Aligning artificial intelligence with human values: reflections from a phenomenological perspective

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

Published in:
AI and Society

DOI:
10.1007/s00146-021-01247-4

Published: 20/07/2021

Document Version
Final published version

Document License
CC BY

AI & SOCIETY
https://doi.org/10.1007/s00146-021-01247-4

ORIGINAL ARTICLE

Aligning artificial intelligence with human values: reflections from a phenomenological perspective

Shengnan Han[1] · **Eugene Kelly**[2] · **Shahrokh Nikou**[1,3] · **Eric-Oluf Svee**[1]

Received: 12 March 2021 / Accepted: 21 June 2021

Abstract
Artificial Intelligence (AI) must be directed at humane ends. The development of AI has produced great uncertainties of ensuring AI alignment with human values (AI value alignment) through AI operations from design to use. For the purposes of addressing this problem, we adopt the phenomenological theories of material values and technological mediation to be that beginning step. In this paper, we first discuss the AI value alignment from the relevant AI studies. Second, we briefly present what are material values and technological mediation and reflect on the AI value alignment through the lenses of these theories. We conclude that a set of finite human values can be defined and adapted to the stable life tasks that AI systems will be called upon to accomplish. The AI value alignment can also be fostered between designers and users through technological mediation. Upon that foundation, we propose a set of common principles to understand the AI value alignment through phenomenological theories. This paper contributes the unique knowledge of phenomenological theories to the discourse on AI alignment with human values.

Keywords Artificial intelligence (AI) · AI value alignment · Human values · Material values · Phenomenology · Technological mediation

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 by means of developing technologies to serve our needs and satisfy our desires. Artificial intelligence (AI) has grown explosively in recent years. Berente et al. (2019) define AI as machines performing cognitive functions that we typically associate with humans, including perceiving, reasoning, learning, and interacting with others. They emphasize that "AI is not confined to one or a few applications, but rather is a pervasive economic, societal, and organizational phenomenon" (p. 1). We have seen the technological advances in AI developments. The results of Alpha Go (Silver et al. 2017) demonstrate the great scientific advances in deep-mind research and provide strong evidence that AI can achieve human level (or above human level) performance without human interventions. Sophia, the social robot provided by Hanson Robotics[1] has travelled around the world and presented her thoughts on AI and on interesting organizational, political, and societal questions. In addition, quantum computing has a multifactor increase in processing speed over conventional computers (Trabesinger 2017). Quantum intelligence algorithms have proven to be more competitive than traditional intelligence algorithms and shown huge potential by

✉ Shengnan Han
shengnan@dsv.su.se

Eugene Kelly
ekelly@nyit.edu

Shahrokh Nikou
shahrokh.nikou@abo.fi

Eric-Oluf Svee
eric-svee@dsv.su.se

[1] Department of Computer and Systems Sciences, Stockholm University, Nodhuset, Borgarfjordsgatan 12, Postbox 7003, 164 07 Kista, Sweden

[2] Department of Social Science, New York Institute of Technology, Northern Boulevard, Old Westbury, NY 11568, USA

[3] Faculty of Social Science, Business and Economics, Åbo Akademi University, Fänriksgatan, 3B, 20500 Turku, Finland

[1] Hanson Robotics' most advanced human-like robot, Sophia, personifies our dreams for the future of AI. https://www.hansonrobotics.com/sophia

Published online: 20 July 2021

simulating quantum computing (Li et al. 2020). Schneider (2018) proposes that AI consciousness may simply go hand-in-hand with sophisticated computation which results in a "singularity" when machine intelligence exceeds the computing power of human brains. Scientists believe that, within our lifetime, machines will obtain the artificial general intelligence (AGI) that can be applied across different domains (e.g., Tegmark 2017). AGI differs from "narrow AI" in that, unlike narrow AI, which focuses on producing programs that display intelligence in a single domain, it focuses on concurrently building a software program that can solve a range of complicated problems in multiple areas and that can operate independently of human interventions and have its own thoughts, anxieties, feelings, strengths, and weaknesses (Pennachin and Goertzel 2007, p. 2). The AGI, however, has been thought to produce a bimodal distribution of results (both negative and positive outcomes) (Worley 2019, p. 226). In this perspective, the potential negative outcomes for humanity, organizations and societal development may be exceedingly undesirable. Christian (2020) tells us of a study of a proprietary software called COMPAS which is used in some US states to estimate the potential recidivism on a scale of 1–10 of persons coming before judges for the purpose of setting bail or granting parole. The system was found to have a pervasive bias against Black suspects or convicts as measured by empirical outcomes. Christian notes that the debate the report caused raised questions "not only about algorithmic risk assessment, but about the very concept of fairness itself. How, exactly, do we define—in statistical and computational terms—the principles, rights, and ideals articulated by the law?" (Christian 2020, pp. 8–9). What is to count in assessing fairly a human person's tendency towards crime or violence? Surely not skin color! Bostrom (2003) has also argued we would be better off focusing more on avoiding negative consequences rather to attaining positive results, even if it means missing out on much that may be of positive value. Given these uncertainties, the exponential growth of AI has been met with confusion and anxiety and yet often also with approbation (e.g., Aleksander 2017; Galanos 2019).

AI should be aimed at making this a better world using its highly optimized mechanistic functions and super intelligence to serve human needs, satisfy human desires and to maximize the realization of human values (e.g., Yudkowsky 2011). This is also proposed as the AI value alignment principles (e.g., Christian 2020; Gabriel 2020; Russell et al. 2015; Russell 2019). One fundamental and critical question is raised and intensively debated: how can we ensure AI alignment with human values through AI operations from design to use? Sotala and Yampolskiy (2017) argue that because of the unresolved disagreements in the disciplines of philosophy and axiology regarding the nature and content of human values, the question of how to align these values in regulating and designing AI, is also moot. We propose here a new way of thinking about the problem and offer a step forward in resolving it.

