Reflections on Artificial Intelligence Alignment with Human Values: A Phenomenological Perspective

Han, Shengnan; Kelly, Eugene; Nikou, Shahrokh; Svee, Eric-Oluf

Published in:
Proceedings of the 28th European Conference on Information Systems (ECIS), An Online AIS Conference, June 15-17, 2020

Published: 01/01/2020

Document Version
(Peer reviewed version when applicable)

Document License
Publisher rights policy

REFLECTIONS ON ARTIFICIAL INTELLIGENCE ALIGNMENT WITH HUMAN VALUES: A PHENOMENOLOGICAL PERSPECTIVE

Research paper

Han, Shengnan, Stockholm University, Stockholm, Sweden, shengnan@dsv.su.se

Kelly, Eugene, Now York Institute of Technology, New York, USA, ekelly@nyit.edu

Nikou, Shahrokh, Åbo Akademi University, Turku, Finland, Shahrokh.Nikou@abo.fi

Svee, Eric-Oluf, Stockholm University, Stockholm, Sweden, eric-sve@dsv.su.se

Abstract

The need for a systematic approach to work with artificial intelligence (AI) is current and rapidly growing. It is important that Information Systems researchers get ahead of public sentiment and be able to provide proactive commentary about the current state-of-the-art, as well as solutions for future systems. One critical question is how can we ensure value alignment between AI and human values through AI operations from design to use? For the purposes of this discussion, we adopt the phenomenological theories of material values and technological mediation to be that beginning step. In this paper, we firstly analyze the AI phenomenon from selected resources from the top IS research outlets (basket of 8 journals and 5 AI journals in IS). Secondly, we briefly present what are material values and technological mediation and reflect on the AI value alignment principle through the lenses of these theories. Supported by these new understandings and reflections, we propose to build a common principle of human values to understand the AI value alignment principle through phenomenological theories. The paper contributes unique aspect of material values that are not addressed in the current AI research.

Keywords: human values, AI, material values, technological mediation, phenomenology.

1 Introduction

In the long history of human development, we have continuously sought to extend our physical and mental reach beyond our current limitations, especially in developing technologies to serve our needs and satisfy our desires. In recent years, artificial intelligence (AI) has grown explosively. The results from Alpha Go (Silver et al., 2017) demonstrate the great scientific advances in deep mind research and further provide strong evidence that AI can achieve human level (or above human level) performance without human interventions. Sophia, the social robot (Sophia, 2019) has travelled around the world and presented her thoughts on AI and on interesting organizational, political, and societal questions. Quantum computing now has a multi-factor increase in processing speed over conventional computers (Giles, 2019). This "quantum supremacy" can perform mathematical calculations in 200 seconds that would take a supercomputer 10,000 years (Herman, 2019). Scientists believe that, within our lifetime, machines will obtain the general flexible intelligence that can learn and do things across different domains, so-called artificial general intelligence (AGI) (e.g. Tegmark, 2017). In tandem with the development of AI and humans' efforts to expand the limits of the possible, we will sooner or later reach a kind of singularity where the machine intelligence that we have created exceeds the computing power of human brains. The singularity can be seen as a result of humans succeeding in extending themselves beyond their intellectual limitations. However, this singularity could be detrimental to human society, and robots could eventually dominate humans (Kurzweil, 2005). Still others also argue that we overrate AI and its impacts on humans. Aleksander (2017) has argued AI's "algorithmic" mind is different in kind from the human mind. AI should be realistically considered as a "helpful" technology, with highly

optimized mechanistic functions and super intelligence to address human needs regardless of its "human-like" appearance (eyes, face, ears, etc.).

Nonetheless, one fundamental and critical question is raised and intensively debated: *how can we align AI with human values?* This value alignment principle is pioneered by professor Stuart Russell (e.g. Russell et al., 2015), one of the leading computer scientists in AI research. He formulates the question as "*how can we build autonomous systems with values that 'are aligned with those of the human race?"* This is also named the AI value alignment principle. In another formulation, "highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation" (Conn, 2017). The objective of AI is to maximize the realization of human values (e.g. Yudkowsky, 2011). Several influential scientists and business leaders have shared their concerns about this question (e.g. Aleksander, 2017; Galanos, 2019). In their critical review and analysis Sotala and Yamposkiy (2017) present both risks and solutions for designing and implementing a safe AI. They conclude that, because of the unresolved disagreements in the disciplines of philosophy and axiology regarding what human values are and how to align these values in regulating and designing AI, the solutions provided or discussed in research are not easily implemented. In the context of these disputes, Schneider (2018) proposes an open question about the problems of AI consciousness: "*does consciousness simply go hand-in-hand with sophisticated computation?*" First an AI or Artificial General Intelligence (AGI) may have an architecture that bypasses consciousness altogether. Second, consciousness could be limited to carbon substrates, due to its ability to form stronger and more stable bonds with other compounds. In fact, the second question has already been answered: a carbon nanotube-based chip is now pushing the limits of silicon for expanding Moore's Law (Hills et al., 2019). If AI (or AGI) is a conscious being, then it is imperative to align human values with AI to assure AI's safe and beneficial behavior towards us.

The need for a systematic approach for working with artificial intelligence is current and growing. The near exponential growth of AI, where it seems that such systems surpass some human milestone weekly, has often been met with confusion, fear, and approbation (e.g. Aleksander, 2017; Galanos, 2019). The Information Systems (IS) discipline has the unique ability and distinctiveness to expand across different discipline boundaries to study and understand new technological phenomena through the sociotechnical axis of cohesion (Saker et al., 2019). Surprisingly, the IS community has not yet paid sufficient attention to this AI phenomenon and has contributed a very limited understanding of human values in general and AI alignment with human values in particular (See Appendix). We acknowledge that there are quite a lot of literature outside IS the treats values and AI, especially in AI research, computer science, and machine ethics and philosophy. This calls serious attention from IS community to engage more fruitful research in AI and human values. IS may not necessary be the central place of such reflections, but the contribution from our unique sociotechnical axis of cohesion will certainly generate more profound understanding and advance the knowledge of AI and human values. Hence, IS research is in the great need of developing a common principle by which to qualify and understand human values in the coming age of AI. To achieve the AI value alignment principle, we need to develop an inclusive and common understanding of human values, to study AI relations with humans, and to build a common language that designers and users can communicate to reach value consensus and to increase the likelihood of AI value alignment with human values.

In this paper, we firstly analyze the AI phenomenon from selective resources from the top IS research outlets (basket of 8 journals and 5 AI journals in IS). Secondly, we briefly present what material values and technological mediation are and reflect on the AI value alignment principle through the lens of these theories. Upon this understanding and these reflections, we propose to build a common principle of human values to understand the AI alignment principle through the lens of the phenomenology theories. This paper argues that phenomenological theories bring new interpretations and understanding of the AI alignment principle. Through the lens of material values, simplified AI "artificial" values that are prioritized in AI's algorithmic mind are not possible to be aligned with the richness of human values. The alignment can be fostered between designers and users through technological mediation. Upon that foundation, we propose a common principle that we think the IS community can use as the beginning

step. Impactful studies on human values should be promoted in the coming AI age in order to ensure that AI will be a "helpful" technology instead of being "detrimental".

