Classical Ethics in A/IS

The task of the Committee for Classical Ethics in Autonomous and Intelligent Systems is to apply classical ethics methodologies to considerations of algorithmic design in autonomous and intelligent systems (A/IS) where machine learning may or may not reflect ethical outcomes that mimic human decision-making. To meet this goal, the Committee has drawn from classical ethics theories as well as from the disciplines of machine ethics, information ethics, and technology ethics.

As direct human control over tools becomes, on one hand, further removed, but on the other hand, more influential than ever through the precise and deliberate design of algorithms in self-sustained digital systems, creators of autonomous systems must ask themselves how cultural and ethical presumptions bias artificially intelligent creations, and how these created systems will respond based on such design.

By drawing from over two thousand years' worth of classical ethics traditions, the Classical Ethics in Autonomous and Intelligent Systems Committee will explore established ethics systems, addressing both scientific and religious approaches, 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 toward an address of human morality in the digital age. In doing so the Committee will critique assumptions around concepts such as good and evil, right and wrong, virtue and vice and attempt to carry these inquiries into artificial systems decision-making processes.

Through reviewing the philosophical foundations that define autonomy and ontology, the Committee will address the potential for autonomous capacity of artificially intelligent systems, posing questions of morality in amoral systems, and asking whether decisions made by amoral systems can have moral consequences. Ultimately, it will address notions of responsibility and accountability for the decisions made by autonomous systems and other artificially intelligent technologies.

Disclaimer: *While we have provided recommendations in this document, it should be understood these do not represent a position or the views of IEEE but the informed opinions of Committee members providing insights designed to provide expert directional guidance regarding A/IS. In no event shall IEEE or IEEE-SA Industry Connections Activity Members be liable for any errors or omissions, direct or otherwise, however caused, arising in any way out of the use of this work, regardless of whether such damage was foreseeable.*

Section 1 — Definitions for Classical Ethics in Autonomous and Intelligent Systems Research

Issue:

Assigning foundations for morality, autonomy, and intelligence.

Background

Classical theories of economy in the Western tradition, starting with Plato and Aristotle, embrace three domains: the individual, the family, and the *polis*. The forming of the individual character (*ethos*) is intrinsically related to others, as well as to the tasks of administration of work within the family (*oikos*) and eventually all this expands into the framework of the *polis*, or public space (*poleis*). This means that when we discuss ethical issues of autonomous and intelligent systems we should consider all three traditional economic dimensions that evolved in modernity into an individual morality disconnected from economics and politics. This disconnection was partly questioned by thinkers such as Adam Smith, Hegel, Marx, and others. In particular, Immanuel Kant's ethics located morality within the subject (see: categorical imperative) and separated morality from the outside world and the consequences of being a part of the outside world. The moral autonomous subject of modernity became thus a worldless isolated subject. This process is important to understand in terms of ethics for artificial intelligence since it is, paradoxically, the kind of autonomy that is supposed to be achieved by intelligent machines in the very moment in which we, humans, begin to change our being into digitally networked beings.

There lies a danger in uncritically attributing classical concepts of anthropomorphic autonomy to machines, including using the term *artificial intelligence* to describe them since, in the attempt to make them "moral" by programming moral rules into their behavior, we run the risk of assuming economic and political dimensions that do not exist, or that are not in line with contemporary human societies. As noted above, present human societies are being redefined in terms of digital citizenship via digital social networks. The present public debate about the replaceability of human work by *intelligent* machines is a symptom of this lack of awareness of the economic and political dimensions as defined by classical ethics, reducing ethical thinking to the "morality" of a worldless and isolated machine (a mimic of the modern subject).

Classical Ethics in A/IS

Candidate Recommendations

- Via a return to classical ethics foundations, enlarge the discussion on ethics in autonomous and intelligent systems (A/IS) to include a critical assessment of anthropomorphic presumptions of ethics and moral rules for A/IS. Keep in mind that machines do not, in terms of classical autonomy, comprehend the moral or legal rules they follow, but rather move according to what they are programmed to do, following rules that are designed by humans to be moral.
- Enlarge the discussion on ethics for A/IS to include an exploration of the classical foundations of economy, outlined above, as potentially influencing current views and assumptions around machines achieving isolated autonomy.

Further Resources

- Bielby, J., ed. "Digital Global Citizenship." *International Review of Information Ethics* 23 (November 2015).
- Bendel, O. "Towards a Machine Ethics." Northwestern Switzerland: University of Applied Sciences and Arts, 2013.
- Bendel, O. "Considerations about the Relationship Between Animal and Machine Ethics." *AI & Society* 31, no. 1 (2016): 103–108.
- Capurro, R., M. Eldred, and D. Nagel. Digital Whoness: Identity, Privacy and Freedom in the Cyberworld. Berlin: Walter de Gruyter, 2013.
- Chalmers, D. "The Singularity: A Philosophical Analysis." *Journal of Consciousness Studies* 17, (2010): 7–65.

Issue:

Distinguishing between agents and patients.

Background

Of concern for understanding the relationship between human beings and A/IS is the uncritically applied anthropomorphistic approach toward A/IS that many industry and policy makers are using today. This approach erroneously blurs the distinction between moral agents and moral patients (i.e., subjects), otherwise understood as a distinction between "natural" self-organizing systems and artificial, non-self-organizing devices. As noted above, A/IS devices cannot, by definition, become autonomous in the sense that humans or living beings are autonomous. With that said, autonomy in machines, when critically defined, designates how machines act and operate independently in certain contexts through a consideration of implemented order generated by laws and rules. In this sense, A/IS can, by definition, qualify as autonomous, especially in the case of genetic algorithms and evolutionary strategies. However, attempts

Classical Ethics in A/IS

to implant true morality and emotions, and thus accountability (i.e., autonomy) into A/IS is both dangerous and misleading in that it encourages anthropomorphistic expectations of machines by human beings when designing and interacting with A/IS.