In this paper, we first provide a review of literature on AI alignment and human values and discuss the AI alignment principles. Second, we briefly present a phenomenology of material values and technological mediation are and reflect on the AI value alignment principles through the lens of these theories. We argue that phenomenology brings a new interpretation and understanding of human values and assists in the construction of the AI alignment principles. Simplified values derived from a phenomenology of material values and their order of relative value, when they are prioritized in AI's algorithmic mind, will not at first, be perfectly aligned with the richness and variety of the human experience of values as they function in the various practices of disparate human communities. Nor do we need to assume that human values exhibit an entirely uniform structure across all of humankind, but they are generally fixed within stable life worlds. Thus, we conclude that a set of finite human values can be defined and adapted to the stable life tasks that AI systems will be called upon to accomplish. The AI value alignment and good human behaviors can also be fostered between designers and users through technological mediation. Upon that foundation, we propose a set of common principles that we think the research community can use as the beginning step. Impactful studies on human values should be promoted in AI research in the coming AI age to ensure that AI is aimed to create a life that we prefer and to make a better world.

2 Literature review

2.1 The AI value alignment principles

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 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). Russell (2017, 2019) has proposed three principles for creating a safe and beneficial AI (i.e., the AI value alignment principles).

(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 humans would "prefer their life to be like."

(2) A law of humility: AI agents are initially uncertain of what human values are, but it may learn those values and preferences by observing how human beings act in stable contexts. 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 capable of expansion as AI progresses. This is the case for humans, where children live and act within a single cultural paradigm altering and extending their knowledge of values as they grow. This simple value-nexus can be probed by phenomenological philosophy.

(3) The ultimate source of information about human preferences is human behavior. To achieve the value alignment between AI and humans, we, in this process, must learn to be better persons, or, perhaps, simpler. The aim should be ensuring the supply to the less fortunate at least the possession of a minimum of the lower but essential value goods such as safety, healthcare, food and shelter, and meaningful work. AI agents can be programmed to make such values primary where and when needed.

To achieve the three AI alignment principles, AI researchers have made advances in our understanding of the nature and content of human values, have proposed technical solutions to the alignment problem and have attempted to insure "good human behaviors" in AI development. In the next section, we present a review of these studies.

2.2 Understandings towards human values and AI alignment

Computer sciences and software engineering research have accumulated some limited knowledge of how to integrate human factors and values in system development. For examples, consumer values (Holbrook 1999), non-economic values (Afuah and Tucci 2000) and internal values (Ilayperuma and Zdravkovic 2010), among others. Kluckhohn's (Kluckhohn 2013) definition of values—that they are a conception, explicit or implicit, distinctive or an individual or characteristic of a group, of the desirable, which influences the selection from available models, and means of action—has been adopted by studies on values in the context of software development (Rescher 1982; Fishbein and Ajzen 2005). The research from Value Sensitive Design (VSD) by the human computer interaction community (Friedman and Hendry 2019) define human values as "what a person or group of people consider important in life", which is in line with what Russell (2017, 2019) has defined, i.e., human values as what humans would "prefer their life to be like." Yudkowsky (2011) argued that we need complex values systems for ensuring positive outcomes of AI research. Sotala (2016) conceptualizes "human values as mental representations that encode the brain's value function (in the reinforcement learning sense) by being imbued with a context-sensitive affective gloss." Sarma and Hay (2017) propose the notion of "mammalian value systems" that define human values and informally distinguish them in three types: (1) mammalian values, (2) human cognition, and (3) several millennia of human social and cultural evolution. The common mammalian values, e.g., seeking and play can be the very rough or approximate initial assumptions of human values that the AI agent can learn and refine its models of human values. The influences of human cognition and the cultural and social effects should be considered in defining the human values. Muehlhauser and Helm (2012) conclude that human values are dynamic, complex and difficult to specify. They recommend an ideal preference theory of value as a promising approach for reaching the AI alignment. This theory is valuable for AI research to make choice modelling by extracting human preferences from human behaviors and human brain activities.

As AI's ability to make independent decisions grows, the most crucial consideration should most likely be a reassessment of safe and responsible AI design. Dignum (2017), for example, asserted that design methodologies that embrace ethical principles and address societal concerns are required to ensure that systems maintain human values. The author concluded that AI must be able to consider social values, moral and ethical issues, balance the relative importance of values held by various stakeholders and in multicultural environments, explain its reasoning, and ensure transparency in all areas of application (Dignum 2017, p. 8). Etzioni and Etzioni (2016b) argued that because AI systems are constantly collecting data, undertaking data mining, and using experience to improve their performance, they may deviate significantly from the standards set forth by their programmers. Therefore, the authors suggested that to develop AI systems that follow human values, they require some sort of oversight, not by our fellow mortals, but by a new kind of AI system which must be provided by AI, or in other words, AI needs to be guided by AI (p. 30). Riedl and Harrison (2016) asserted that value alignment is a property of an intelligent agent. In other words, AI or an artificial agent can only pursue goals that benefit humans, and successful value alignment should ensure that an AI or artificial general intelligence cannot undertake behaviors that harm humans, either purposefully or inadvertently. To alleviate one of the drawbacks of value alignment, the authors stated that system designers and developers should employ the implicit and explicit sociocultural knowledge encoded in stories to create a value-aligned reward signal for reinforcement learning agents (Riedl and Harrison 2016). Armstrong (2019)

assumed that human values can be theoretically considered as equal to human reward function, which is approximately the same as a rational agent's reward function. Consequently, human values and preferences can be synchronized into utility functions which can be adapted to AI design. In a recent study, Hendrycks et al. (2020, p. 8) discovered that aligning machine learning systems with human values is difficult. The authors stated that future systems may indulge in "reward hacking", in which our preferences are only superficially fulfilled, if we do not incorporate all our values into a machine's value function. As such, some authors stated we need to pursue a formal bottom-up approach to value learning (Soares et al. 2015) or take a more empirical approach and use inverse reinforcement learning (Ng and Russell 2000; Russell 2019; Christian 2020).

