IEEE Standard Review – Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems

Kyarash Shahriari, *PEng, PhD, Senior MIEEE*
Senior Control Systems Engineer
Aversan Inc., ON, Canada
Email: kyarash.shahriari@ieee.org

Mana Shahriari, *Student MIEEE*
Computer Vision and Systems Laboratory
Department of Electrical and Computer Engineering
Laval University, QC, Canada
Email: mana.shahriari.1@ulaval.ca

***Abstract*—In September 2009, the IEEE Board of Directors approved the new IEEE tagline – Advancing Technology for Humanity – as recommended by the IEEE Public Visibility Committee. Aligned with the IEEE tagline, IEEE Standards Association takes the initiative to address ethics in engineering design under "Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems" focusing on Artificial Intelligence and Autonomous Systems (AI/AS). The intention is to cover, as much as possible, the ethical concerns on AI/AS through a rigorous regard to the problem from different perspectives. The ultimate objective of this ongoing initiative is to provide guidelines/procedures/standards to prioritize human wellbeing in the forthcoming evolutions on artificial intelligence and autonomous systems. This article reviews different aspects addressed in Version 1 of this initiative.**

Keywords: Ethically Aligned Design, Artificial Intelligence, Autonomous Systems

I. Introduction

Artificial Intelligence (AI) and Autonomous Systems (AS) are ever-increasingly getting integrated into our daily lives through recent technical/technological innovations. The major difference between the integration of AI/AS into our lives with previous technological advancements is the embedded capability of these systems to make decision independently without human assistance. This new dimension could potentially interfere with ethical and moral values already addressed by human intelligence and responsibility. Delegating the decision-making capability to AI/AS requires fundamental considerations in design process which is largely disregarded and ignored in the current mainstream engineering design practices. Recently, IEEE Standards Association takes the initiative to address ethics in engineering design under "Ethically Aligned Design: A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems", shortly called the IEEE Global Initiative in this article. The purpose of this initiative as is "to ensure every technologist is educated, trained, and empowered to prioritize ethical considerations in the design and development of autonomous and intelligent systems" [1]. In other words, the initiative is (1) to raise public awareness and start open discussions on the impact that AI/AS systems can have on human wellbeing in both a positive and negative way; (2) how the AI/AS design can be aligned to moral values and ethical principles; and (3) to encourage engineers to prioritize ethical considerations in their design and to promote moral values and ethical principles in creation of AI/AS. Engineers, as the designers, companies as promoters, and authorities as regulators must be aware that they all should go beyond technical issues and provide concrete assurance to the public that new developed AI/AS systems are not harmful to the public in any ways and are for the human wellbeing. Aristotle elucidated human wellbeing as "the highest virtue for a society. Translated roughly as "flourishing", the benefits of eudemonia begin by conscious contemplation, where ethical considerations help us define how we wish to live. By aligning the creation of AI/AS with the values of its users and society we can prioritize the increase of human wellbeing as our metric for progress in the algorithmic age" [1].

II. Problem Statement

Artificial Intelligence (AI) itself is a young field which was officially introduced by Alan Turing in 1940s, an English computer scientist, mathematician, logician, cryptanalyst and theoretical biologist. Though the field is relatively new, it has inherited ideas, viewpoints and techniques from other some very old disciplines such as philosophy, theories of reasoning, learning, physics, mathematics and some relatively new disciplines such as logic probability, decision making, computation, psychology, linguistics, and computer science [19]. Due to AI's nature, inheritance from various disciplines, different people think of AI differently; "Are you concerned with thinking or behavior? Do you want to model humans or work from an ideal standard" [18]. The field of AI attempt to understand intelligent entities, as well as striving to constructing intelligent entities that are able to perceive, understand, manipulate and predict the world; thus unlike philosophy and psychology it is not only concerned about understanding intelligence. In concise, AI can be defined as the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence and to exhibit it by machines [20]. Machine Intelligence can therefore

978-1-5090-6264-5/17/$31.00 ©2017 IEEE

be seen as any kind of cognitive function e.g. reasoning, learning, and problem solving, perception, rationality, thought process, and natural language processing that a machine mimics human intelligence. Although no one can predict the future of AI in details, it is clear that " computers with human-level intelligence (or better) would have a huge impact on our everyday lives and on the future course of civilization" [18].

With the recent advancements in AI, and specifically in machine learning, the growth in design and development of Autonomous Systems (AS) especially in uncertain, rapidly-changing, and potentially adversarial environments is becoming more feasible. One examples of recent innovation in AI/AS is autonomous-driving cars.