Thus, an adequate assessment of expectations and language used to describe the human-A/IS relationship becomes critical in the early stages of its development, where unpacking subtleties is necessary. Definitions of autonomy need to be clearly drawn, both in terms of A/IS and human autonomy. On one hand A/IS may in some cases manifest seemingly ethical and moral decisions, resulting for all intents and purposes in efficient and agreeable moral outcomes. Many human traditions, on the other hand, can and have manifested as fundamentalism under the guise of morality. Such is the case with many religious moral foundations, where established cultural mores are neither questioned nor assessed. In such scenarios, one must consider whether there is any functional difference between the level of autonomy in A/IS and that of assumed agency (the ability to choose and act) in humans via the blind adherence to religious, traditional, or habitual mores. The relationship between assumed moral customs (mores), the ethical critique of those customs (i.e., ethics), and the law are important distinctions.

The above misunderstanding in definitions of autonomy arise in part because of the tendency for humans to shape artificial creations in their own image, and our desire to lend our human experience to shaping a morphology of artificially intelligent systems. This is not to say that such terminology cannot be used metaphorically, but the difference must be maintained, especially as A/IS begins to resemble human beings more closely. Terms like "artificial intelligence" or "morality of machines" can be used as metaphors, and it does not necessarily lend to misunderstanding to do so. This is how language works and how humans try to understand their natural and artificial environment.

However the critical difference between human autonomy and autonomous systems involves questions of free will, predetermination, and being (ontology). The questions of critical ontology currently being applied to machines are not new questions to ethical discourse and philosophy and have been thoroughly applied to the nature of human *being* as well. John Stuart Mill, for example, is a determinist and claims that human actions are predicated on predetermined laws. He does, however, argue for a reconciliation of human free will with determinism through a theory of compatibility. Millian ethics provides a detailed and informed foundation for defining autonomy that could serve to help combat general assumptions of anthropomorphism in A/IS and thereby address the uncertainty therein (Mill, 1999).

Candidate Recommendation

When addressing the nature of "autonomy" in autonomous systems, it is recommended that the discussion first consider free will, civil liberty, and society from a Millian perspective in order to better grasp definitions of autonomy and to combat general assumptions of anthropomorphism in A/IS.

Classical Ethics in A/IS

Further Resources

- Capurro, Rafael. "Toward a Comparative Theory of Agents." *AI & Society* 27, no. 4 (2012): 479–488.
- King, William Joseph, and Jun Ohya. "The representation of agents: Anthropomorphism, agency, and intelligence." Conference Companion on Human Factors in Computing Systems. ACM, 1996.
- Hofkirchner, W. "Does Computing Embrace Self-Organization?" in *Information and Computation, Essays on Scientific and Philosophical Understanding of Foundations of Information and Computation,* edited by G. Dodig-Crnkovic, M. Burgin, 185–202. London: World Scientific, 2011.
- International Center for Information Ethics.
- Mill, J. S. *On Liberty.* London: Longman, Roberts & Green, 1869.
- Verbeek, P.-P. What Things Do: Philosophical Reflections on Technology, Agency, and Design. University Park, PA: Penn State Press, 2010.

Issue:

There is a need for an accessible classical ethics vocabulary.

Background

Philosophers and ethicists are trained in vocabulary relating to philosophical concepts and terminology. There is an intrinsic value placed on these concepts when discussing ethics and AI, since the layered meaning behind the terminology used is foundational to these discussions, and is grounded in a subsequent entrenchment of values. Unfortunately, using philosophical terminology in cross-discipline instances, for example, in conversation with technologists and policymakers is often ineffective since not everyone has the education to be able to encompass the abstracted layers of meaning contained in philosophical terminology.

However, not understanding a philosophical definition does not detract from the necessity of its utility. While ethical and philosophical theories should not be over-simplified for popular consumption, being able to adequately translate the essence of the rich history of ethics traditions will go a long way in supporting a constructive dialogue on ethics and A/IS. As access and accessibility concerns are also intricately linked with education in communities, as well as secondary and tertiary institutions, society needs to take a vested interest in creating awareness

Classical Ethics in A/IS

for government officials, rural communities, and school teachers. Creating a more "user-friendly" vocabulary raises awareness on the necessity and application of classical ethics to digital societies.

Candidate Recommendation

Support and encourage the efforts of groups raising awareness for social and ethics committees whose roles are to support ethics dialogue within their organizations, seeking approaches that are both aspirational and values-based. A/IS technologists should engage in cross-discipline exchanges whereby philosophy scholars and ethicists attend and present at non-philosophical courses. This will both raise awareness and sensitize non-philosophical scholars and practitioners to the vocabulary.

Further Resources

- Capurro, R. "Towards an Ontological Foundation of Information Ethics." *Ethics and Information Technology* 8, no. 4 (2006): 175–186.
- Flinders, D. J. "In Search of Ethical Guidance: Constructing a Basis for Dialogue 1." *Qualitative Studies in Education* 5, no. 2 (1992): 101–115.
- Saldanha, G. S. "The Demon in the Gap of Language: Capurro, Ethics and Language in Divided Germany." *Information Cultures in the Digital Age*. Wiesbaden, Germany: Springer Fachmedien, 2016. 253–268.

Issue:

Presenting ethics to the creators of autonomous and intelligent systems.

Background

The question arises as to whether or not classical ethics theories can be used to produce meta-level 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.

This problem can be addressed by providing students with job-aids such as checklists, flowcharts, and matrices that help them select and use a principal ethical framework, and then exercise use of those devices with steadily more complex examples. In such an iterative process, students will start to determine for themselves what examples do not allow for perfectly clear decisions, and in fact require some interaction between frameworks. Produced outcomes such as videos, essays, and other formats – such as project-based learning activities – allow for a didactical strategy which proves effective in artificial intelligence ethics education.

Classical Ethics in A/IS

The goal is to provide students a means to use ethics in a manner analogous to how they are being taught to use engineering principles and tools. In other words, the goal is to help engineers tell the story of what they're doing.

- Ethicists should use information flows and consider at a meta-level what information flows do and what they are supposed to do.
- Engineers should then build a narrative that outlines the iterative process of ethical considerations in their design. Intentions are part of the narrative and provide a base to reflect back on those intentions.
- The process then allows engineers to better understand their assumptions and adjust their intentions and design processes accordingly. They can only get to these by asking targeted questions.