Notably, the current understanding of human values provides a weak foundation for AI development (Sotala and Yampolskiy 2017). Turchin (2019) also claims that the psychological and behavioral theories of human values are mostly descriptive, informal and underdefined that failed to support AI researchers to define the finite and stable set of human values of which can be applied to ensure the AI alignment. We have seen the progress in formalizing the AI agent value learning abilities; however, according to Etzioni and Etzioni (2016a, p. 155), there is a rising concern regarding how society and individuals can be convinced that AI-enabled instruments will not make unethical decisions. Furthermore, the introduction of AI into an increasing number of instruments makes them considerably smarter, more efficient, and more effective. The authors concluded that in the process, these AI enabled instruments gain some autonomy in the sense that they make numerous decisions on their own, often in direct opposition to the programmers' instructions and guidelines.

In contrast, we argue that the human/AI relation is a reciprocal one, however, where human behavior also mimics the behavior of machines. In the pursuit of more advanced technical solutions for safe and beneficial AI, we need also to address fully the challenge, defined by Heidegger (1954) and others, of determining how technology affects us, its users, in our very being. How can we align AI with human values? and "Which human values should be aligned with the technology we use?" Therefore, AI research needs developing a common principle by which to qualify and understand human values and achieve the alignment principles in the coming AI age. As Gabriel (2020) has argued, however, human preferences, as they are normally understood, may not be capable of providing guidance to an artificial agent in achieving desired outcomes. For preferences in humans are always embedded in a range of values, some of which have nothing to do with what is immediately preferred. It may, therefore, be necessary to consider the more complex and articulated set of values that condition our specific preferences at some given time, as phenomenology has attempted to do.

3 Method

We adopt the phenomenological approach to reflect the AI value alignment principles. Phenomenology is a philosophical approach that seeks to uncover humans' active 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 purposes of this paper. A phenomenological perspective has also been adapted to investigate the desired alignment of AI agents to humans and societies. Scheler's characterization of phenomenology is worth considering: Scheler once wrote that phenomenology, the methodological premises of which are taken as a point of departure for the research undertaken here, is defined only as an attitude and a perspective upon philosophical problems that uniformly informs the spirit of a circle of scholars. "A philosophy based in phenomenology must have as its fundamental character the most living, intensive and immediate experiential intercourse with the world itself—that is, with the objects that are the focus of its concern." (Scheler 1957, p. 380). The witness to that openness to the world and the flexibility of its approach is the work of the great phenomenologists themselves, whose work responded to their own deepening insights into the possibilities of the method and the evolving conditions in the world around them. Unlike linguistic analysis, which is generally tied to a single language, English, phenomenology has adherents beyond Europe itself. It turns itself toward exhibiting the challenges to human values posed by contemporary problems such as ecological balance, global warming, and the possibility of "minds in machines" (a bibliography of works of this kind could be easily assembled). Scheler himself was a careful student of the biological, psychological, and physical theories of his age. Phenomenology is precisely oriented toward the grasping and appropriation of the essential structures of whatever phenomena are brought before it. It is not the lonely occupation of isolated geniuses but, like the best philosophy, requires constant interaction with other thinkers of their own and of foreign cultures.

This alignment between AI and human values requires a balanced relation between the two. We recall that Martin Heidegger's (1954) lecture "Die Frage nach der Technik"

(classical phenomenology) asserts that "the essence of technology is nothing technological." We must grasp instead, he proposes, technology's fundamental impact on what is forgotten about our "being in the world." When technology "holds sway" over us, as it does today, it determines the way we interpret the Being of beings in terms of technology even as we use technology to reveal the world. In addition, 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 falling victim either to the techno-centric construct that humans and society will be dominated and determined by technology, or to the anthropocentric view that technology is merely an instrument and a tool (Verbeek 2011) and does not affect human consciousness of the world. Instead, this mediation theory pays strong attention to the mutual shaping of technology and humanity. 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 mutual shaping of technology and humanity as follows: "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). Therefore, our method seeks out ways in which technology imposes itself upon humankind and humankind imposes values upon the world, while we seek value sets that we believe are deeply hidden beneath technology's values of pragmatic efficiency.

4 A phenomenological perspective upon human values and technological mediation

4.1 Material value ethics

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 2007, 2011). Other phenomenologists such as Edmund Husserl and Nicolai Hartmann developed similar theories of material values during the early phenomenological movement. Schutz (1958) describes the premise of Scheler's phenomenological material value-ethics (MVE) as follows: material or "… concrete values and their hierarchical order form a realm of material, aprioristic data which is disclosed to us by emotional intuition" (p. 486). Stated very simply, phenomenological MVE demonstrates that we can obtain a fairly detailed picture of what human being value if we abstract values from valued objects, as we might abstract the color blue from blue objects and focus upon the colors and their parameters themselves. The result will be a typology of values and disvalues and their relative worth. Values constitute a particular class of ideal objects, like numbers and figures that are objective and immutable. Accordingly, the emotional acts that intend them have epistemic value, and yet are beyond the grasp of the rational intellect alone. That fact does not suggest that our knowledge of values is random or chaotic, for feelings are not without an a priori order. Scheler quotes Pascal's famous observation, "the heart has its reasons." Indeed, it has been shown experimentally that the feelings of simple injustice in very small children are aligned with those of adults (cf. Bloom 2013; McAuliffe et al. 2020).