2 Method

We adopt the phenomenological approach to reflect the AI value alignment principle: Phenomenology is a philosophical approach that we can use to analyze humans' relationship to the world. This is a method to describe the world to reveal reality in the full and original richness of meaning (Merleau-Ponty, 1962). Max Scheler (1874-1928) was one of the leading scholars in the German-Austrian school of phenomenology. He developed in great detail the material ethics of values and theorized within his philosophy how values (material values) guide human actual actions. We follow Kelly's (2011) interpretation of Scheler's theory of material values for the discussion of this paper. A phenomenological perspective has also been adapted to investigate technology's relations to humans and societies. Alignment between AI and human values can be interpreted as a balanced relation between AI and humans. This motives us to adapt phenomenological perspective to understand the AI alignment principle. Martin Heidegger's "Die Frage nach der Technik" (lecture, 1954) (classical phenomenology) asserts that "the essence of technology is nothing technological" and that we must grasp technology's fundamental impact on what is forgotten while we are using it to reveal the world. The post-phenomenological philosophy of technology emphasizes the mediating role of technological artifacts in human/world relations (Ihde, 1990). Technological mediation can be investigated without following the techno-centric construct that humans and society will be dominated and determined by technology, but also without adhering to an anthropocentric view that technology is merely an instrument and a tool (Verbeek, 2011). Instead, this mediation theory pays strong attention to the mutual shaping of technology and human. This approach takes actual technologies and technological development as a starting point for both empirical investigations and philosophical analysis (Rosenberger and Verbeek, 2015).

Verbeek (2003) formulates this approach as such: "technologies co-constitute these relationships by shaping people's perceptions and interpretations (the ways in which reality can be present for humans) on the one hand and their actions and forms of engagement with reality (the ways in which humans can be present in their world) on the other. Human interpretations of, and ways of being involved with, reality are mediated by technological artefacts" (p 93).

3 The Phenomenon of AI Challenges

Science has advanced our knowledge of intelligence. The legendary computer scientist, John von Neumann said in the 1950s that "the ever-accelerating progress of technology … gives the appearance of approaching some essential singularity in the history of the race beyond which human affairs, as we know them, could not continue" (cited from Ulam, 1958). Two important ideas spring from this quote: (1) human progress is exponential; and (2) such exponential growth can turn explosive and profoundly transformative. Max Tegmark, the leading physicist in future AI research, argues that if we achieve artificial general intelligence (AGI), then humans will extend the limits of our own intelligence and create more value for the life that we prefer (Tegmark, 2017). In recent years, we have seen the accelerated changes of AI and we have estimated various impacts of this "possible" singularity on the human race. For example, we may enter a posthuman era (Bostrom, 2008) where we may experience the merging of biological and machine intelligence, thus a human-machine co-evolution (Kurzweil, 2005) or where fears of human extinction brought about by the coming AI paralyze further efforts (see Galanos, 2019). Stuart Russell (2017), a leading computer scientist in AI research, has proposed three principles for creating a safe and beneficial AI (i.e. AI value alignment principle).

1. A principle of altruism: the AI's only objective is to maximize the realization of human values. Here, human values are defined as what human would prefer their life to be like.

2. A law of humility: AI is uncertain of what human values are, and thus, the AI observes and learns the values of what we prefer our life to be like. AI obtains information primarily by observation of human

choices, so our choices reveal what we prefer our lives to be like. The challenge here is that there are many different "we", and that values are diversified, and culturally and socially rooted. This may require a *simplification* of human values and the creation of a single cultural paradigm as convergence theory has proposed.

3. In order to achieve the value alignments between AI and humans, we, in this process, must learn to be better persons, or, perhaps, simpler. The aim should be insuring the supply to the less fortunate at least the minimum possession of things having "lower" value-goods such as healthcare, food and shelter, and meaningful work.

Currently, technical solutions have been proposed to solve this alignment problem. For example, in research into inverse reinforcement learning and inferring human preferences, and also into the design of algorithms that AI systems can learn from human behavior and then infer and realize the intended goal of their users, designers, and society in general. In the pursuit of more advanced technical solutions for safe and beneficial AI, we also need to address the social challenges that are triggered by AI development.

The State of the Art: IS Research on AI and Human Values

Berente et al. (2019) define AI as machines performing the cognitive functions typically associated with humans, including perceiving, reasoning, learning, interacting, etc. They emphasize that "AI is not confined to one or a few applications, but rather is a pervasive economic, societal, and organizational phenomenon". Surprisingly, the IS community has not yet paid sufficient attention to this AI phenomenon and has provided very limited understanding of human values in general, and AI alignment with human values in particular (See Appendix). Most of the papers (Appendix, Table 1) limit their contributions towards AI technical problems, and very few have implicitly discussed AI's impacts on humans, organizations, and society in general. For example, Ransbotham et al. (2016, p. 1) argue that while on the one hand Information Technology provides many advantages and benefits to these humans, organizations and society in general, on the other hand it has potential to create new vulnerabilities such as online harassment, incivility, a merely algorithmic ethics and bias. In another study, Aleksander (2017) argues that, as the robots and machines operate in an algorithmic way and not in a truly cognitive and conscious human way, AI in general can impose serious threats to humanity if the algorithms are biased. Elkins et al. (2013) demonstrated that using artificial technology and integrating the AI into advanced expert systems inadvertently imposes threats to human experts, as well as inhibit the expertise and ability of users to accept the technology. Moreover, Aleksander (2004) argued that, while AI technology is developing more and more, advances in AI potentially can overcome some of the unforeseen difficulties in the pursuit of very ambitious targets. Glezer (2003, p. 65) argued that using artificial intelligence for automation of tasks is problematic, for the software agents often interfere with the human ability to specify the amount of control they would like over the agent's behavior. Saltz and Dewar (2019) recently conducted a literature review on ethical considerations when big data, machine learning, and data science are used. The authors concluded that, while these approaches are emerging in disciplines that involve the analysis of data to solve problems and develop insights, they can also introduce many ethical issues and concerns such as the possible loss of privacy or the harming of a sub-category of the population via a classification algorithm (Saltz and Dewar, 2019). In particular, when referring to data- and model-related challenges, they found that the preparation, storage, and dissemination of data and data science models introduce privacy issues which have the potential to cause harm.

In IS research, for example, value-sensitive design has introduced human values into design (e.g. Friedman et al., 2008) Values, such as privacy, the level of freedom, informed consent, and trust, have been represented by properties or measures or design features. Essentially, such values should support our wellbeing and the quality of human lives (van der Hoven, 2009). Sutcliffe and Thew (2014) also propose to understand users' values and then design appropriate software and system architecture, while Tuunanen and Kuo (2015) emphasize the important of incorporating users' value in requirement engineering. In Green IS research, values rarely have been studied or have been mentioned at a

superfluous level (Paulsson et al., 2019). For example, Watson et al. (2010) acknowledge self-interest as a frequent driver, which may not work in favor of creating a sustainable civilization in the long term. Self-interest is inevitable in any self-aware system, but its effects can be mitigated. Melville (2010) mentions several critical factors that are essentially value-based, such as the relevance to practice, choice of metrics, and belief formation, among others. However, there is little clarification and understanding regarding how these values should be incorporated, measured or at all represented. Moreover, the question remains as to what specific set of values we should incorporate and choose for a specific design of an artifact. Hence, IS research is lacking a common principle by which to qualify and understand human values in the coming age of AI.