This process, one with which engineers are quite familiar, is basically Kantian and Millian ethics in play.

The aim is to produce what in computer programming lexicon is referred to as a *macro*. A macro is code that takes other code as its input(s) and produces unique outputs. This macro is built using the Western ethics tradition of virtue ethics.

Candidate Recommendation

Find ways to present ethics where the methodologies used are familiar to engineering students. As engineering is taught as a collection of *techno-science, logic, and mathematics*, embedding ethical sensitivity into these objective and non-objective processes is essential.

Further Resources

- Bynum, T. W., and S. Rogerson. *Computer Ethics and Professional Responsibility.* Malden, MA: Wiley-Blackwell, 2003.
- Seebauer, E. G., and R. L. Barry. *Fundamentals of Ethics for Scientists and Engineers.* New York: Oxford University Press, 2001.
- Whitbeck, C. "Teaching Ethics to Scientists and Engineers: Moral Agents and Moral Problems." *Science and Engineering Ethics* 1, no. 3 (1995): 299–308.
- Zevenbergen, B. et al. "Philosophy Meets Internet Engineering: Ethics in Networked Systems Research." GTC workshop outcomes paper. Oxford, U.K.: Oxford Internet Institute, University of Oxford, 2015.
- Perez Á., and M. Ángel, "Teaching Information Ethics." *International Review of Information Ethics* 14 (12/2010): 23–28.
- Verbeek, P-P. *Moralizing Technology: Understanding and Designing the Morality of Things*. Chicago: University of Chicago Press, 2011.

Classical Ethics in A/IS

Issue:

Access to classical ethics by corporations and companies.

Background

Many companies, from start-ups to tech giants, understand that ethical considerations in tech design are increasingly important, but are not quite sure how to incorporate ethics into their tech design agenda. How can ethical considerations in tech design become an integrated part of the agenda of companies, public projects, and research consortia? Many corporate workshops and exercises that attempt to consider ethics in technology practices present the conversation as a carte blanche for people to speak about their opinions, but serious ethical discussions are often lacking. As it stands, classical ethics is not accessible enough to corporate endeavors in ethics, and as such, are not applicable to tech projects. There is often, but not always, a big discrepancy between the output of engineers, lawyers, and philosophers when dealing with computer science issues and a large difference in how various disciplines approach these issues. While this is not true in all cases, and there are now several interdisciplinary approaches in robotics and machine ethics as well as a growing number of scientists that hold double and interdisciplinary degrees, there remains a vacuum for the wider understanding of classical ethics theories in the interdisciplinary setting.

Candidate Recommendation

Bridge the language gap between technologists, philosophers, and policymakers. Understanding the nuances in philosophical language is critical to digital society from IoT, privacy, and cybersecurity to issues of Internet governance.

Further Resources

- Bhimani, A. "Making Corporate Governance Count: The Fusion of Ethics and Economic Rationality." *Journal of Management & Governance* 12, no. 2 (2008): 135–147.
- Carroll, A. B. "A History of Corporate Social Responsibility." in *The Oxford Handbook of Corporate Social Responsibility,* edited by Chrisanthi A., R. Mansell, D. Quah, and R. Silverstone. Oxford, U.K.: Oxford University Press, 2008.
- Lazonick, W. "Globalization of the ICT Labor Force." in *The Oxford Handbook of Information and Communication Technologies,* edited by Chrisanthi A., R. Mansell, D. Quah, and R. Silverstone. Oxford, U.K.: Oxford University Press, 2006.
- IEEE P7000™, Model Process for Addressing Ethical Concerns During System Design. This standard will provide engineers and technologists with an implementable process aligning innovation management processes, IS system design approaches and software engineering methods to minimize ethical risk for their organizations, stakeholders and end users. The Working Group is currently in process, and is free and open to join.

Classical Ethics in A/IS

Issue:

Impact of automated systems on the workplace.

Background

The impact of A/IS on the workplace and the changing power relationships between workers and employers requires ethical guidance. Issues of data protection and privacy via big data in combination with the use of autonomous systems by employers is an increasing issue, where decisions made via aggregate algorithms directly impact employment prospects. The uncritical use of A/IS in the workplace in employee/employer relations is of utmost concern due to the high chance for error and biased outcome.

The concept of responsible research and innovation (RRI), a growing area, particularly within the EU, offers potential solutions to workplace bias and is being adopted by several research funders such as the EPSRC, who include RRI core principles in their mission statement. RRI is an umbrella concept that draws on classical ethics theory to provide tools to address ethical concerns from the outset of a project (design stage and onwards).

Quoting Von Schomberg, "Responsible Research and Innovation is a transparent, interactive process by which societal actors and innovators become mutually responsive to each other with a view to the (ethical) acceptability, sustainability and societal desirability of the innovation process and its marketable products (in order to allow a proper embedding of scientific and technological advances in our society)."[1]

When RRI methodologies are used in the ethical considerations of A/IS design, especially in response to the potential bias of A/IS in the workplace, theoretical deficiencies are then often exposed that would not otherwise have been exposed, allowing room for improvement in design at the development stage rather than from a retroactive perspective. RRI in design increases the chances of both relevance and strength in ethically aligned design.

Candidate Recommendation

It is recommended that through the application of RRI, as founded in classical ethics theory, research in A/IS design utilize available tools and approaches to better understand the design process, addressing ethical concerns from the very beginning of the design stage of the project, thus maintaining a stronger more efficient methodological accountability throughout.

Further Resources

- Burget, M., E. Bardone, and M. Pedaste. "Definitions and Conceptual Dimensions of Responsible Research and Innovation: A Literature Review." *Science and Engineering Ethics* 23, no. 1 (2016): 1–9.

1 Von Schomberg (2011) 'Prospects for Technology Assessment in a framework of responsible research and innovation' in: M. Dusseldorp and R. Beecroft (eds). Technikfolgen abschätzen lehren: Bildungspotenziale transdisziplinärer Methoden, Wiesbaden: Vs Verlag, in print, P.9.