As stated in the introduction, technical challenges are not the only problems in developing AI/AS systems. A quick Google search on autonomous-driving cars dilemma results in ambiguous, contradictory, and very concerning outcomes from ethical and moral perspective. Only few are listed here:

- The self-Driving Dilemma: Should Your Car Kill you to Save Others [14]?
- Why Self-Driving Cars Must Be Programmed to Kill [15]?
- The Robot Car of Tomorrow May Just be Programmed to Hit You [16].
- Mercedes-Benz's Self-Driving Cars Would Choose Passenger Lives Over Bystanders [17].

This simple question demonstrates just a very small portion of ethical/moral challenges we are facing by adopting AI/AS in our lifestyle. The concerns cannot definitely be answered by companies' internal design guidelines or a piece of legislation from local or federal governments. The start point would therefore be to clearly state the problem, to understand its scope, and to define its dimensions before heading toward any proposal and solutions.

III. The Mission of the IEEE Global Initiative

To prioritize ethical considerations in design, manufacturing, and development of AI/AS, we need to make sure that "anyone involved in the research, design, manufacture or messaging around AI/AS including universities, organizations, and corporations making these technologies a reality for society are well educated" [1].

As stated before, this paper is a review on Version 1 of [1], where it represents the collective input of over one hundred global thought leaders in the fields of Artificial Intelligence, law and ethics, philosophy, and policy form academia, science, and the government and corporate sectors. Version 1 of [1] seeks mainly two goals: first, to provide insights and recommendation from these peers and that provide a key reference for the work of AI/AS technologists in the coming years; second, to provide recommendations for IEEE Standards based on Ethically Aligned Design IEEE P7000 (TM) – Model Process for Addressing Ethical Concerns During System Design [2] was the first IEEE Standard Project (approved and in development) inspired by The Initiative.

Two further Standards Projects, IEEE P7001(TM) – Transparency of Autonomous Systems [3] and IEEE P7002(TM) – Data Privacy Process [4], have been approved, demonstrating The Initiative's pragmatic influence on issues of AI/AS ethics. In addition, other related standards in this regard are IEEE P7004(TM) – Standard for Child and Data Governance [5], IEEE P7005(TM) – Standard for Transparent Employer Data Governance [6], and IEEE P7006(TM) – Standard for Personal Data Artificial Intelligence (AI) Agent [7].

The Version 1 of [1] was prepared using an open, collaborative and consensus building approach, following the processes of the Industry Connections program, a program of the IEEE Standards Association. The Industrial Connection initiative promotes the collaborations among individuals and organizations regarding emerging technologies towards developing potential new standards.

IV. Structure and Content of the Initiative

Ethically Aligned Design Document Version 1 includes eight sections to cover ethical/moral related AI/AS. Each topic has its dedicated committee of the IEEE Global Initiative and has been largely discussed in the respective committee. Four new committees have recently been added to eight initial committees. Below is a brief summary of the committees and the issues covers in each committee.

A. General Principles

The General Principles Committee is looking forward to providing high-level guiding principles that apply to all types of artificial intelligence and autonomous systems. This committee has articulated three high-level ethical concerns: (1) how to embody the highest ideas of human rights; (2) how to prioritize the maximum benefits to humanity and the natural environment; (3) and to mitigate risks and negative impacts as AI/AS evolve as socio-technical systems. As stated in [1] the intention is that "Principles, Issues, and Candidate Recommendations will eventually serve to underpin and scaffold future norms and standards within a new framework of ethical governance for AI/AS design".

In this regard, the main concerns to tackle with are: human benefits, i.e. how to ensure that AI/AS do not infringe human rights; responsibility, i.e. how to assure that AI/AS are accountable to avoid any potential harm; transparency, i.e. how to ensure that AI/AS are transparent to make sure why the system made a particular decision; finally education and awareness. i.e. how to extend the benefits and minimize the risks of AI/AS technology being misused.

As an example and for the human right issue upon first Principle being Human Rights, all AI/AS must comply to Documents such as the Universal Declaration of Human Rights [8], the International Covenant for Civil and Political Rights [9], the Convention on the Rights of the Child [10], Convention on the Elimination of all forms of Discrimination against Women [11], Convention on the Rights of Persons with Disabilities [12], and the Geneva Convention [13]. The AI/AS must also be safe and secure throughout their lifetime and they

must be traceable to discover the root cause should they cause any harm.