4 A Phenomenological Perspective to Understand Human Values and Human Relationships with Technological Artefacts

The Material Ethics of Values

The theory of material ethics of values is presented in Max Scheler's Formalism in Ethics and the Material Ethics of Values, with the first part published in 1913 and the second in 1916 (Kelly, 2011). To clarify what is material value and its key philosophical extensions, we need to acknowledge that many phenomenologists such as Edmund Husserl and Nicolai Hartmann developed their own theory of values as part of the whole phenomenological movement. Schutz (1968) interprets Scheler's material values as "… concrete values and their hierarchical order form a realm of material, aprioristic data which is disclosed to us by emotional intuition" (p.486). Values, according to Scheler, are a particular class of ideal objects that are objective, eternal, and immutable. They are cognized in acts of feeing, which accordingly have epistemic value. He calls this system "ethical absolutism and objectivism". The values are intentional objects of our feelings and the mode in which values are knowable to us is beyond the grasp of the intellect. For example, colors exist as aprioristic data, as colors are given to us as a material quality (intentional object) via our visual perception. The theory of material values also states that values are independent of our subjective bodily states. For instance, our understanding of the phenomenon of sadness may remain unchanged while our subjective emotional state changes from heartbroken to composed. Finally, this theory claims that values are independent of things and relationships of all kinds which are their carriers: the so-called "goods". Values may, however, appear on a single example of a good. For example, friendship will remain unchanged as a material value although your friend betrayed you. The colour blue remains the same visible quality when we are not observing it. A value such as utility is similarly independent of our having something in use. Both can be thematised by phenomenology by repeating the intentional acts in which they are given. Thus, we can describe the material content of the value in question, just as Aristotle tries to describe in *NE* the content of the phenomenon of courage. By this reasoning, it is possible for us to discover the realm of concrete values in an aprioristic way and that this knowledge is by no means derived from our experience of goods. (cf. Scheler, 2009, p. 35-36)

Scheler also outlined the aprioristic structure of the realm of values. First, all values are either positive or negative. Second, the whole realm of values is graded in an order of ranks in terms of where the values stand to one another in the relation of "higher" and "lower". The gradation of this rank is self-existent and absolute, as it is immutable and entirely independent of the historically or individually variable notions of the gradation. And this order of rank is knowable to us in particular acts of feeling without any support on the part of the intellect. According to Scheler, the specific act of feeling by which the rank of a value is revealed is the act of preferring. Preferring refers to the felt relationships among values: it is an immediate feeling of the relationships prevailing among values. Scheler ascertains two different extant orders of rank of values. He first places values in accordance with their carriers; for example, personal values may have a higher rank than values for which goods are the carriers. The second order is the "modes" of values, where the lower value is founded upon the higher one, which means the higher value is the axiological condition of the lower one. The values order of rank of modes,

from the lowest to the highest, are: (1) the values revealed by sensory feelings, e.g. pleasure; (2) the values revealed by the class of vital feelings (utility), e.g. the feelings of health and sickness, courage, anxiety etc.; (3) the class of spiritual values (e.g. beauty, goodness) and (4) the values of the holy and the unholy (the sacred and the profane). We all recognize, for example, that collegiality and camaraderie are lower values than friendship – in communities where these values function at all.

Scheler does not include the moral values "good" and "evil" into his four classes of modes of values. He states that good and evil are material values, and as such are objects of intentional feelings. The will to realize a positive value is a good whereas the will to realize a negative one an evil. Therefore, the possible moral value of the volition depends upon the insight into the values that are carried by the choice, and upon the insight into their order of preference. Based on this doctrine, Scheler further argues that any ought-to-be (or ideal object) is founded upon a value. As we have said, values are, to Scheler, ideal objects and are entirely independent of the real existence of their carriers. But, the ideal ought-to-be generates an obligation (ought-to-do), which refers to a potential volition aiming at the realization of the ideal value content. To Scheler, a person is a unique type of being who experiences a priori values of an object and also subjectively ranks the values in a situated condition. Moreover, the person is an absolute value, the concrete unity of acts of different types and natures. Because personhood is present in each and every act, she constitutes the whole of these actions and consequently is morally accountable for them. Finally, values themselves have no power, for Scheler, to realize themselves in action. However human beings may sense themselves to be "called" to realize positive values possible in some situation. People may even sense that they are called upon to realize some values that are "good for them alone." Scheler states that the call to realize positive values is not based upon a universal rule or duty or achieving a certain consequence, but may instead be radically individual, or rather unique (based on a value perceived by that individual alone). But Scheler notes, such realization of a "good for me" should be done in conformity to universal moral rules (do not kill, etc.). Scheler calls this experience as a process of individualization, value personalism. An example is given in Kelly (2011, p.116) "[W]hen I experience emotionally the kindness of some action of a person towards some other person or other sentient creature, the moral value of kindness is given to me, and I respond to it in a specific act of affirmation. Similarly, once I grasp the validity of a demonstration of a theorem in mathematics, I naturally respond not only with intellectual assent, but also with a determination to use the theorem with confidence as a premise in further demonstrations".

Finally, Kelly (2011) claimed that material values, fundamental to phenomenological axiology, bring both concrete and synthetic understandings of human values. It offers a systematic means towards a personal response to the Socratic question: how should we live? "there are many ***incompatible ways*** of living successfully and happily, but they are all (should be) founded in the ***right knowledge of the values*** themselves" (Authors' emphasis).

Technological Mediation Theory

Another analysis of the interface between human being and the technology we generate, which offers insight to our present efforts to explore the question of AI and values, was advanced by Don Ihde (1990). He argues in criticism of the classical phenomenological analyses of technology by Martin Heidegger is too abstract and alienates technology from human use. Heidegger is not aiming at practical use but at the Sinn des Seins that hold sway in eras characterized by the dominance of technology and that disrupts our ability to let things appear as the things they are. This means that this classical theory does not pay sufficient attention to the actual experiences people have of the roles of technologies in human existence. To address this concern, he develops the technological mediation theory which emphasizes that technology mediates human experiences and perceptions with the lifeworld. This theory investigates technology in terms of the relations between human beings and technological artifacts with the focus on interpreting the different ways that technologies shape relations between human beings and the world (environment). This mediation theory regards technologies as the mediators of human experiences and practices rather than merely as functional and instrumental "objects".

Ihde (1990) has distinguished four types of relations in his analysis of human-technology relations. First, technologies can be embodied by the users (embodiment relation), such as the glasses worn to see better. Second, they can be terminus of our experience (hermeneutic relation), for example, we can buy a bus ticket from a ticket machine. Third, technologies can give a representation of reality (alterity relation), for instance, a thermometer presents a number of temperatures without producing the reality of heat or cold; and fourth, technologies can play a role at the background of our experience, creating a context for our perceptions, such as public video surveillance systems installed in many big cities.