Classical Ethics in A/IS

- Von Schomberg, R. "Prospects for Technology Assessment in a Framework of Responsible Research and Innovation," in *Technikfolgen Abschätzen Lehren: Bildungspotenziale Transdisziplinärer Methode*, 39–61, Wiesbaden, Germany: Springer VS, 2011.
- Stahl, B. C. et al. "From Computer Ethics to Responsible Research and Innovation in ICT: The Transition of Reference Discourses Informing Ethics-Related Research in Information Systems." *Information & Management* 51, no. 6 (2014): 810–818.
- Stahl, B. C., and B. Niehaves. "Responsible Research and Innovation (RRI)."
- IEEE P7005™, Standard for Transparent Employer Data Governance is designed to provide organizations with a set of clear guidelines and certifications guaranteeing they are storing, protecting, and utilizing employee data in an ethical and transparent way. The Working Group is currently in process, and is free and open to join.

Section 2 — Classical Ethics From Globally Diverse Traditions

Issue:

The monopoly on ethics by Western ethical traditions.

Background

As human creators, our most fundamental values are imposed on the systems we design. It becomes incumbent on a global-wide community to recognize which sets of values guide the design, and whether or not A/IS will generate problematic (e.g., discriminatory) consequences without consideration of non-Western values. There is an urgent need to broaden traditional ethics in its contemporary form of "responsible innovation" (RI) beyond the scope of "Western" ethical foundations, e.g., utilitarianism, deontology, and virtue ethics; and include other traditions of ethics in RI, including those inherent to, for example, Buddhism, Confucianism, and Ubuntu traditions.

However, this venture poses problematic assumptions even before the issue above can be explored, when, in classifying Western values, we also group together thousands of years of independent and disparate ideas originating from the Greco-Roman philosophical tradition with its Christian-infused cultural heritage.

What is it that one refers to by the term *Western ethics*? By Western ethics, does one refer to philosophical ethics (ethics as a scientific discipline) or is the reference to Western morality?

The *West* (however it may be defined) is an individualistic society, arguably more so than much of the rest of the world, and thus in some aspects should be even less collectively defined than say, "Eastern" ethical traditions. If one is referring to Western values, one must designate which values, and values of which persons and institutions. Additionally, there is a danger in intercultural information ethics (however unconsciously or instinctively propagated) to not only group together all Western traditions under a single banner, but to negatively designate any and all Western influence in global exchange to representing an abusive collective of colonial-influenced ideals. Just because there exists a monopoly of influence by one system over another does not mean that said monopoly is devoid of value, even for systems outside itself. In the same way that culturally diverse traditions have much to offer Western tradition(s), so too do they have much to gain from them.

In order to establish mutually beneficial connections in addressing globally diverse traditions, it is of critical import to first properly distinguish between subtleties in Western ethics (as a discipline) and morality (as its

Classical Ethics in A/IS

object or subject matter). It is also important to differentiate between philosophical ethics (as scientific ethics) and theological ethics. As noted above, the relationship between assumed moral customs (mores), the ethical critique of those customs (i.e., ethics), and the law is an established methodology in scientific communities. Western and Eastern philosophy are very different, as well as are Western and Eastern ethics. Western philosophical ethics uses scientific methods, e.g., the logical, discursive, dialectical approach (models of normative ethics) and the analytical and hermeneutical approach. The Western tradition is not about education and teaching of social and moral values, but rather about the application of fundamentals, frameworks, and explanations. However, several contemporary globally relevant community mores are based in traditional and theological moral systems, requiring a conversation around how best to collaborate in the design and programming of ethics in A/IS amidst differing ethical traditions.

While experts in Intercultural Information Ethics, such as Pak-Hang Wong, highlight the dangers of the dominance of "Western" ethics in AI design, noting specifically the appropriation of ethics by liberal democratic values to the exclusion of other value systems, it should be noted that those same liberal democratic values are put in place and specifically designed to accommodate such differences. However, while the accommodation of differences are, in theory, accounted for in dominant liberal value systems, the reality of the situation reveals a monopoly of, and a bias toward, established Western ethical value systems, especially when it comes to standardization. As Wong notes:

> Standardization is an inherently value-laden project, as it designates the normative criteria for inclusion to the global network. Here, one of the major adverse implications of the introduction of value-laden standard(s) of responsible innovation (RI) appears to be the delegitimization of the plausibility of RI based on local values, especially when those values come into conflict with the liberal democratic values, as the local values (or, the RI based on local values) do not enable scientists and technology developers to be recognized as members of the global network of research and innovation (Wong, 2016).

It does however become necessary for those who do not work within the parameters of accepted values monopolies to find alternative methods of accommodating different value systems. Liberal values arose out of conflicts of cultural and subcultural difference and are designed to be accommodating enough to include a rather wide range of differences.

Responsible innovation (RI) enables policy-makers, scientists, technology developers, and the public to better understand and respond to the social, ethical, and policy challenges raised by new and emerging technologies. Given the historical context from which RI emerges, it should not be surprising that the current discourse on RI is predominantly based on liberal democratic values. Yet, the bias toward liberal democratic values will inevitably limit

Classical Ethics in A/IS

the discussion of RI, especially in the cases where liberal democratic values are not taken for granted. Against this background, it is important to recognize the problematic consequences of RI solely grounded on, or justified by, liberal democratic values.

Candidate Recommendation

In order to enable a cross-cultural dialogue of ethics in technology, discussions in ethics and A/IS must first return to normative foundations of RI to address the notion of "responsible innovation" from value systems not predominant in Western classical ethics, including nonliberal democratic perspectives. Pak-Hang Wong's paper, "Responsible Innovation for Decent Nonliberal Peoples: A Dilemma?" demonstrates the problematic consequences of RI solely grounded on, or justified by, liberal democratic values and should be consulted as a guide to normative foundations in RI.