B. Embedding Values into Autonomous Intelligence Systems

In order to develop successful Autonomous Intelligent Systems (AIS) that will benefit our society, it is essential for such systems to be designed to adopt to human norms and moral values of the community they serve. To help designers, the Embedding Values into AIS Committee has broken this objective into a three-pronged approach that are: to Identify the norms and values of a specific community affected by AIS; to implement the norms and values of that community within AIS; and to evaluate the alignment and compatibility of those norms and values between the humans and AIS within that community.

In this regard, we may face several challenges that are: values to be embedded are not universal, but rather largely specific to user communities and tasks; moral overload, i.e. AIS are usually subject to a multiplicity of norms and values that may conflict with each other and designers need to prioritize these constraints; AIS can have built-in data or algorithmic biases that disadvantage members of certain groups; it is not trivial to build values into a computational architecture once the relevant sets of norms (of AIS's specific role in a specific community) have been identified; norms embedded in AIS need to correspond closely to the given community; due to the lack of transparency and verifiability, achieving a correct level of trust between human and AIS is very much challenged; and finally third-party evaluation of AIS's value alignment.

C. Methodologies to Guide Ethical Research and Design

In order to serve humans and to enhance their wellbeing, system design methodologies need to put great emphasis on human rights and values. As stated in the Universal Declaration of Human Rights, human right is a primary form of human value and machines should be in human's service and not the other way around. Therefore, values-aligned design methodologies must become a part of the design process and all new developed AI/AS and final product must be aligned with ethical guidelines.

The main issues in this regard are: ethics is not part of degree programs due to the ambiguity inherent in ethical language which cannot be translated into the formal languages of mathematics and computer programming; we need models for interdisciplinary and intercultural education to account for the distinct issues of AI/AS and to bring engineers and designers in contact with ethics and social scientists; the need to differentiate culturally distinctive values embedded in AI design, i.e. how do different cultures view privacy; lack of value-based ethical culture and practices for the development and deployment of products for industry; lack of values-aware leadership on what human values should be respected in the design of a system; lack of empowerment to raise ethical concerns between engineers and design teams regarding design or design specifications; lack of ownership or responsibility from tech community where the broader set of social concerns raised by the public, legal and social science communities; need to include stakeholders and practitioners for best context of AI/AS due to their insights to incorporate; poor documentation on systems, related date flows, their performance and limitation and risks; inconsistent or lacking oversight for algorithms which results in the lack of transparency on how a certain algorithm or system came to its conclusion; lack of an independent review organization on how AI/AS are marketed and on their actual performance and application; use of black-box components which may not be fully understood.

D. Safety and Beneficence of Artificial General Intelligence (AGI) & Artificial Super-intelligence (ASI)

Future highly capable AI systems, i.e. artificial general intelligence (AGI), may have fundamental impact equal to agricultural or industrial revolution. On the other hands as AI systems capabilities increase, small defects in AI architecture, training, or implementation, as well as mistaken assumptions could result in unanticipated results. Thus, there is no guarantee that the AI/AS transformation will be a positive one without a safety mindset to develop systems which are safe by design.

Major issues in this regard are: as AI systems capabilities increase dramatically, measured by the ability to optimize more complex objective functions with greater autonomy across a wider variety of domains, unanticipated or unintended behavior becomes increasingly dangerous; retrofitting safety into future more generally capable AI systems may be difficult, and developing AI systems without certain concerns could result in systems with high levels of technical debt; researchers and developers will be challenged by a progressively more complex set of ethical and technical safety issues in the development and deployment of increasingly autonomous and capable AI systems; future AI systems may have capacity to impact the world by transforming not only the economy, but the global political landscape.

E. Personal Data and Individual Access Control

The artificial intelligence and autonomous systems have widespread access to our personal information, yet we remain isolated from gains we could obtain from our data. In other words, the insights derived from our lives is more of an asset to others than it is to us. To tackle the data asymmetry, people need a data environment where they can control their sense of self, define, access, and manage their personal data. The main goal is to envision the tools and evolved practices that will eliminate this data asymmetry.

The issues we are facing in this regard are mainly: how can an individual define and clarify his/her identity and organize personal date in the algorithmic era; what is the definition and scope of personally identifiable information; how to define control regarding personal data; how to redefine data access to honor the individual and their privacy rights; how to redefine consent regarding personal data so it honors the individual; how can data handlers ensure the consequences (positive or negative) of accessing and collecting data are explicit to an

individual in order for truly informed consent to be given; and could a person have a personalized AI or algorithmic guardian.