Furthermore, for Scheler material values are independent of our subjective bodily states and are intended by "pure" (that is, not visceral) emotional acts. That is why our knowledge of values remains the same even as our bodily states may vary chaotically from one moment to the next. For instance, our understanding of the phenomenon of sadness may remain unchanged, while our subjective or visceral emotional state changes from heartbroken to composed, and friendship will remain unchanged as a material value although we suffer when a friend betrays us. Values are independent of things and relationships of all kinds which are their carriers: the so-called "goods." A value such as utility is similarly independent of our having something in use. Both can be thematized 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 reflecting phenomenologically upon the level and intentionality of our emotions, it is possible to discover the realm of concrete values in an aprioristic way without deriving values from visceral feelings that the perception of empirical goods may cause in us. (cf. Scheler 2009, pp. 35–36).

No doubt for Scheler all mental awareness of self and world is made possible through a form of emotional affirmation and receptivity that is summarized for him in the word "love." But values are given in specific emotional acts that re mediated by the structure of human sensibility. That structure is developed in Scheler's *Formalismus in der Ethik*, but it would take us too far afield to present that structure in this paper. It is always operative in any attempt to exhibit the content of material values given in acts of feeling. Perhaps it suffices to point out that the kind and level of feelings that intend the value of courage in an act of self-sacrifice is quite different from the receptive emotionality of hearing a piece of music or the feelings directed at the various values

revealed in the taste of a fine wine. Such values are given in emotional acts, as mathematical objects are given in rational intentional acts and do not exist apart from them. It is just because there is overlap in human sensibility that we can understand the values that function in cultures foreign to our own.

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 rank is disclosed in the phenomenology of the pure emotions of "preferring" and "thinking less of." Preferring refers to felt relationships among sets of values; it is an immediate feeling of the relationships of higher and lower 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 a cognition of the higher one, which means the higher value is the axiological condition of the lower one. The order of rank of values, 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.

Based on this analysis, Scheler further argues that any ought-to-be (or ideal object) is founded upon some specific value or values. One ought (ideally, that is, in the absence of a specific case currently facing us) to be courageous and seek to save a drowning child, for he is a human being whose life possesses intrinsic value; I ought to be generous and give money to a beggar, for he is needy. Courage and generosity are called forth in such instances by the positive value of human life or the negative value of human neediness. We see then, how ideal entities, values, can become functional in human efforts to imagine and to achieve a world worth having. Values may be ideal objects and entirely independent of the real existence of their carriers. However, 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. A person is a unique type of being who perceives intuitively these ideal a priori values that are carried upon possible objects and who also ranks the values as she acts within a situation. Moreover, the person is an absolute value, the concrete unity of intentional acts of different types and natures. Because personhood is present in each and every act, the acting person constitutes the whole of her actions and consequently can be morally accountable for them. Finally, values themselves have no power, as in Plato, to realize themselves in action. However, human beings sense themselves to be "called" to realize positive values possible in some situation. This call—an "ought to do"—does not emanate from a universal Kantian-type Categorical Imperative, nor from a duty to achieve a certain beneficial consequence, as in utilitarianism, but rather from the values themselves that the human agent perceives as the highest for him to realize in this situation. Of course, an agent's behavior must be limited by moral rules: one should not commit murder regardless of the values realized from such an act. Scheler situates this experience of obligation within a process of individualization: laws must be flexible, for they are made for persons, not persons for laws. 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".

To grasp a material value is not the same as having those values function as an a priori within one's own culture's world- and value view in guiding its moral and other evaluative behaviors. We can understand the values functioning in an ancient Athenian's patriotism (e.g., Pericles' Funeral Oration in Thucydides), although those values may not function in the patriotism common today. We can understand what might have driven men a century or two ago to fight duels, though their values of honor and manliness, easily comprehensible to us, hardly function in the ethos of most cultures today. Of course the *functionality* of values evolves, although the essential *content* of the values themselves does not. Human cultures possess enormous diversity, though there is an internal structure of all value systems. In planning the alignment of AI and values by developing descriptions of key values we must always consider how they will function in machines, that is, how the material content of some values may function as an "a priori" to guide the "choices" among possible courses of action in contexts in which these robots will be put to use. It is important to bear in mind that the intentional acts that are aimed a phenomenon are only indirectly relevant to the project of this paper. It is the material content of any value described by phenomenologists that is of concern, not human re-enactments of the emotional acts in which they are given. As yet there is no equivalent of intentionality in AI.

Finally, Kelly (2011) claims that material values, fundamental to phenomenological axiology, bring both concrete and synthetic understanding to 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). Consequently, any moral agent must be capable of understanding the nature of some set of values, of perceiving values as "carried upon" objects, processes or actions, and of having the means for realizing specific valued objects and processes. Human beings are also able to relate their actions to their own specific history and to have the flexibility to order their actions with reference to that personal history. That is the foundation of human integrity.

4.2 Technological mediation theory and its significance for AI

We offer here another analysis of the interface between human being and the technology that gives us insight to our present efforts to explore the question of AI and values. This theory was advanced by Don Ihde (1990), a critic within the aegis of phenomenology. In criticism of Heidegger's (1954) account of technology, he argues that his theory is too abstract and alienates technology from human use. Heidegger (1954) is not aiming at practical use but at the meaning of being that holds sway in eras characterized by the dominance of technology and that disrupts our ability to let things appear as the things they are. Heidegger's (1954) 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 demonstrates how technology mediates human experiences and perceptions with the lifeworld. Technology is analyzed 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 their world (environment). It regards technologies as the mediators of human experiences and practices rather than merely as functional and instrumental "objects".