Further Resources

- Bielby, J. "Comparative Philosophies in Intercultural Information Ethics." *Confluence: Journal of World Philosophies* 2 (2016).
- Hongladarom, S. "Intercultural Information Ethics: A Pragmatic Consideration." *Information Cultures in the Digital Age*, 191–206. Wiesbaden, Germany: Springer Fachmedien, 2016.
- Rodríguez, L. G., and M. Á. P. Álvarez. Ética *Multicultural y Sociedad en Red*. Fundación Telefónica, 2014.
- Wong, P.-H. "What Should We Share?: Understanding the Aim of Intercultural Information Ethics." *ACM SIGCAS Computers and Society* 39, no. 3 (2009): 50–58.
- Wong, P.-H. "Responsible Innovation for Decent Nonliberal Peoples: A Dilemma?" *Journal of Responsible Innovation* 3, no. 2 (2016): 154–168.
- Zeuschner, R. B. *Classical Ethics, East and West: Ethics from a Comparative Perspective*. Boston: McGraw-Hill, 2000.
- Mattingly-Jordan, S., Becoming a Leader in Global Ethics, IEEE, 2017.

Issue:

The application of classical Buddhist ethical traditions to AI design.

Background

According to Buddhism, ethics is concerned with behaving in such a way that the subject ultimately realizes the goal of Liberation. The question "How should I act?" is answered straightforwardly; one should act in such a way that one realizes Liberation (nirvana) in the future, achieving what in Buddhism is understood as "supreme happiness." Thus Buddhist ethics are clearly goal-oriented. In the Buddhist tradition, people attain Liberation when they no longer endure

Classical Ethics in A/IS

any unsatisfactory conditions, when they have attained the state where they are completely free from any passions, including desire, anger, and delusion (to name the traditional three), which ensnare one's self against freedom. In order to attain Liberation, one engages oneself in mindful behavior (ethics), concentration (meditation), and what in Buddhism is deemed as *wisdom*, a term that remains ambiguous in Western scientific approaches to ethics.

Thus ethics in Buddhism is concerned exclusively with how to attain the goal of Liberation, or freedom. In contrast to Western ethics, Buddhist ethics is not concerned with theoretical questions concerning the source of normativity or what constitutes the good life. What makes an action a "good" action in Buddhism is always concerned with whether the action leads, eventually, to Liberation or not. In Buddhism, there is no questioning as to why Liberation is a good thing. It is simply assumed. Such an assumption places Buddhism, and ethical reflection from a Buddhist perspective, in the camp of mores rather than scientifically led ethical discourse, and it is approached as an ideology or a worldview.

While it is critically important to consider, understand, and apply accepted ideologies such as Buddhism in A/IS, it is both necessary to differentiate the methodology from Western ethics, and respectful to Buddhist tradition not to require it be considered in a scientific context. Such assumptions put it at odds with, and in conflict with, the Western foundation of ethical reflection on mores. From a Buddhist perspective, one does not ask why supreme happiness is a good thing; one simply accepts it. The relevant question in Buddhism is not about methodological reflection, but about how to attain Liberation from the necessity for such reflection.

Thus, Buddhist ethics contains potential for conflict with Western ethical value systems which are founded on ideas of questioning moral and epistemological assumptions. Buddhist ethics is different from, for example, utilitarianism, which operates via critical analysis toward providing the best possible situation to the largest number of people, especially as it pertains to the good life. These fundamental differences between the traditions need to be first and foremost mutually understood and then addressed in one form or another when designing A/IS that span cultural contexts.

The main difference between Buddhist and Western ethics is that Buddhism is based upon a metaphysics of relation. Buddhist ethics emphasizes how *action* leads to achieving a *goal*, or in the case of Buddhism, the final Goal. In other words, an action is considered a good one when it contributes to realization of the Goal. It is relational when the value of an action is relative to whether or not it leads to the Goal, the Goal being the reduction and eventual cessation of suffering. In Buddhism, the self is constituted through the relationship between the synergy of bodily parts and mental activities. In Buddhist analysis, the self does not actually exist as a self-subsisting entity. Liberation, or nirvana, consists in realizing that what is known to be the self actually consists of nothing more than these connecting episodes and parts. To exemplify the above, one can draw

Classical Ethics in A/IS

from the concept of privacy as oft explored via intercultural information ethics. The Buddhist perspective understands privacy as a protection, not of self-subsisting individuals, because such do not exist ultimately speaking, but a protection of certain values which are found to be necessary for a well-functioning society and one which can prosper in the globalized world.

The secular formulation of the supreme happiness mentioned above is that of the reduction of the experience of suffering, or reduction of the metacognitive state of suffering as a result of lifelong discipline and meditation aimed at achieving proper relationships with others and with the world. This notion of the reduction of suffering is something that can resonate well with certain Western traditions, such as epicureanism and the notion of ataraxia, freedom from fear through reason and discipline, and versions of consequentialist ethics that are more focused on the reduction of harm. It also encompasses the concept of phronesis or practical wisdom from virtue ethics.

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.

Candidate Recommendation

In considering the nature of human and autonomous systems interactions, the above notion of "proper relationships" through Buddhist ethics can provide a useful platform that results in ethical statements formulated in a relational way, instead of an absolutist way, and is recommended as an additional methodology, along with Western values methodologies, to addressing human/computer interactions.

Further Resources

- Capurro, R. "Intercultural Information Ethics: Foundations and Applications." *Journal of Information, Communication & Ethics in Society* 6, no. 2 (2008): 116.
- Ess, C. "Ethical Pluralism and Global Information Ethics." Et*hics and Information Technology* 8, no. 4 (2006): 215–226.
- Hongladarom, S. "Intercultural Information Ethics: A Pragmatic Consideration," in *Information Cultures in the Digital Age* edited by K. M. Bielby, 191–206. Wiesbaden, Germany: Springer Fachmedien Wiesbaden, 2016.
- Hongladarom, S. et al. "Intercultural Information Ethics." I*nternational Review of Information Ethics* 11 (2009): 2–5.