F. Reframing Autonomous Weapons Systems

Autonomous systems that are designed to cause physical harm, can and should have additional ethical ramifications and higher standard as compared to both traditional weapons and autonomous systems that aren't designed to cause damage.

There are several concerns in this area which can be summarized as: the lack of clear definitions and potential confusions regarding important concepts in artificial intelligence (AI), autonomous systems (AS), and autonomous weapons systems (AWS); AWS are by default amenable to covert and non-attributable use; there are multiple ways in which accountability for AWS's actions can be compromised; legitimizing AWS development sets precedents that are geopolitically dangerous in the medium-term; exclusion of human oversight from the battle-space can too easily lead to inadvertent violation of human rights and escalation of tensions; the variety of direct and indirect customers of AWS will lead to a complex and troubling landscape of proliferation and abuse; by default, the type of automation in AWS encourage rapid escalation of conflicts; there are no standards for design assurance verification of AWS; and understanding the ethical boundaries of work on AWS and semi-autonomous weapons systems can be confusing.

G. Economics / Humanitarian Issues

Technologies, methodologies, and systems that aim at reducing human intervention in our day-to-day lives are evolving dramatically. These factors have transform the lives of individuals in multiple ways. However, the public feels unprepared both personally and professionally at this rapid pace. It is therefore impotent to "identify the key drivers shaping the human-technology global ecosystem and address economical and humanitarian ramifications, and to suggest key opportunities for solutions that could be implemented by unlocking critical choke points of tension" [1].

Main concerns are: misinterpretation AI/AS in media tend to obnoxious oversimplifications which is also confusing to the public; the complexities of employment structure are being neglected regarding AI/AS and solely focuses on the number of jobs lost and gained; the current pace of technological change is too fast for excising methods of (re)training the workforce and the employment structure; it is important to regulate emerging technology effects on the society, however this does not necessarily means that any artificial intelligence policy slows down innovation; distribution of the benefits of AI/AS are not equally available worldwide; lack of access and understanding regarding personal information, privacy, and safety issues; and finally the advent of AI/AS can exacerbate the economic and power-structure differences between and within developed and developing nations.

H. Law

The design, development and distribution of AI/AS has given rise to many complex ethical issues. These ethical problems have almost always directly translated into concrete legal challenges or have given rise to collateral legal. The preliminary objective to make sure that the development of AI/AS fully complies with both the international and domestic laws. Though looks obvious, "this simple observation obscures the many deep challenges AI/AS pose to legal systems; global, national, and local-level regulatory capacities; and individual rights and freedoms" [1]. These concerns mainly fall into three areas that are government and liability, societal impact, and human in the loop.

Main concerns in this regard are: AI/AS have a degree of uncertainty, so we need to improve the accountability and verifiability in these systems; how can we ensure that AI is transparent and respects individual rights? For example, international, national, and local governments are using AI which impinges on the rights of their citizens who should be able to trust the government, and thus the AI, to protect their rights; how can AI systems be designed to guarantee legal accountability for harms caused by these systems; how can autonomous and intelligent systems be designed and deployed in a manner that respects the integrity of personal data?

V. New Committees

The recently added committees to the IEEE Global Initiative are:

A. Classical Ethics in Information & Communication Technologies

The focus of this Committee will be on examining classical ethics ideologies (utilitarianism, etc) in light of AI and autonomous technologies. Some of the issues that are being considered are: how classical ethics can help and act in AI/AS, how ethical considerations can be prioritized among tech companies, public projects and research consortiums; how responsibilities in AI/AS are defined in classical ethics; and whether or not value guided designed would cause problematic consequences such as unintended discrimination.

B. Mixed Reality Committee

This Committee will be devoted to the notion of identity and reality over the next generation, as "these technologies infiltrate more and more aspects of our lives, from work to education, from socializing to commerce" [1]. Mixed reality (MR) media combined with intelligent and autonomous systems have the potential to alter the reality we see, hear and experience. Thus, this Committee will work on discovering "the methodologies that could provide this future with an ethical skeleton and the assurance that the rights of the individual, including control over one's increasingly multifaceted identity, will be reflected in the encoding of this evolving environment" [1]. The main focus of this committee is: how to avoid the eradication of the positive effects of serendipity within the forthcoming AI/AS; how to prevent the mind to be fooled by MR; what long term effects would be on decision-making considering that the provided contents in MR are algorithm sensitive; and what the cultural effect would be on individual behaviours.