Multistability is one of the key concepts in post-phenomenological research. As Ihde puts it, "no technology is 'one thing,' nor is it incapable of belonging to multiple contexts" (1999, 47). In technological contexts, multistability means that the "same" technology can have multiple instantiations in history or across cultures. The concept also means that a technology can be put to multiple purposes and can be meaningful in different ways to different users. When we apply the notion of multistability to the analysis of human-technology relations, technology can be understood to potentially support multiple embodiment relations or hermeneutic relations (or other relations). It is remarkable how the systems functioning in living organisms have been altered and adapted by the evolutionary process to function in new ways in different organisms at different temporal points. Technology developed for some purpose in the functioning of routines may similarly be repurposed as the complexity of AI grows.

Given that fact, this multistability of technologies makes it nearly impossible for designers, to a great extent, to predict the ways in which given technologies will influence human actions and to evaluate this influence in a value system. Because of the multistability, designers are not able to maintain an equivocal relationship between their activities and the mediating role of the technologies they are designing; alternatively, the technological mediations emerge in a complex interplay between technologies and their users. The identity of technologies is not fixed, technologies are defined in their context of use and are always "interpreted" and "appropriated" by their users. Verbeek (2011) explains the figure that shows the mediated human actions and interpretations: " (1) the agency of the human being performing the action or making the decision, in interaction with the technology and appropriating the technological artefact in a specific way: (2) the agency of the designer, who, either unintentionally or in deliberate delegations, give a shape to the technology and thus helps to shape its eventual mediating roles; and (3) the agency of the technology mediating human actions and decisions, sometimes in unforeseen ways" (p.99). To handle the complexity of technological mediation, designers should make a connection between the context of design and the context of use with the aim to not only formulate technical features, such as technical objects, affordances and symbolic expressions (e.g. Markus and Silver, 2008), but also to obtain an informed prediction of the technology's future mediating roles.

5 Phenomenological Reflections on the AI Alignment Principle

In the following, we reflect on the AI value alignment principle from these two theories.

Understanding the AI alignment principle through the lens of theory of material values

Firstly, most of us accept the idea that "values are diversified, culturally and socially rooted." Scheler was very insistent on this point, telling us that values function differently in different societies depending upon the "real factors" that are present in each – the way each earns its living, the political organization of the people, the presence or absence of strong family structures, and the like. These determine the different "ideal factors" that function in their ethos, their art, and their religion. How can computer intelligence-AI- be programmed so as also to respond to such real factors? We assume that after the "singularity, AI systems will still not be members of communities that are embedded in a peculiar geographical region, have to earn a living or have a history and a tradition that expresses their values. The danger is that since such sensitivity to peculiar geographical, social, and environmental milieu may be impossible to build in future computers, or any AI systems, their masters (for example, big tech companies such as Google, IBM, Amazon, Facebook) will choose, perhaps not entirely arbitrarily, an ethos that attempts to express a "common denominator" of all cultures, perhaps that of enlightenment liberalism. This could lead to a dangerous simplification of our intellectual and moral environment. This

reflection leads us to conclude that AI, regardless of its general/super intelligence, will not possess the sensitivity of humans, and therefore its "artificial" values, perhaps created by several dominating computer companies, will lose significantly the alignment with the richness of human values.

Secondly, since computers are not sentient creatures, how can they have preferences that emerge from a sensitive "feeling" of values? Although AI may be conscious, their values are still artificial, algorithmic and not equal to human feelings. Furthermore, would these AI systems value human persons, their own personhood (if they should develop them) and if so, in what way? According to Scheler, as we note, the highest non-personal value is that of the sacred. Will AI systems, after the singularity, have some sort of sense of the transcendental or the holy, or will they be entirely secular in their Weltanschauung? Scheler suggests that our feeling of any values has, as a necessary and sufficient condition, the sense that there is absolute being. This reflection further points out that an AI system, even after the singularity, has "artificial" values that are algorithmic. An AI system will not be able to feel these artificial values is the same way as humans feel human values. This probability further challenges the alignment of AI "artificial value" with human values, as well as question humans feeling and realization of any "artificial values" that may be possessed by any AI systems in human actions.

No doubt some work has been done in recent years about the emotional capacities of computers, Rosalind Picard, head of the Affective Computing Research Group at MIT has explored such issues (2010). At last check, she has not managed to create a computer that feels values carried by things, but she is able to program computers to recognize emotions in human faces. Could it be that such a computer could learn to read values carried by a face (its ugliness, perhaps), even moral values carried by actions? But if a machine can be taught to recognize the values on things, it might be able also to prioritize them according to Scheler's scale of values and apply a set of axioms he borrowed from Brentano to the making of decisions. The set is as follows. It is clear that at least the first ground of these "axioms" can be thought of as algorithms, and could be reformulated as commands:

I.

1. The existence of a positive value is itself a positive value.
2. The nonexistence of a positive value is itself a negative value.
3. The existence of a negative value is itself a negative value.
4. The nonexistence of a negative value is itself a positive value.

II.

1. Good is the value in the sphere of will that is attached to the realization of a positive value.
2. Evil is the value in the sphere of will that is attached to the realization of a negative value.
3. Good is the value that in the sphere of will is attached to the realization of a higher (or the highest) value.
4. Evil is the value that in the sphere of will is attached to the realization of a lower (or the lowest) value.

III.

The criterion of 'good' (and 'evil') consists in this sphere in the agreement (disagreement) of the value intended in the realization with the preferred value, or in the disagreement (in the agreement) with the value not preferred" (cited in *Formalism*, 26). Of course, in any such ranking of values as Scheler and Hartmann attempt the value of the human being and of the human person must be counted as the highest value. The preservation of human life and functions must be higher in value than, say, the preservation of a great painting or a sacred vessel. This codable rule might have a stronger positive effect on AI than having AI machines study actual human behavior or human conversation as in the case of Tay, below.

Finally, we see today that AI systems are being used not only to provide functions that contribute to human well-being, but also to spread disinformation, to undermine democratic processes or to demolish the capacity of their armies for war, or at least for preserving the nation. By describing phenomenological the values and disvalues that function in human communities, can we prevent the misuse of the sophisticated AI systems of the future for such purposes? Knowing what is objectively valuable, while inspiring many of us to find our basic human values on that knowledge and to attempt to bring objects, actions or events that are more valuable than those that currently exist and to destroy

what has a lesser value than what could exist in its place. This reflection brings significant implications for the AI alignment principle, which is to align designers' values and users' values. We believe this is the alignment we should foster and build as the necessary conditions for beneficial AI development. Designers and users should strive to establish an inclusive value consensus and thereby both be responsible for AI development. And this responsibility should not be only taken by designers and scientists, but the users should contribute. The phenomena of misuse of computer systems are naturally made by users, however, designers should reflect upon these events and integrate human values better into the system, which use AI systems as the mediators to transmit "good" values to users. This can result in some degree of value alignments between designers and users. This alignment has emphasized by Russell (2017), since AI will learn about human behaviors and infer our preferences, if we behave badly, then the AI will become evil too.

The Case of Tay.