Ihde (1990) distinguished in his analysis four types of relations between technology and human beings. First, technologies can be embodied by the users (embodiment relation), such as the glasses worn to see better. Second, they can be the terminus of our experience (a hermeneutic relation), for example, we can buy a bus ticket from a ticket machine. Third, technologies can give a representation of reality (an alterity relation), for instance, a thermometer measures 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.

Ihde calls his approach to values "post-phenomenological." His concept of multi-stability is relevant here. For, he notes, "no technology is 'one thing,' nor is it incapable of belonging to multiple contexts" (1999, p. 47), that is, the same technology can have multiple instantiations in history or across cultures, each of which may be stable in each instance. Multi-stability also means that a technology can be put to multiple purposes within multiple constellations of values and thus be relevant and useful in different ways to different users. The concept of multi-stability in human–technology relations functions within multiple embodiment or hermeneutic relations in a given human praxis. It is remarkable how the living systems functioning in organisms have been altered and adapted by the evolutionary process to function in new ways in different organisms at different temporal points. Technology that was developed for specific purposes in the functioning of routines may similarly be repurposed as the complexity of AI grows.

Given that fact, this multi-stability of technologies makes it nearly impossible for designers to anticipate the ways in which given technologies will influence human actions and then to evaluate this influence in a system of values. Who could have predicted, in 1903, to what uses the Wright brothers' invention would be put and how it would transform our lifeworld and even determine the values that function a priori in our consciousness of ourselves in that world? Because of the multi-stability factor, designers are not able to maintain an equivocal relationship between their activities and the mediating role of the technologies they are designing; moreover, the technological mediations emerge in a complex interplay between technologies and their users. Technologies have no fixed identity, for they are defined in the context of their use and are always "interpreted" and "appropriated" by their users. Verbeek (2011) describes in general how the forms of agency that appear through technologically mediated human actions may be interpreted. There is "(1) the agency of the human being performing the action or making the decision to do so in interaction with the technology and appropriating the technological artefact in a specific way; (2) the agency of the designer, who, either unintentionally or [deliberately], give a shape to the technology and thus helps to shape its eventual mediating roles; and (3) the agency of the technology [that] mediate[s] 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 not only to formulate technical features, such as technical artefacts, affordances, and symbolic expressions (e.g., Markus and Silver 2008), but also to obtain at least an informed prediction of the technology's future mediating roles. Consequently, the role of a material value ethics in guiding the alignment of AI with human

values must be as flexible and relative to circumstances as are human values themselves, which may become functional in new ways as conditions evolve and become subject to unexpected injections of new values in a non-closed system of values.

At this point, we may summarize. Technological mediation theory argues that technologies and humans co-constitute the context, feelings and experiences as humans design and use technology. People "feel" the world about them in new ways and thus discover new values or new functions for those that are already recognized among them. Thus, this new material content of values in our lifeworld may generate different "feelings" and experiences of values for different persons, as well the dynamic development of knowledge of values in the process. In this reasoning, we think that Ihde's work paves the way for us to understand the material content of values that are discovered and functionalized by technology and humans in tandem, as humans design technology and use it to (act upon) the world and as the new world contexts they have created act back upon them. AI must situate itself within this process so that we may align it with our developing experience of values.

5 Phenomenological reflections on the AI alignment principles

In the following, we reflect on the AI value alignment principles from these two theories and note key difficulties.

5.1 Understanding the AI alignment principles through the lens of theory of material values

First, we must accept the idea that "values are diversified, culturally and socially rooted." Max 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 in which each community earns its living, the kind of technology used, the political organization of the people, the presence or absence of strong family structures, and the like. Think of how the material value of "motherhood" has come to function in myriad ways in the cultures of the world and how its material content has entered into new configurations with other values, say "women's liberation." And of course, these societies will be constituted in part by their level of technology. These social factors 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 themselves 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. Since such sensitivity to peculiar geographical, social, and environmental milieu may be impossible to build in future computers, we must continually reconstruct the AI-value alignment as we encounter changes in the real factors. Of course, for any AI system, its "masters of technology" (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 and even to dogmatism or moral paralysis. This reflection warns us to be vigilant, for AI, whatever its general/super intelligence, will not possess the sensitivity of humans to values, and therefore, its politically imposed values, perhaps created by several dominating computer companies, will lose significantly their value alignment with human values in all their riches.

Second, since computers are not sentient creatures, how can they have preferences that emerge from a sensitive "feeling" of values? Although AI agents may become in some measurable fashion conscious, their values will still be artificial, algorithmic and not founded upon human-like feelings. Furthermore, would these AI agents value human persons or even their own personhood (if they should develop something analogous to personhood) and if so, in what way? According to Scheler, as we noted, the highest non-personal value is that of the sacred. Will AI agents, after the singularity, have some sort of sense of the transcendental or the holy, or will they be entirely secular in their Weltanschauung? For, again, however, "sensitive" the AI agent may be to shifts in the values in its environment and, however, "intelligent" its willingness to adapt itself to them may be, an AI agent will nonetheless not be able to feel these higher or "spiritual" values, such as beauty, truth, or goodness in the same way as human sensibility does. This probability further challenges the process of aligning the AI agent's "artificial" values with human behavior and desires. If knowledge of values is given only in emotional acts, as material value-ethics holds, an AI agent would have to feel emotions if it were to become independent of its human programmers.