Classical Ethics in A/IS

- Nakada, M. "Different Discussions on Roboethics and Information Ethics Based on Different Contexts (Ba). Discussions on Robots, Informatics and Life in the Information Era in Japanese Bulletin Board Forums and Mass Media." *Proceedings Cultural Attitudes Towards Communication and Technology* (2010): 300–314.
- Mori, Ma. *The Buddha in the Robot.* Suginami-ku, Japan: Kosei Publishing, 1989.

Issue:

The application of Ubuntu ethical traditions to A/IS design.

Background

In his article, "African Ethics and Journalism Ethics: News and Opinion in Light of Ubuntu," Thaddeus Metz frames the following question: "What does a sub-Saharan ethic focused on the good of community, interpreted philosophically as a moral theory, entail for the duties of various agents with respect to the news/opinion media"? (Metz, 2015, 1). When that question is applied to A/IS) *viz*: "If an ethic focused on the good of community, interpreted philosophically as a moral theory, is applied to autonomous and intelligent systems, what would the implications be on the duties of various agents"? Agents in this regard would therefore be the following:

1. Members of the A/IS research community
2. A/IS programmers/computer scientists
3. A/IS end-users
4. Autonomous and intelligent systems

Ubuntu is a Sub-Saharan philosophical tradition. Its basic tenet is that a person is a person through other persons. It develops further in the notions of caring and sharing as well as identity and belonging, whereby people experience their lives as bound up with their community. A person is defined in relation to the community since the sense of being is intricately linked with belonging. Therefore, community exists through shared experiences and values: "to be is to belong to a community and participate" also *motho ke motho ka batho* "A person is a person because of other people."

Very little research, if any at all, has been conducted in light of Ubuntu ethics and A/IS, but its focus will be within the following moral domains:

1. Between the members of the A/IS research community
2. Between the A/IS community/programmers/computer scientists and the end-users
3. Between the A/IS community/programmers/computer scientists and A/IS
4. Between the end-users and A/IS
5. Between A/IS and A/IS

Classical Ethics in A/IS

Considering a future where A/IS will become more entrenched in our everyday lives, one must keep in mind that an attitude of sharing one's experiences with others and caring for their well-being will be impacted. Also by trying to ensure solidarity within one's community, one must identify factors and devices that will form part of their lifeworld. If so, will the presence of A/IS inhibit the process of partaking in a community, or does it create more opportunities for doing so? One cannot classify A/IS as only a negative or disruptive force; it is here to stay and its presence will only increase. Ubuntu ethics must come to grips with and contribute to the body of knowledge by establishing a platform for mutual discussion and understanding.

Such analysis fleshes out the following suggestive comments of Desmond Tutu, renowned former chair of South Africa's Truth and Reconciliation Commission, when he says of Africans, "(we say) a person is a person through other people... I am human because I belong" (Tutu, 1999). I participate, I share. Harmony, friendliness, and community are great goods. Social harmony is for us the *summum bonum* — the greatest good. Anything that subverts or undermines this sought-after good is to be avoided (2015:78).

In considering the above, it is fair to state that community remains central to Ubuntu. In situating A/IS within this moral domain, it will have to adhere to the principles of community, identity and solidarity with others. While virtue ethics questions the goal or purpose of A/IS and deontological ethics questions the duties, the fundamental question asked by Ubuntu would be "how does A/IS affect the community in which it is situated"? This question links with the initial question concerning the duties of the various moral agents within the specific community. Motivation becomes very important, because if A/IS seek to detract from community it will be detrimental to the identity of this community, i.e., in terms of job losses, poverty, lack in education and skills training. However, should A/IS seek to supplement the community, i.e., ease of access, support systems, etc., then it cannot be argued that it will be detrimental. It therefore becomes imperative that whosoever designs the systems must work closely both with ethicists and the target community/audience/end-user to ascertain whether their needs are identified and met.

Candidate Recommendations

- It is recommended that a concerted effort be made toward the study and publication of literature addressing potential relationships between Ubuntu ethical traditions and A/IS value design.
- A/IS designers and programmers must work closely with the end-users and target communities to ensure their design aims are aligned with the needs of the end-users and target communities.

Further Resources

- Lutz, D. W. "African Ubuntu Philosophy and Global Management." *Journal of Business Ethics* 84 (2009): 313–328.

Classical Ethics in A/IS

- Metz, T. "African Ethics and Journalism Ethics: News and Opinion in Light of Ubuntu," *Journal of Media Ethics: Exploring Questions of Media Morality* 30 no. 2 (2015): 74–90. doi: 10.1080/23736992.2015.1020377
- Tutu, D. *No Future Without Forgiveness*. London: Rider, 1999.

Issue:

The application of Shinto-influenced traditions to A/IS design.

Background

Alongside the burgeoning African Ubuntu reflections on A/IS, other indigenous techno-ethical reflections boast an extensive engagement. One such tradition is Japanese Shinto indigenous spirituality, (or, *Kami-no-michi*), often cited as the very reason for Japanese robot and autonomous systems culture, a culture more prevalent in Japan than anywhere else in the world. Popular Japanese AI, robot and video-gaming culture can be directly connected to indigenous Shinto tradition, from the existence of *kami* (spirits) to puppets and automata.

The relationship between A/IS and a human being is a personal relationship in Japanese culture and, one could argue, a very natural one. The phenomenon of *relationship* in Japan between humans and automata stands out as unique to technological relationships in world cultures, since the Shinto tradition is arguable the only animistic and naturalistic tradition that can be directly connected to contemporary digital culture and A/IS. From the Shinto perspective, the existence of A/IS, whether manifested through robots or other technological autonomous systems, is as natural to the world as are rivers, forests, and thunderstorms. As noted by Spyros G. Tzafestas, author of *Roboethics: A Navigating Overview*, "Japan's harmonious feeling for intelligent machines and robots, particularly for humanoid ones," (Tzafestas, 2015, 155) colors and influences technological development in Japan, especially robot culture.