The incident of Microsoft chatbot Tay has told us the "real" story. Created in collaboration between Microsoft's Technology and Research team and its Bing team, the Tay chatbot was an "experiment with and conduct research on conversational understanding" targeted at 18 to 24 years-old in the US. Microsoft claims Tay's conversational abilities were built with relevant, publicly available data that had been anonymized and filtered as its primary source, which was then combined with input from editorial staff, "including improvisational comedians." The bot was supposed to learn and improve as it talked to people, so theoretically it would both become more natural and better at understanding input over time. Microsoft also made Tay able to respond to a handful of specific requests beyond straightforward chatting. It could tell jokes and stories, make memes out of photos, deliver a horoscope, and also play an emoji guessing game. However, in its launch, Tay's conversation quickly devolved to racist, inflammatory, and political statements. Within 18 hours, Microsoft disconnected the bot withdrew it from the market. The event, Microsoft president Brad Smith writes, provided a lesson "not just about cross-cultural norms but about the need for stronger AI safeguards" (Hamilton 2019). Tay's Twitter conversations reinforced the so-called Godwin's law – that as an online discussion goes on, the probability of a comparison involving the Nazis or Hitler grows – with Tay having been encouraged to repeat variations on "Hitler was right" as well as "9/11 was an inside job".

The case has given strong evidence about how AI may be used in a context of use which failed to behave "well". The learning algorithm embedded in Tay functioned effectively to facilitate conversations among humans. Tay could learn about human conversations and infer our preferences. However, the users may not know the importance of "teaching" Tay to behave well, or they intended to bring the "bad" conversations to Tay's algorithmic mind which led the whole experiment to completely fail. AI4People (Floridi et al., 2018), forecasts that, in the coming AI era, more and more people will become the "end users" of AI. Therefore, in addition to the ethical principles that AI researchers or developers have to follow, there will be need to study and incorporate vast amounts of "end-users'' values into the design and development of AI.

The Case of Alpha Go.

Returning to the example of Alpha Go, it was recently announced that the human Go champion the system defeated, Lee Se-dol, was retiring from gameplay. It would appear that a value conflict was the proximate cause. While Alpha Go was trained to learn, compute, and move, it did not value the game itself. In contrapoise, Lee stated in interviews that "Even if I become the number one, there is an entity that cannot be defeated" (Yonhap, 2019). He had lost his love for the game—the values of play, for exploration, for pride— and could not work through those feelings of loss. One can conjecture that, had Alpha Go been programmed with the value of sportsmanship rather than solely with that of a competition where the only aim is to win, there could have been a collaboration of sorts between it and the Go master Lee. Instead we are left with the possibility that a game enjoyed for millennia could fall by the wayside, for if winning is not an option, what is the point of playing?

Understanding the AI alignment principle through the lens of technological mediation

We have strongly argued, through the lens of material values, that this simplified AI "artificial" values that are prioritized in AI's algorithmic mind cannot possibly be aligned with the richness of human values. The alignment can be fostered between designers and users through technological mediation.

Through technological mediation, we can interpret AI alignment principles through the relations that are shaped by AI systems between users and their environment. AI systems are mediators and can be used to bridge human practices and experiences. Since human values will be disclosed and felt in these practices and experiences, then certain degree of alignments can be achieved. Verbeek (2011) discusses new technology and human relations with regard to AI, for example, cyborgs. In this relation, technologies merge with the human body instead of merely being embodied, for example, artificial heart valves and pacemakers to support human's heartbeat. This cyborg association produces a new entity which physically alters the human. This human-technology relation is "bionic" beings-half –organic and half-technological. However, this entity is still literally a "human" which means human can dominate the half of "AI" to perform good and reach the human's preferences. However, given the emergence of new human enhancement, AI technology will make possible a new relation between human and AI, i.e. a full cyborg relation.

The Case of Peter 2.0.

Very recently, Dr. Peter B. Scott-Morgan informed the world that he is transferring himself to Peter 2.0, a full cyborg. He made this decision because of his chronic neuromuscular diseases and the likelihood his muscles will lose the power to function. He wants his life to be enhanced so that "he" can live longer in the world. This type of human enhancement technology goes beyond the medical treatment of diseases, but rather is an early attempt to optimize instead human beings' physical, cognitive, and psychological abilities (Pariseau-Legault et al., 2018). In this case of Peter 2.0, computer intelligence has dominated the organic-human beings to respond to the real factors where human values can function in a human world. This new "full" cyborg relation has raised the AI value alignment to a new moral level that makes possible the questioning of human existence, the meaning of life, and our being-in-the-world.

6 Discussions

A Common Language for Understanding AI Alignment with Human Values

Phenomenological reflection upon the AI alignment principle has resulted in the following principle that the IS community can use as a first step. First, material values refer to the a priori content of values. Knowledge of this content is given in human cognitive feelings and functions in actions where values may be realized. Since there are incompatible ways of feeling and realizing the same human values in a given situation, we must acknowledge the differences among humans in how values function in their world views, and appreciate these differences as enhancing the richness of human openness to values. Instead of arguing that there is no common understanding of values, we need to turn our attention to building a value consensus or an inclusive value concept that we can increase the likelihood of AI value alignment with humans. Second, AI systems may possess simplified "artificial" values (nonpersonal values) in their algorithmic minds which are created by dominant technology companies, which will not be able to align with the richness of human values and will not be able to fully comprehend the complexity of human nature. Instead, we are in danger of making biased, discriminatory and "evil" AIs which can't serve human needs inclusively. Third, AI will shape more and more emerging human experiences and practices. AI will mediate and form new relations with humans which challenge established value framework. How AI aligns with human values will be influenced by these emerging relations among peoples and human experience and practices. Following the phenomenological approach, the alignment itself revealed in these relations will be diversified and situation-dependent.

Fourth, due to the complexity of the multistability of technologies, the values in design and values in use (AI systems in use) will be not seamlessly transmitted from designers and users. However, misuse of AI systems is not evitable. The value alignment between designers and users is more important/critical than AI "artificial" value alignment with human values. In addition to building ethical code for designers

(responsible/ethical design), users must be re-educated and trained to have "good" appropriation of AI systems to ensure value alignments between these two groups. More importantly, an inclusive and common value consensus should be made that can be shared across cultures and societies. Fifth, human "intuitions" of reality are constituted by technology. Technology mediates our existence and experiences and the advance of AI systems will co-constitute a "new" reality that will be studied by scientists in different disciplines. The role of AI systems will have greater impact upon how we shape our ways of access to reality. This will influence the clarity of human beings' cognition of values, their preferences, and their determination to act based upon knowledge of values. Finally, the safe and beneficial AI codes (e.g. Floridi et al., 2018) [AI4People] should be shared with the "vast amount" of users to educate and re-skill users to behave "well" in the age of AI. AI systems will not be able to embrace "multistability" when shaping and mediating human relations to the world. As Russell (2017) has emphasized, in the process, we "must learn to be better persons". This can lead to an "effective symbiosis" between human and AI systems (Aleksander, 2017).