Yet in fact such agents even today seem to be able to emulate or mimic the emotions of men and women, and there has been reflection upon the possible emotional capacities of computers. Rosalind Picard, head of the Affective Computing Research Group at MIT has explored such issues (Picard 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. More recently, work proposed by Höök (2018) has focused on somaesthetic design, whereby individuals are enabled by the use of technological designs to make better sense of their own felt bodily experiences. As opposed to the more instrumental approach common in present day design, where ICT

is designed for embodiment, some aesthetic designs take an approach that focusses more on the corporeal essence of the lived experience of "living in" a human body. Could it be that such a computer could learn to read values carried by a face (its relative ugliness, perhaps), or 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 the scale of values proposed by MVE and apply according to their relative worth a set of axioms he borrowed from Brentano for the making of decisions. The set is as follows. It is clear that at least the first two groups of these "axioms" can be thought of as algorithms, whose variables are material values, and they could be reformulated as commands:

I.

- The existence of a positive value is itself a positive value.
- The nonexistence of a positive value is itself a negative value.
- The existence of a negative value is itself a negative value.
- The nonexistence of a negative value is itself a positive value.

II.

- Good is the value in the sphere of will that is attached to the realization of a positive value.
- Evil is the value in the sphere of will that is attached to the realization of a negative value.
- Good is the value that in the sphere of will is attached to the realization of a higher (or the highest) value.
- 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).

There are challenges to our procedure that must be met. 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 a nation's armies for war or at least for preserving the nation. By describing phenomenologically, the ranked values and disvalues that function in human communities, we can prevent or make difficult the misuse of the sophisticated AI systems of the future for such purposes. For humans, to know reflectively what is objectively valuable and to seek out new knowledge of values, inspires us to base our behavior upon 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. We can judge abstractly which values, in which contexts of human action, can be made to function in ways that foster the general moral health of their persons and community. True, human greed can override such knowledge; we see the right way, yet we choose the wrong way. We wish in general to align our actions with our value knowledge, and just such alignment between values and actions may be easier to create in an AI agent than in us. Designers and users can establish an inclusive and coherent value consensus such as the multifaceted and articulated value-nexus suggested here. The phenomena of misuse of computer systems are naturally made by users; however, designers should reflect upon these events and attempt to integrate higher human values into the system, which uses AI systems as the mediators to transmit "good" values to users. This can result in some degree of value alignment between designers and users. Of course, according to MVE 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. This alignment problem of codable rules over against the study of human behavior by AI agents 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 bad too.

5.2 Understanding the AI alignment principles through the lens of technological mediation

Let us return to through Ihde's concept of technological mediation to reflect further upon how value-alignment may be shaped by the relations of AI systems to their 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 within these practices and experiences, a certain degree of alignment 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 carried by it. For example, artificial heart valves and pacemakers are used to support a human's heartbeat by physically altering the patient. This human–technology relation is "bionic"—half organic and half technological, although the human element dominates the AI agent to satisfy its masters' preferences. Still, AI technology

in this medical context will make possible a new relation between human and AI, i.e., a full cyborg relation. 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. 2019). In this case, computer intelligence dominates the organic—human beings—to realize the real factors so that 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 A set of common principles for creating and evaluating AI alignment with human values

Phenomenological reflection upon the AI alignment principles has resulted in the following principles that the AI 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 this knowledge functions in actions by which higher values may be realized or lower ones eliminated. Since there are incompatible sets of values that become functional in a given praxis, 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, AI researchers need to turn our attention to building a value consensus within the range of competence of the AI agent's operation so that we can increase the likelihood of AI/value alignment with human desires and intentions in that area and minimize conflict between them.

Second, AI systems may possess simplified "artificial" or nonpersonal values in their algorithmic "minds," which are created by dominant technology companies. Such partial systems, if they are to align with the richness of human value, need to be integrated with larger and more encompassing value systems. Otherwise, we are in danger of making biased, or discriminatory AI agents whose routines will not be able to 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 affect and change human praxis and will thus make obsolete established value frameworks. How AI aligns with human values will be influenced by these emerging relations among peoples and human experience and practices. The alignment functioning in these relations will be diversified and situation-dependent.

Fourth, due to the complexity of the multi-stability 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 inevitable. The value alignment between designers and users is more important/critical than AI "artificial" value alignment with human values. In addition to building an ethical code for designers, users must be re-educated and trained to make humane appropriations of AI systems to ensure value alignment between these two groups. More importantly, an inclusive and common value consensus should be made that can be developed by all interested parties and shared across cultures and societies.

Fifth, human reality is constituted on a deep level of the human psyche and is still only in part aligned with the technology we value. Technology nonetheless 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. We may learn from machines to be better than we currently are.

And thus, finally, to make a better world with AI, we "must learn to be better persons"—that is the third AI alignment principle that Russell (2017, 2019) has emphasized. There are three stages characteristic of achieving the conditions of this moral progress. The state, on its various levels, has the responsibility to provide the people with the foundations of a free life: justice, education, and employment, healthcare. The culture of the nation, on the other hand, is the responsibility of the people. Thirdly, we must borrow from the many independent and opposed systems of values that have been developed down through the ages, to which men and women still look to for guidance: the Socratic questioner, the Confucian sage, the Buddhist freedom from craving, the practical American entrepreneur, the pious monk, the submitter to Allah. This seeking for meaning cannot be aided by artificial intelligence, except that AI may liberate more people and allow them to go off on their own thoughtful way, pursuing happiness, where they think that they can find it. In this process, 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 alone be able to embrace "multi-stability" when shaping and mediating human relations to the world. This embrace depends on humans to achieve AI value alignment and to create a better world.