The word Shinto can be traced to two Japanese concepts, Shin, meaning spirit, and "to", the philosophical path. Along with the modern concept of the android, which can be traced back to three sources — one, to its Greek etymology that combines "άνδρας": andras (man) and gynoids, "γυνή": gyni (woman); two, via automatons and toys as per U.S. patent developers in the 1800s, and three to Japan, where both historical and technological foundations for android development have dominated the market since the 1970s — Japanese Shinto-influenced technology culture is perhaps the most authentic representation of the human-automaton interface.

Shinto tradition is an animistic religious tradition, positing that everything is created with, and maintains, its own spirit (*kami*) and is animated by that spirit, an idea that goes a long way to defining autonomy in robots from

Classical Ethics in A/IS

a Japanese viewpoint. This includes on one hand, everything that Western culture might deem natural, including rivers, trees, and rocks, and on the other hand, everything artificially (read: *artfully*) created, including vehicles, homes, and automata (i.e., robots). Artifacts are as much a part of nature in Shinto as are animals, and are considered naturally beautiful rather than falsely artificial.

A potential conflict between Western concepts of nature and artifact and Japanese concepts of the same arises when the two traditions are compared and contrasted, especially in the exploration of *artificial* intelligence. Where in Shinto, the artifact as *artificial* represents creation and authentic being (with implications for defining autonomy), the same is designated as secondary and oft times unnatural, false, and counterfeit in Western ethical philosophical tradition, dating back to Platonic and Christian ideas of separation of form and spirit. In both traditions, culturally presumed biases define our relationships with technology. While disparate in origin and foundation, both Western classical ethics traditions and Shinto ethical influences in modern A/IS have similar goals and outlooks for ethics in A/IS, goals that are centered in *relationship*.

Candidate Recommendation

Where Japanese culture leads the way in the synthesis of traditional value systems and technology, we recommend that efforts in A/IS ethics explore the Shinto paradigm as representative, though not necessarily as directly applicable, to global efforts in understanding and applying traditional and classical ethics methodologies to ethics for A/IS.

Further Resources

- Holland-Minkley, D. F. "God in the Machine: Perceptions and Portrayals of Mechanical Kami in Japanese Anime." PhD Diss. University of Pittsburgh, 2010.
- Jensen, C. B., and A. Blok. "Techno-Animism in Japan: Shinto Cosmograms, Actor-Network Theory, and the Enabling Powers of Non-Human Agencies." *Theory, Culture & Society* 30, no. 2 (2013): 84–115.
- Tzafestas, S. G. *Roboethics: A Navigating Overview*. Cham, Switzerland: Springer, 2015.

Section 3 — Classical Ethics for a Technical World

Issue:
Maintaining human autonomy.

Background

Autonomous and intelligent systems present the possibility for a digitally networked intellectual capacity that imitates, matches, and supersedes human intellectual capacity, including, among other things, general skills, discovery, and computing function. In addition, A/IS can potentially acquire functionality in areas traditionally captured under the rubric of what we deem unique human and social ability. While the larger question of ethics and AI looks at the implications of the influence of autonomous systems in these areas, the pertinent issue is the possibility of autonomous systems imitating, influencing, and then determining the norms of human autonomy. This is done through the eventual negation of independent human thinking and decision-making, where algorithms begin to inform through targeted feedback loops what it is we *are* and what it is we should decide. Thus, how can the academic rigor of traditional ethics speak to the question of maintaining human autonomy in light of algorithmic decision-making?

How will AI and autonomous systems influence human autonomy in ways that may or may not be advantageous to the good life, and perhaps even if advantageous, may be detrimental at the same time? How do these systems affect human autonomy and decision-making through the use of algorithms when said algorithms tend to inform ("in-form") via targeted feedback loops?

Consider, for example, Google's autocomplete tool, where algorithms attempt to determine one's search parameters via the user's initial keyword input, offering suggestions based on several criteria including search patterns. In this scenario, autocomplete suggestions influence, in real-time, the parameters the user phrases their search by, often reforming the user's perceived notions of what it was they were looking for in the first place, versus what they might have actually originally intended.

Targeted algorithms also inform as per emerging IoT applications that monitor the user's routines and habits in the analog world. Consider for example that our bio-information is, or soon will be, available for interpretation by autonomous systems. What happens when autonomous systems can inform the user in ways the user is not even aware of, using one's bio-information in targeted advertising campaigns that seek to influence the user in real-time feedback loops

Classical Ethics in A/IS

based on the user's biological reactions (pupil dilation, body temperature, emotional reaction), whether positive or negative, to that very same advertising, using information *about* our being to *in-form* (and re-form) our being?

On the other hand, it becomes important not to adopt dystopian assumptions concerning autonomous machines threatening human autonomy. The tendency to think only in negative terms presupposes a case for interactions between autonomous machines and human beings, a presumption not necessarily based in evidence. Ultimately the behavior of algorithms rests solely in their design, and that design rests solely in the hands of those who designed them. Perhaps more importantly, however, is the matter of choice in terms of how the *user* chooses to interact with the algorithm. Users often don't know when an algorithm is interacting with them directly, or their data which acts as a proxy for their identity. The responsibility for the behavior of algorithms remains with both the designer and the user and a set of well-designed guidelines that guarantee the importance of human autonomy in any interaction. As machine functions become more autonomous and begin to operate in a wider range of situations, any notion of those machines working for or against human beings becomes contested. Does the machine work *for* someone in particular, or for particular groups but not for others, and who decides on the parameters? The machine itself? Such questions become key factors in conversations around ethical standards.

Candidate Recommendation

- An ethics by design methodology is the first step to addressing human autonomy in AI, where a critically applied ethical design of autonomous systems preemptively considers how and where autonomous systems may or may not dissolve human autonomy.
- The second step is a pointed and widely applied education curriculum that encompasses school age through university, one based on a classical ethics foundation that focuses on providing choice and accountability toward digital being as a priority in information and knowledge societies.

Further Resources

- van den Berg, B. and J. de Mul. "Remote Control. Human Autonomy in the Age of Computer-Mediated Agency," in: *Autonomic Computing and Transformations of Human Agency. Philosophers of Law Meeting Philosophers of Technology,* edited by Mireille Hildebrandt and Antoinette Rouvroy, 46–63. London: Routledge, 2011.
- Costa, L. "A World of Ambient Intelligence," Chapter 1 in *Virtuality and Capabilities in a World of Ambient Intelligence,* 15–41. Cham, Switzerland: Springer International, 2016.