Implications for IS Research

As AI and its abilities increase, it is important that IS researchers anticipate of public sentiment and become able to provide proactive comments about the current state-of-the-art, along with solutions for future systems. When people read that AlphaGo has surpassed all human knowledge and skill—skills that took millennia to acquire—they necessarily become concerned (Silver et al., 2017). When they read that AI is not necessarily a prejudice-free, benevolent tool but that it can repeat and even exacerbate current problems, they quite rightly become concerned (Hills et al., 2019; Metz, 2019) This topic is growing within IS research. Baskerville et al. (2019) has proposed that our research field begin to design systems based on a "digital first" ethos. They state that:

"the classical view of an information system is that it represents and reflects physical reality". We suggest this classical view is increasingly obsolete: digital technologies are now creating and shaping physical reality… (sic) this ontological reversal is where the digital version is created first, and the physical version second (if needed).

Baskerville et al.'s (2019) work speaks directly to the singularity, a time where computers will have surpassed human abilities; they will be beyond merely calculating mathematical proofs but will actually possess human traits. Baskerville et al. (2019) state that the time is rapidly approaching—if not already here—where engineering will lose pride of place and be replaced by concerns more pertinent to the growth of the digital world first. The discussions of the emergence of digital reality in comparison to physical reality pose new challenge for IS community to understand the impact of AI on human values (e.g. freedom and autonomy) (Baskerville et al., 2019).

Contributions and Limitations

In this paper, we have argued that phenomenological theories, material values and technological mediation, bring new interpretations and understanding of the AI alignment principle. Thus, the paper contributes unique knowledge of material values in relation with AI alignment principles that are not addressed by the current ethical guidelines for AI. Upon that foundation, we propose a common principle that we think the IS community can use as the beginning step. Impactful studies on human values should be promoted in the coming AI age in order to ensure AI as a "helpful" technology instead of being "detrimental". Because the inherent complexity and the lack of a model of human values in current AI research, we are in the great challenges to achieve the AI alignment principles. Since the main purpose of the paper is to discuss the phenomenological understanding of human values and AI alignment (relations) with human, we pay exclusive attention to "human values". We deliberately avoid the discussions of "ethics" and "morality" though the theories of material values and technological mediation are originally developed to analysis ethics and moralizing technologies. We argue that the understanding of human values is the first fundamental step for further developing AI related ethics and morality. Future research can departure from our results and make more comprehensive understanding of human values from other ethical theories. As Sotala and Yamposkiy (2017) recommend that "studies which aim to uncover the roots of human morals and preferences also seem like candidates for research

that would benefit the development of safe AI, as do studies into computational models of ethical reasoning" (p.71).

Phenomenology provides a philosophical rationale for studying human experiences in its own terms, thus the results are not normative that can be measured across different contexts. Scheler ascertains the values order of rank of modes without formulating a list of "universal" human values. Milton Rokeach (1973) has articulated a list of universal values, and that list has been revised and used currently in psychology (e.g. Schwartz, 1994 Schwartz et al., 2012). Future research can combine the human values theory from Rokeach (1973) and Schwartz (1994) with the aim to build a value consensus (e.g. Schwartz and Sagie, 2000) and an inclusive value concept that can be shared by humankind. The IS community has built rich knowledge base regarding users' and organizations' behaviors towards technologies. This knowledge base can be extracted and analyzed to identify the fundamental human values that guide these behaviors. These may help us to understand more of human values explicitly, especially, how human values may guide the future development of AI systems. Further phenomenological inquiry can contribute to establishing the material content of values such as those on the list assembled by Rokeach (1973).

7 Conclusion

The rise of AI requires human values to be embodied into, or to become an inseparable function that can influence, how AI learn and behavior and fulfil human objectives and values. The right knowledge of the values that are felt and exist in each person's emotional consciousness should the only reliable value "codes" that we should input to AI algorithms. The theories of material value and technological mediation provide a common language grounded in philosophical traditions that allows it to access thousands of years of thought. In utilizing these theories our community can discuss and clarify the sociotechnical issues that arise as part of the age of AI and the singularity. Such an eventuality depends on the ability of system designers, programmers, researchers, and users to speak a common language, a language that considers the values of human beings. A computer devoid of human values will never be able to become the singularity. However, as of yet, no framework for a unified discussion has been proposed until the present work. The need for a systematic approach for working with artificial intelligence (AI) is current and growing. It is important that IS researchers get ahead of public sentiment and be able to provide proactive comments about the current state-of-the-art, along with solutions for future systems. We propose to build a common language of human values to understand AI through the lens of phenomenological theories. Both academia and business peoples are striving to find solutions to achieve the AI value alignment principle (e.g. Callaghan et al., 2017). This paper contributes a fruitful thought for understanding this principle.

Appendix: The Start of the Art: IS Research on AI and Human Values

We followed Lowry et al. (2004) recommendation to select the journals and articles. In order to include the most relevant articles for the review, we searched databases such as Web of Knowledge, INFORM, Science Direct, Google Scholar, and Emerald Insight. In order to secure a profound literature search, we separately also searched all the 13 journals. We searched for articles which were published between January 2000 to November 2019 in English. For the inclusion of an article in our search database, we did not account for the number of article's citations per year. The initial database search retrieved 151 articles which could potentially be added in our review sample. In the next step, we excluded 138 articles that employ ethical issues from general perspectives, ethics in Information Technology or discussed the human value from the standpoint of AI. The most frequent reason for excluding an article from the review was that, although drawing to some extend on AI, the article did not use AI in the context of human value primarily. This selection process identified 13 AI articles in 13 journals, and we downloaded the full-text (See Table 1). After reviewing those 13 articles, it has become clear that none

of the articles discussed or approached artificial intelligence AND human value similar to our current approach. Nonetheless, the 13 articles are briefly presented in Table 1.

Table 1. Related AI research in IS (2000-2019)

Title	Description	Source	Reference
Partners of humans: a realistic assessment of the role of robots in the foreseeable future	Robots performing human-like tasks depending on success of IT in the area of AI	Journal of Information Technology	Aleksander (2017)
Special section introduction—ubiquitous IT and digital vulnerabilities	IT creates digital vulnerabilities in four areas that are, or could become, significant societal problems with implications at multiple levels of analysis: Online harassment and incivility, technology-driven economic inequality, industrial Internet of Things, and algorithmic ethics and bias	Information Systems Research	Ransbotham et al. (2016).
Are Users Threatened by Credibility Assessment Systems?	As more and more systems increase integration of AI and inadvertently assail the expertise and abilities of users, threat and self-evaluative concerns will become an impediment to technology acceptance	Journal of Management Information Systems	Elkins et al. (2013)
Artificial immune systems for the detection of credit card fraud: an architecture, prototype and preliminary results	Researchers have been studying techniques such as neural networks and genetic algorithms for computational intelligence and their applications to such complex problems. The problem of security management is one of the major concerns in the development of e-Business services and networks	Information Systems Journal	Wong et al. (2012).
Incorporating software agents into supply chains: Experimental investigation with a procurement task	Using of intelligent software agents in enterprise supply chains. Software agents combine and integrate capabilities of several IT classes in a novel manner that enables SCM and decision making in modes not supported previously by IT and not reported in IS	MIS Quarterly	Nissen and Sengupta (2006)
Advances in intelligent information technology: re-branding or progress towards conscious machines?	There is a growing concern on using AI to model the human brain and the possibility of designing systems with the brain's ability to create conscious thought. As AI developed, progress was achieved by overcoming unforeseen difficulties in the pursuit of very ambitious targets, not just a re-branding of promises. This process not only advanced AI but also fed into the mainstream of computing that underpins the IT	Journal of Information Technology	Aleksander (2004)
A conceptual model of an interorganizational intelligent meeting-scheduler (IIMS)	Not relevant	The Journal of Strategic Information Systems	Glezer (2003)
Profiling Web usage in the workplace: A behavior-based AI approach	The use of a behavior-based artificial intelligence system to profile employee Web usage behavior	Journal of Management Information Systems	Anandarajan (2002)
Training for crisis decision-making: Psychological issues and computer-based solutions	AI in keywords, but no discussion on AI. Not relevant	Journal of Management Information Systems	Sniezek et al. (2002)
The role of AI-based technology in support of the KM value activity cycle	The study illustrates both the potential and the limitations of AI technologies in terms of their capability to support the KM process	The Journal of Strategic Information Systems	Fowler (2000)
A new weighted pathfinding algorithm to reduce the search time on grid maps	Artificial Intelligence (AI) techniques are utilized widely in the field of Expert Systems (ES) - as applied to robotics, video games self-driving vehicles. AI techniques which are used in Expert System as decision making functions for the purpose of solving problems that would otherwise require human competence or expertise	Expert Systems with Applications	Algfoor et al. (2017)
Echoic log-surprise: A multi-scale scheme for acoustic saliency detection	For AI systems it is advantageous to mimic some of the traditional algorithms and mechanisms such as visual saliency algorithms. This kind of algorithms have been successfully employed in tasks such as medical diagnosis, detection of violent scenes, environment understanding made by robots, etc	Expert Systems with Applications	Rodriguez-Hidalgo et al. (2018)

