about AI/AS is that these technologies may evolve to exist beyond the control of their creators. Citing fictional tales of "evil robots", there is public concern that AI will become violent towards humanity. While a goal of establishing the principles and guidelines contained in *EAD* is to prevent intentional creation of maleficent AI, one way to articulate and expand upon this goal is to embrace the first of the 5 sacred vows of Jains-- *ahimsa*, non-violence.

The teachings of Mahavira argue that the highest goal is that of utter harmlessness and this spirit of non-harm should permeate throughout the lives of adherents. Harmlessness becomes an all encompassing principle that guides all actions for committed Jains, from dress to diet.

The value of ahimsa holds promise as a principle that expands upon commitments already elaborated upon in the *EAD* document, such as ensuring the safety and beneficence of AI/AS. Following the doctrine of ahimsa, AI systems should be as close to harmless to their designers, users, and those whose lives are indirectly affected by AI/AS, as possible. Augmenting the oft-cited Asimov principles, ahimsa may also serve as a powerful reminder for the ultimate purpose of creating autonomous weapons-- the eclipse of warfare-- by reminding those who design such systems to minimize harm to the least possible level.

Buddhism
Of the traditions mentioned as elaborating upon a set of principles that could expand the western-centric traditions often cited in the *EAD* documents, Buddhism was the most specifically cited. While schools of Buddhist thought are many and nuanced, the respondents to the public call cited Buddhist values as important to consider in the next round of revisions. A full account of all of the potential virtues in the Buddhist traditions being well beyond the scope of this brief section, three values are elevated as potential companions to principles already within the EAD set.

Pratītyasamutpāda/ Dependent Arising
How can AI "care"? And how should humans care for AI/AS? While conventional expectations are that those individuals and groups which design AI have created something which is then separate from them and their wellbeing, some respondents to the *EADv1* suggested that a more mutually dependent relationship is at play. That is, AI is created by and creates its designers. As discussed in the popular and technical press, it is possible that there will be a time when AI creates other AI. This eventuality has caused considerable public concern about liability and responsibility for AI.

The Buddhist idea of dependent arising is that there is nothing which exists that does not have a relationship to all else that exists. No thing exists in isolation and nothing is the root of cause and responsibility. The idea of a reciprocal or mutual causation is not foreign to the western-centered ethical system, but the Buddhist argument for dependent arising captures both a theory of causation, epistemology, and a theory of being. In relationship to AI, the dependence of the created on the creator and creators on their creations is epistemic and causal. The individuals and groups who design AI/AS are dependent upon dense networks of emerging ideas, codes, and capacities, which when used, change the intellect of the creators and users.

With this principle in mind, the idea of a separation of responsibilities – that that an individual or group is responsible for the acts of an AI – is a misspecification of the relationship. AI and the designers are collaboratively or dependently responsible for AI. Adoption of the doctrine of dependent arising also stipulates that the authors of the *EAD* documents are also creating and created by AI, the document itself, and the process of creating *EAD*.

Viraaga/ Non-attachment/ Non-desire

Reviewers of *EADv1* offered criticism concerning absence of a strongly asserted commitment to making AI/AS open source material to benefit humanity free of profit-making. While the *EAD* document and its comments are freely available to the connected public, the charge of making AI/AS free is not one taken up in this version. As the possibilities of an open source AI environment are more clearly explored, the Buddhist value of non-attachment could form a principled basis for such an environment. In Buddhist doctrine, suffering is the result of attachment, whether to things, status, or even to self. Non-attachment is the organizing principle for a life lived in such a way as to avoid causing suffering, whether to self or others. Non-attachment does not connote, in all cases, asceticism. Instead, in this arena non-attachment means minimizing profit, pride, or prestige to only those levels necessary to sustain one’s activities and improvements towards an ideal, ethical, goal.

āryāṣṭāṅgamārga/ 8 fold path

Respondents who recommend a more global, western and non-western, approach to the ethical principles enshrined in AI have pointed out that it is not clear that the principles and practices captured in the *EAD* document lead towards wellbeing, eudaimonia, or equal levels of human happiness around the globe. The *Ethically Aligned Design* document is a statement of a full program of practices and guidelines necessary to foster ethical AI development. Standing alone, no one program or potential standard will ensure that ethical AI/AS are developed. However, in conjunction with other standards, educational materials and opportunities, and fruitful public discussion, it is possible that we will come closer to the goal of benevolent, wellbeing promoting, AI and AS.

A unifying principle for the organization of virtuous speech, practices and attitudes is the Buddhist Noble Eightfold Path. Intended to lead followers towards liberation of self and others (boddhisatva way), the elements of the eightfold path are: Right Understanding, Right Thought, Right Speech, Right Action, Right Livelihood, Right Effort, Right Mindfulness and Right Concentration. Similar to the mission of The IEEE Global Initiative – to ensure every technologist is educated, trained and empowered to prioritize ethical considerations in the design and development of autonomous and intelligent systems – each of these elements come together in an iterative way to form the path to liberation.

Conclusion

The many thoughtful reviewers of the first version of *Ethically Aligned Design* expanded the horizon of principles, resources, conversations, and languages to be included in subsequent drafts. This iterative process of development will undoubtedly be made better by their contributions, which are available <HERE>. With particular respect to inclusion of non-Western ethical principles, much work is left to be done. However, as is described above, the distance between principles expressed and those which will meaningfully augment and compliment them is not so great that it cannot be overcome through this collaborative, interdisciplinary exercise in inclusion.

Sara R. Mattingly-Jordan
Assistant Professor
Center for Public Administration & Policy
Virginia Tech