Classical Ethics in A/IS

- Verbeek, P.-P. "Subject to Technology on Autonomic Computing and Human Autonomy," in *The Philosophy of Law Meets the Philosophy of Technology: Autonomic Computing and Transformations of Human Agency,* edited by. M. Hildebrandt and A. Rouvroy. New York: Routledge, 2011.

Issue:

Applying goal-directed behavior (virtue ethics) to autonomous and intelligent systems.

Background

Initial concerns regarding A/IS also include questions of function, purpose, identity, and agency, a continuum of goal-directed behavior, with function being the most primitive expression. How can classical ethics act as a regulating force in autonomous technologies as goal-directed behavior transitions from being externally set by operators to being indigenously set? The question is important not just for safety reasons, but for mutual productivity. If autonomous systems are to be our trusted, creative partners, then we need to be confident that we possess mutual anticipation of goal-directed action in a wide variety of circumstances.

A virtue ethics approach has merits for accomplishing this even without having to posit a "character" in an autonomous technology, since it places emphasis on habitual, iterative action focused on achieving excellence in a chosen domain or in accord with a guiding purpose. At points on the goal-directed continuum associated with greater sophistication, virtue ethics become even more useful by providing a framework for prudent decision-making that is in keeping with the autonomous system's purpose, but allows for creativity in how to achieve the purpose in a way that still allows for a degree of predictability. An ethics that does not rely on a decision to refrain from transgressing, but instead to prudently pursue a sense of purpose informed by one's identity, might provide a greater degree of insight into the behavior of the system.

Candidate Recommendation

Program autonomous systems to be able to recognize user behavior as being those of specific types of behavior and to hold expectations as an operator and co-collaborator whereby both user and system mutually recognize the decisions of the autonomous system as virtue ethics based.

Further Resources

- Lennox, J. G. "Aristotle on the Biological Roots of Virtue." *Biology and the Foundations of Ethics,* edited by J. Maienschein and M. Ruse, 405–438. Cambridge, U.K.: Cambridge University Press, 1999.
- Boden, M. A., ed. *The Philosophy of Artificial Life*. Oxford, U.K.: Oxford University Press, 1996.

Classical Ethics in A/IS

- Coleman, K. G.. "Android Arete: Toward a Virtue Ethic for Computational Agents." *Ethics and Information Technology* 3, no. 4 (2001): 247–265.

Issue:

A requirement for rule-based ethics in practical programming.

Background

Research in machine ethics focuses on simple moral machines. It is deontological ethics and teleological ethics that are best suited to the kind of practical programming needed for such machines, as these ethical systems are abstractable enough to encompass ideas of non-human agency, whereas most modern ethics approaches are far too human-centered to properly accommodate the task.

In the *deontological model*, duty is the point of departure. Duty can be translated into rules. It can be distinguished into rules and meta rules. For example, a rule might take the form "Don't lie!", whereas a meta rule would take the form of Kant's categorical imperative: "Act only according to that maxim whereby you can, at the same time, will that it should become a universal law."

A machine can follow simple rules. Rule-based systems can be implemented as formal systems (also referred to as axiomatic systems), and in the case of machine ethics, a set of rules is used to determine which actions are morally allowable and which are not. Since it is not possible to cover every situation by a rule, an inference engine is used to deduce new rules from a small set of simple rules (called axioms) by combining them. The morality of a machine comprises the set of rules that are deducible from the axioms.

Formal systems have an advantage since properties such as decidability and consistency of a system can be effectively examined. If a formal system is decidable, every rule is either morally allowable or not, and the "unknown" is eliminated. If the formal system is consistent, one can be sure that no two rules can be deduced that contradict each other. In other words, the machine never has moral doubt about an action and never encounters a deadlock.

The disadvantage of using formal systems is that many of them work only in closed worlds like computer games. In this case, what is not known is assumed to be false. This is in drastic conflict with real world situations, where rules can conflict and it is impossible to take into account the totality of the environment. In other words, consistent and decidable formal systems that rely on a closed world assumption can be used to implement an ideal moral framework for a machine, yet they are not viable for real world tasks.

One approach to avoiding a closed world scenario is to utilize self-learning algorithms, such as case-based reasoning approaches.

Classical Ethics in A/IS

Here, the machine uses "experience" in the form of similar cases that it has encountered in the past or uses cases which are collected in databases.

In the context of the *teleological model*, the consequences of an action are assessed. The machine must know the consequences of an action and what the action's consequences mean for humans, for animals, for things in the environment, and, finally, for the machine itself. It also must be able to assess whether these consequences are good or bad, or if they are acceptable or not, and this assessment is not absolute: while a decision may be good for one person, it may be bad for another; while it may be good for a group of people or for all of humanity, it may be bad for a minority of people. An implementation approach that allows for the consideration of potentially contradictory subjective interests may be realized by decentralized reasoning approaches such as agent-based systems. In contrast to this, centralized approaches may be used to assess the overall consequences for all involved parties.

Candidate Recommendation

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.

Further Resources

- Bendel, O. *Die Moral in der Maschine: Beiträge zu Roboter-und Maschinenethik*. Heise Medien, 2016.
- Bendel, O. "LADYBIRD: the Animal-Friendly Robot Vacuum Cleaner." *The 2017 AAAI Spring Symposium Series*. Palo Alto, CA: AAAI Press, 2017.
- Fisher, M., L. Dennis, and M. Webster. "Verifying Autonomous Systems." *Communications of the ACM* 56, no. 9 (2013): 84–93.
- McLaren, B. M. "Computational Models of Ethical Reasoning: Challenges, Initial Steps, and Future Directions." *IEEE Intelligent Systems* 21, no. 4 (2006): 29–37.
- Perez Alvarez, M. A. "Tecnologías de la Mente y Exocerebro o las Mediaciones del Aprendizaje," 2015.