Decision support for ethical problem solving: A multi-agent approach	Concurrently, the artificial intelligence community engineered the multi-agent concept and developed the Belief–Desire–Intention (BDI) model. This research suggests that computer-based support for ethical problem solving can be provided by integrating this knowledge in a dynamic computational model, providing support in the roles of Advisor, Group Facilitator, Interaction Coach, and Forecaster	Decision Support Systems	Robbins and Wallace, (2007)

References

Aleksander, I. (2004). "Advances in intelligent information technology: re-branding or progress towards conscious machines? *Journal of Information Technology*, 19 (1), 21-27.

Aleksander, I. (2017). "Partners of humans: a realistic assessment of the role of robots in the foreseeable future." *Journal of Information Technology*, 32 (1), 1-9.

Algfoor, Z. A., M. S. Sunar and A. Abdullah (2017). "A new weighted pathfinding algorithm to reduce the search time on grid maps." *Expert Systems with Applications*, 71, 319-331.

Anandarajan, M. (2002). "Profiling Web usage in the workplace: A behavior-based artificial intelligence approach." *Journal of Management Information Systems*, 19 (1), 243-266.

Baskerville, R., M. Myers and Y. Yoo (2019). "Digital First: The Ontological Reversal and New Challenges for Information Systems Research." *MIS Quarterly* (2019).

Berente, N., B. Gu, J. Recker and R. Santhanam (2019). "*Managing AI*." Call for Papers, *MIS Quarterly.*

Bostrom, N. (2008). Why I want to be a posthuman when I grow up. In Gordijn B. and Chadwick R. (eds) *Medical enhancement and posthumanity* (pp. 107-136). Springer, Dordrecht.

Callaghan, V., J. Miller, R. Yampolskiy and S. Armstrong (2017). *Technological Singularity*. Springer. Conn A. (2017). How Do We Align Artificial Intelligence with Human Values? https://futureoflife.org/2017/02/03/align-artificial-intelligence-with-human-values/. [visited on 03/18/2020].

Elkins, A. C., N. E. Dunbar, B. Adame and J. F. Nunamaker (2013). "Are users threatened by credibility assessment systems?" *Journal of Management Information Systems*, 29 (4), 249-262.

Floridi, L., J. Cowls., M. Beltrametti, R. Chatila., P. Chazerand., V. Dignum and B. Schafer (2018). "AI4People—An ethical framework for a good AI society: opportunities, risks, principles, and recommendations." *Minds and Machines*, 28 (4), 689-707.

Fowler, A. (2000). "The role of AI-based technology in support of the knowledge management value activity cycle." *The Journal of Strategic Information Systems*, 9 (2-3), 107-128.

Friedman, B., P. H. Kahn and A. Borning (2008). "Value sensitive design and information systems," in *The handbook of information and computer ethics*, K. E. Himma, and H. T. Tavani (eds.). Hoboken, N.J.: Wiley, pp. 69-101.

Galanos V. (2019). "Exploring expanding expertise: artificial intelligence as an existential threat and the role of prestigious commentators, 2014–2018." *Technology Analysis & Strategic Management*, 31 (4), 421-432.

Giles, M. (2019). Google researchers have reportedly achieved "quantum supremacy" *Technology Review*, 20 Sept 2019, [visited on 03/18/2020].

Glezer, C. (2003). "A conceptual model of an interorganizational intelligent meeting-scheduler (IIMS)." *The Journal of Strategic Information Systems*, 12 (1), 47-70.

Hamilton, I. A. (2019). Taylor Swift once threatened to sue Microsoft over its chatbot Tay, which Twitter manipulated into a bile-spewing racist, *Business Insider*, 10 Sept 2019, [visited on 3/18/2020].
Herman, A. (2019). The Quantum Computing Threat to American Security, https://www.wsj.com/articles/the-quantum-computing-threat-to-american-security-11573411715 [visited on 3/18/2020].

Hills, G., C. Lau., A. Wright., S. Fuller., M. Bishop., T. Srimani., P. Kanhaiya., R. Ho., A. Amer., Y. Stein., D. Murphy., A. Chandrakasan and M. Shulaker (2019). "Modern microprocessor built from complementary carbon nanotube transistors." *Nature*, (572) 595-602.

Ihde, D. (1990). *Technology and the lifeworld: From garden to earth*. Indiana University Press.

Ihde, D. (1999). *Expanding Hermeneutics: Visualism in Science*. Evanston, IL: Northwestern University Press.

Kelly, E. (1997). "Revisiting Max Scheler's formalism in ethics: virtue-based ethics and moral rules in the non-formal ethics of value." *The Journal of Value Inquiry*, 31 (3), 381-397.

Kelly, E. (2011). *Material Ethics of Value: Max Scheler and Nicolai Hartmann*, Dordrecht, Netherlands: Springer.

Kurzweil, R. (2005). *The singularity is near: When humans transcend biology*. Penguin Group USA.

Lowry, P., D. Romans and A. Curtis (2004). "Global Journal Prestige and Supporting Disciplines: A Scientometric Study of Information Systems Journals." *Journal of the Association for Information Systems*, 5 (2), 29-77.

Markus, M. L. and M. S. Silver (2008). "A foundation for the study of IT effects: A new look at DeSanctis and Poole's concepts of structural features and spirit." *Journal of the Association for Information systems*, 9 (10/11), 609-632.

Melville, N. P. (2010). "Information systems innovation for environmental sustainability," *MIS Quarterly*, 34 (1), 1-21.