7 Discussion

In this paper, we argue that phenomenological theories, material values and technological mediation offer new interpretations and understanding of the AI alignment principles. We contribute a unique account of material values that are put in alignment with AI principles in a way that has not yet been addressed by the current AI research. Upon that foundation, we propose a common principle that we think the AI community can use as the beginning step. We argue that simplified values, which are prioritized in AI's algorithmic mind, will not at first be well aligned with the human experience of values. The initial alignment should take place on the level of the more practical or utilitarian values. AI will not, we imagine, be contributing to humankind's religion, culture, and philosophy, at least not at first. There are, however, some strides that have been made in fostering the capacities of AGIs for creating interesting works of art and music (Miller 2019). We believe that the alignment can be fostered between designers and users through technological mediation. This is the alignment we should foster as the necessary condition of beneficial AI development. Designers and users should strive to establish an inclusive value consensus and thereby both be responsible for AI development. In addition, this responsibility should not be only taken by designers and scientists, but the users must also contribute to the process. We must learn to become better persons first, then AI can learn and infer our (human) values with the aim to maximize what we prefer, i.e., a model of a better world.

This paper contributes to the literature by proposing common models of material values. The models of the "material" of value predicates, that is, of what fundamental values "contain" of values related to them (e.g., the value of friendship contains such values as intimacy, fellow-feeling, commitment, openness of oneself to the friend), such that Brentano's "laws," listed earlier, can be used to generate actions of AI agents that are preferable to other actions possible in a given context in which choices must be made among possible courses of action, in such wise that the chosen action will be aligned with the values and aims of persons operating in that context. Clearly, this alignment, and also the values on which it operates and the human contexts in which it is applied, must be part of an ongoing project, one guided by the principle outlined in the paper. The way is clear though the achievement is not yet given, for the complex and uncertain technical problems detailed throughout the research upon which this paper is based must still be resolved.

Phenomenology is not the only means of reflection on the contents of values. Rokeach (1973) has also articulated a list of universal material values, and that list has been revised and applied currently in psychology (e.g., Schwartz 1994; Schwartz et al. 2012). Future research can study comparatively the axiology of Scheler, Rokeach (1973) and Schwartz (1994) with the aim of building a value consensus (e.g., Schwartz and Sagie 2000) and an inclusive value concept that can be shared by humankind. The AI community has itself built a rich storehouse 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.

In this context, Walsham (2012, p. 89) argues that what is called critical thinking "involves considering what is wrong with the world, as well as what is right and challenging existing orthodoxies and hierarchies." But what is right and wrong themselves must be subject to phenomenological scrutiny. Here, we believe, the current paper is seeking a new way to align the artificial with the human. For the thrust of the paper is to provide a descriptive account of what is or may be valuable, while at the same time appropriating the phenomenological methodology for describing values and criticizing and revising how they become functional in human practices of governance, production and distribution. This work should be the ongoing task of independent scholars who assess the values that are guiding or ought to be guiding a variety of human practices in contexts that vary from culture to culture. The phenomenological perspective of this paper allows, like all phenomenology, for constant self-correction even at the very foundations of its procedures.

8 Conclusions

The time for this research is come. Baskerville's et al. (2020) 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. They 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 researchers to understand the impact of AI on human values (e.g., freedom and autonomy) (Baskerville et al. 2020). The AI community is faced with a great challenge because of the inherent complexity of human life and the lack of a model of human values in current AI research to achieve the AI alignment principles.

We deliberately avoided here the discussions of "ethics" and "morality," although the theories of material values and technological mediation were originally applied to an

analysis of technologies from an ethical point of view. We argue that the understanding of human values is the first fundamental step for further developing AI-related ethics and morality. Future research can take its point of departure from our results and make more comprehensive understanding of human values from other ethical theories. Sotala and Yampolskiy (2017) also note that empirical "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).

The rise of AI makes it essential that human values become embedded in or inseparable from the functions of the processes in which an AI system learns to make evaluative choices in the safe fulfilment of human objectives and the values that guide their realization. The right knowledge of the values that are felt and come to function in each situated person's emotional consciousness should be the only reliable value "codes" that we should input to AI algorithms. The theories of MVE and technological mediation provide a theoretical understanding grounded in phenomenological philosophical traditions. 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. A computer devoid of human values will never be able to become the singularity. However, as of yet, no principle for a unified discussion has been proposed until the present work. We propose to build a common principle to understand human values and AI alignment problem through the lens of phenomenological theories. Both academia and business are striving to find solutions to achieve the human value alignment with AI (e.g., Callaghan et al. 2017) with the aim of making a better world. This paper contributes a fruitful thought for achieving this aim.

Funding Open access funding provided by Stockholm University.

References

Afuah A, Tucci CL (2000) Internet business models and strategies: text and cases. McGraw-Hill Higher Education, Boston

Aleksander I (2017) Partners of humans: a realistic assessment of the role of robots in the foreseeable future. J Inf Technol 32:1–9

Armstrong S (2019) Research Agenda v0.9: synthesizing a human's preferences into a utility function. Blogpost in LessWrong. Available at: https://www.lesswrong.com/posts/CSEdLLEkap2pubjof/research-agenda-v0-9-synthesising-a-human-s-preferences-into. Accessed 14 July 2021

Baskerville RL, Myers MD, Yo YG (2020) Digital first: the ontological reversal and new challenges for information systems research. MIS Q 44:509–523

Berente N, Gu B, Recker J, Santhanam R (2019) Managing AI. Call for papers. MIS Quarterly, pp 1–5

Bloom P (2013) Just babies: The origins of good and evil. New York: Crown Publishers

Bostrom N (2003) Astronomical waste: the opportunity cost of delayed technological development. Utilitas 15:308–314

Callaghan V, Miller J, Yampolskiy R, Armstrong S (2017) Technological singularity. Springer, New York

Christian B (2020) The alignment problem: machine learning and human values. W. W. Norton & Company, New York

Dignum V (2017) Responsible artificial intelligence: designing AI for human values. ITU J ICT Discov 1:1–8

Etzioni A, Etzioni O (2016a) AI assisted ethics. Ethics Inf Technol 18:149–156

Etzioni A, Etzioni O (2016b) Designing AI systems that obey our laws and values. Commun ACM 59:29–31