C. Affecting Computing

This Committee will investigate the impact of AI and autonomous systems on individuals and the society, as these systems are capable of "sensing, modelling, or exhibiting affective behaviour such as emotions, moods, attitudes, and personality can produce" [1]. The ethical concerns surrounding human attachment and the overall impact on the social fabric may be profound and "it is crucial that we understand the trajectories that affective autonomous systems may lead us on to best provide solutions that increase human well-being in line with innovation" [1]. The committee main concerns are: what the consequences of losing individual human autonomy are; how to evaluate system's honesty and truthfulness; how to address intimacy and relations with machine; Cross-Cultural concerns in terms of local values; and emotions within systems.

D. Effective Policymaking for Innovative Communities Involving Artificial Intelligence and Autonomous Systems (EpicAI)

This Committee will address: (1) exploring how effective policymaking employing autonomous and intelligent technologies can be done in a rapidly changing world; (2) generating recommendations on what initiatives the private and public sector should pursue to positively impact individuals and society, and (3) illuminating newer models of policymaking both extant and experiment to support the innovation of AI/AS for shared human benefit [1]. The focus of this committee is to assist public service entities such as governments, public/private partners, and members of the public to adopt more rapidly AI/AS and provide recommendations to explore innovative uses of AI/AS to enhance life quality.

VI. Conclusion

As mentioned previously, "The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems", shortly called "The IEEE Global Initiative" is to provide guidelines and recommendations to prioritize human wellbeing within Artificial Intelligence and Autonomous Systems usage. Under this initiative, "Ethically Aligned Design: A Vision for Prioritizing Wellbeing With Artificial Intelligence And Autonomous Systems" document is in development for which Version 1 has already been released and available for public consultation and discussion. Different aspects of this document were briefly reviewed in this paper. The Global Initiative is primarily made up of individuals from North America and Europe, thus the active representation of developing nations is required. The initiative is hoping to increase its impact factor among public and private sectors in different countries by increasing its collaboration with researchers and engineers from more diverse cultural backgrounds.

References

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

[2] IEEE P7000TM Model Process for Addressing Ethical Concerns During System Design, https://standards.ieee.org/develop/project/7000.html.

[3] IEEE P7001TM Transparency of Autonomous Systems, https://standards.ieee.org/develop/project/7001.html.

[4] IEEE P7002TM Data Privacy Process, https://standards.ieee.org/develop/project/7002.html.

[5] IEEE P7004TM Standard for Child and Data Governance, https://standards.ieee.org/develop/project/7004.html.

[6] IEEE P7005TM Standard for Transparent Employer Data Governance, https://standards.ieee.org/develop/project/7005.html.

[7] IEEE P7006 Standard for Personal Data Artificial Intelligence (AI) Agent, https://standards.ieee.org/develop/project/7006.html.

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

[9] International Covenant for Civil and Political Rights, http://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx.

[10] Convention on the Rights of the Child, http://www.ohchr.org/en/professionalinterest/pages/crc.aspx.

[11] Convention on the Elimination of all forms of Discrimination against Women, http://www.un.org/womenwatch/daw/cedaw.

[12] Convention on the Rights of Persons with Disabilities, http://www.un.org/disabilities/convention/conventionfull.shtml.

[13] Geneva Convention, http://www.cfr.org/human-rights/geneva-conventions/p8778.

[14] The self-Driving Dilemma: Should Your Car Kill you to Save Others? http://www.popularmechanics.com/cars/a21492/the-self-driving-dilemma.

[15] Why Self-Driving Cars Must Be Programmed to Kill? https://www.technologyreview.com/s/542626/why-self-driving-cars-must-be-programmed-to-kill.

[16] The Robot Car of Tomorrow May Just be Programmed to Hit You, https://www.wired.com/2014/05/the-robot-car-of-tomorrow-might-just-be-programmed-to-hit-you.

[17] Mercedes-Benzs Self-Driving Cars Would Choose Passenger Lives Over Bystanders, http://fortune.com/2016/10/15/mercedes-self-driving-car-ethics.

[18] Stuart J. Russell and Peter Norvig. 2003. Artificial Intelligence: A Modern Approach (2 ed.). Pearson Education.

[19] Stuart J. Russell and Peter Norvig. 2009. Artificial Intelligence: A Modern Approach (3 ed.) Upper Saddle River, New Jersey: Prentice Hall

[20] What is Artificial Intelligence? http://www-formal.stanford.edu/jmc/whatisai/whatisai.html