Merleau-Ponty, M. (1962). *Phenomenology of Perception*. Translated by Colin Smith. London: Routledge and Kegan Paul.

Metz, C. (2019). 'Nerd,' 'Nonsmoker,' 'Wrongdoer': How Might A.I. Label You? ImageNet Roulette, a digital art project and viral selfie app, exposes how biases have crept into the artificial-intelligence technologies changing our lives, https://www.nytimes.com/2019/09/20/arts/design/imagenet-trevor-paglen-ai-facial-recognition.html [visited on 03/18/2020].

Nissen, M. E and K. Sengupta (2006). "Incorporating software agents into supply chains: Experimental investigation with a procurement task." *MIS Quarterly*, 30 (1), 145-166.

Pariseau-Legault, P., D. Holmes and S. J. Murray (2019). "Understanding human enhancement technologies through critical phenomenology." *Nursing Philosophy*, 20 (1), e12229.

Paulsson, A., S. Han and E. Svee (2019). "A Review of Subjective Values and Their Implications for Green IS Research." In Proceedings of the International Conference on Information Systems (ICIS 2019), Munich, Germany (Available at: https://aisel.aisnet.org/icis2019/sustainable_is/sustainable_is/13/).

Picard, R. (2010). *Affective Computing*, Cambridge, MA: MIT Press.

Ransbotham, S., R. G. Fichman., R. Gopal and A. Gupta (2016). "Special section introduction—ubiquitous IT and digital vulnerabilities." *Information Systems Research*, 27 (4), 834-847.

Robbins, R. W and W. A. Wallace (2007). "Decision support for ethical problem solving: A multi-agent approach." *Decision Support Systems*, 43 (4), 1571-1587.

Rokeach, M. (1973). *The nature of human values*. Free press.

Rodriguez-Hidalgo, A., C. Peláez-Moreno and A. Gallardo-Antolín (2018). "Echoic log-surprise: A multi-scale scheme for acoustic saliency detection." *Expert Systems with Applications*, 114, 255-266.

Rosenberger, R and P. P. Verbeek (2015). "A field guide to post phenomenology." *Post phenomenological investigations: Essays on human-technology relations*, 9-42.

Russell, S., D. Dewey and M. Tegmark (2015). "Research priorities for robust and beneficial artificial intelligence." *AI Magazine*, 34 (4), 105–114.

Russell, S. (2017). 3 Principles for Creating safer AI. TED talk, https://www.ted.com/talks/stuart_russell_3_principles_for_creating_safer_ai [visited on 03/18/2020].

Saltz, J. S and N Dewar (2019). "Data science ethical considerations: a systematic literature review and proposed project framework." *Ethics and Information Technology*, 21 (3), 197-208.

Sarker, S., S. Chatterjee., X. Xiao and A. Elbanna (2019). "The Sociotechnical Axis of Cohesion for the IS Discipline: Its Historical Legacy and its Continued Relevance." *MIS Quarterly*, 43 (3), 695-719.

Scheler, M. (2009). *The Human Place in the Cosmos*, Northwestern U. Press, 2009

Scheler, M. (2012). *Der Formalismus in der Ethik und die materiale Wertethik*. BoD–Books on Demand.

Schutz, A. (1958). Max Scheler's Epistemology and Ethics: II. *The Review of Metaphysics*, pp.486-501.

Schneider, S. (2018). Artificial Intelligence, Consciousness, and Moral Status in *The Routledge Handbook of Neuroethics*, Johnson, L. S. M., & Rommelfanger, K. S. (Eds.). New York: Taylor & Francis.

Schwartz, S. H. (1994). "Are there universal aspects in the structure and contents of human values?" *Journal of social issues*, 50 (4), 19-45.

Schwartz, S. H., J. Cieciuch., M. Vecchione., E. Davidov., R. Fischer., C. Beierlein and, O. Dirilen-Gumus (2012). "Refining the theory of basic individual values." *Journal of personality and social psychology*, 103 (4), 663.

Schwartz, S. H and G. Sagie (2000). "Value consensus and importance: A cross-national study." *Journal of cross-cultural psychology*, 31 (4), 465-497.

Sniezek, J. A., D. C. Wilkins., P. L. Wadlington and M. R. Baumann (2002). "Training for crisis decision-making: Psychological issues and computer-based solutions." *Journal of Management Information Systems*, 18 (4), 147-168.

Silver, D., J. Schrittwieser., K. Simonyan., L. Antonoglou., A. Huang., A. Guez and Y. Chen (2017). "Mastering the game of go without human knowledge." Nature, 550 (7676), 354-359.

Sophia (robot). (2019). https://www.hansonrobotics.com/sophia [visited on 03/18/2020].

Sotala, K and R. Yamposkiy (2017). Responses to the Journey to the Singularity. In *the Technological Singularity*, V. Callaghan et al. (eds.), pp.25-83, The Frontiers Collection, Springer-Verlag GmbH Germany 2017, DOI 10.1007/978-3-662-54033-6_1.

Sutcliffe, A and S. Thew (2014). The design implications of users' values for software and system architecture. In *Economics-Driven Software Architecture* (pp. 297-321). Morgan Kaufmann.

Tegmark, M. (2017). *Life 3.0: Being human in the age of artificial intelligence*. Knopf.

Tuunanen, T and I. T. Kuo (2015). "The effect of culture on requirements: a value-based view of prioritization." *European Journal of Information Systems*, 24 (3), 295-313.

Ulam, S. (1958). "Tribute to John von Neumann." *Bulletin of the American mathematical society*, 64 (3), 1-49.

van der Hoven, J and N. Manders-Huits (2009). "*Value-Sensitive Design*," in A Companion to the Philosophy of Technology, J. K. B. Olsen, S. A. Pedersen, V. F. Hendricks (eds.), John Wiley & Sons, pp. 477-480.

Verbeek, P. P. (2003). "Material hermeneutics." *Techné: Research in Philosophy and Technology*, 6 (3), 181-184.

Verbeek, P. P. (2011). *Moralizing technology: Understanding and designing the morality of things*. University of Chicago Press.

Verbeek, P. P. (2019). *Mediated autonomy-freedom, equality and solidarity in a digital age*. Presentation at Digital Autonomy Conference, Stockholm University, November 7, 2019.

Vohánka, V. (2017). "The Nature and Uniqueness of Material Value-Ethics Clarified." *Ethical Perspectives*, 24 (2), 225-258.

Watson, R. T., M. C. Boudreau and A. J. Chen (2010). "Information systems and environmentally sustainable development: Energy informatics and new directions for the IS community." *MIS Quarterly*, 34 (1), 23-38.

Wong, N., P. Ray., G. Stephens and L. Lewis (2012). "Artificial immune systems for the detection of credit card fraud: an architecture, prototype and preliminary results." *Information Systems Journal*, 22 (1), 53-76.

Yonhap News Agency. (2019). Go master Lee says he quits unable to win over AI Go players, https://en.yna.co.kr/view/AEN20191127004800315 [visited on 03/18/2020].

Yudkowsky, E. (2011). *Complex Value Systems are Required to Realize Valuable Futures*. The Singularity Institute, San Francisco, CA. *http://intelligence.org/files/ComplexValues.pdf*. [visited on 03/18/2020].