Fishbein M, Ajzen I (2005) Theory-based behavior change interventions: comments on Hobbis and Sutton. J Health Psychol 10(1):27–31

Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, Luetge C, Madelin R, Pagallo U, Rossi F, Schafer B (2018) AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Mind Mach 28:689–707

Friedman B, Hendry DG (2019) Value sensitive design: shaping technology with moral imagination. MIT Press, Cambridge

Gabriel I (2020) Artificial intelligence, values, and alignment. Mind Mach 30:411–437

Galanos V (2019) Exploring expanding expertise: artificial intelligence as an existential threat and the role of prestigious commentators, 2014–2018. Technol Anal Strat Manage 31:421–432

Heidegger M (1954) Die Frage nach der Technik", in Vorträge und Aufsätze, Pfullingen: Günther Neske; translated as "The Question concerning Technology", in The Question Concerning Technology and Other Essays, William Lovitt (trans.). New York: Harper and Row, 1977, pp 3–35

Hendrycks D, Burns C, Basar S, Critch A, Li J, Song D, Steinhardt J (2020) Aligning AI with shared human values. arXiv:2008.02275

Holbrook M (1999) Consumer value: a framework for analysis and research. Routledge, London

Höök K (2018) Designing with the body: somaesthetic interaction design. MIT Press, Cambridge

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

Ihde D (1999) Expanding hermeneutics: visualism in science. Northwestern University Press, Evanston, IL

Ilayperuma T, Zdravkovic J (2010) Exploring business value models from the inter-organizational collaboration perspective. In: Proceedings of the 2010 ACM symposium on applied computing (SAC). Sierre, Switzerland, pp 99–105

Kelly E (1997) Revisiting Max Scheler's formalism in ethics: virtue-based ethics and moral rules in the non-formal ethics of value. J Value Inq 31:381–397
Kelly E (2011) Material ethics of value: Max Scheler and Nicolai Hartmann. Springer, Dordrecht
Kluckhohn C (2013) Values and value-orientations in the theory of action: an exploration in definition and classification. Harvard University Press, Cambridge, pp 388–433
Li Y, Tian M, Liu G, Peng C, Jiao L (2020) Quantum optimization and quantum learning: a survey. IEEE Access 8:23568–23593
Markus ML, Silver MS (2008) A foundation for the study of IT effects: a new look at De-Sanctis and Poole's concepts of structural features and spirit. J Assoc Inf Syst 9:609–632
McAuliffe K, Blake PR, Warneken F (2020) Costly fairness in children is influenced by who is watching. Dev Psychol 56:773–782
Merleau-Ponty M (1962) Phenomenology of Perception. Translated by Colin Smith. Routledge and Kegan Paul, London
Miller AI (2019) The artist in the machine: the world of AI-powered creativity. MIT Press, Cambridge
Muehlhauser L, Helm L (2012) Intelligence explosion and machine ethics. In: Eden A, Søraker J, Moor JH, Steinhart E (eds) Singularity hypotheses: a scientific and philosophical assessment. Springer, Berlin
Ng AY, Russell SJ (2000) Algorithms for inverse reinforcement learning. In: Proceedings of the 22nd International Conference on Machine Learning, pp 896–903. https://doi.org/10.1145/1102351.1102464
Pariseau-Legault P, Holmes D, Murray SJ (2019) Understanding human enhancement technologies through critical phenomenology. Nursing Philos 20:e12229
Pennachin C, Goertzel B (2007) Contemporary approaches to artificial general intelligence. In: Pinkal M, Uszkoreit H, Pennachin C (eds) Artificial general intelligence. Springer, Berlin, Heidelberg, pp 1–30
Picard RW (2010) Affective computing. MIT Press, Cambridge
Rescher N (1982) Moral issues relating to the economics of new knowledge in the biomedical sciences. New knowledge in the biomedical sciences. Springer, Dordrecht, pp 35–45
Riedl MO, Harrison B (2016) Using stories to teach human values to artificial agents. In: Proceedings of the 2nd International Workshop on AI. Phoenix, AZ: Ethics and Society
Rokeach M (1973) The nature of human values. Free Press, New York
Rosenberger R, Verbeek PP (2015) A field guide to post phenomenology. Post phenomenological investigations: essays on human-technology relations. Lexington Publishers, London, pp 9–42
Russell S (2017) 3 Principles for creating safer AI. Available at: https://www.ted.com/talks/stuart_russell_how_ai_might_make_us_better_people. Accessed 14 July 2021
Russell S (2019) Human compatible: artificial intelligence and the problem of control. Penguin, New York
Russell S, Dewey D, Tegmark M (2015) Research priorities for robust and beneficial artificial intelligence. AI Mag 34:105–114
Sarma G, Hay N (2017) Mammalian value systems. Informatica 41(3):1–12. https://doi.org/10.2139/ssrn.2975399
Scheler M (1957) Phänomenologie und Erkenntnistheorie. In: Gesammelte Werke Band 10
Scheler M (2009) The human place in the cosmos. Northwestern University Press, Evanston, p 2009
Scheler M (2012) Der Formalismus in der Ethik und die materiale Wertethik. BoD—books on demand
Schneider S (2018) Artificial intelligence, consciousness, and moral status. In: Johnson LSM, Rommelfanger KS (eds) The Routledge hand-book of neuroethics. Taylor & Francis, New York
Schutz A (1958) Max Scheler's epistemology and ethics: II. Rev Metaphys 11(3):486–501
Schwartz SH (1994) Are there universal aspects in the structure and contents of human values? J Soc Issues 50:19–45
Schwartz SH, Sagie G (2000) Value consensus and importance: a cross-national study. J Cross Cult Psychol 31:465–497
Schwartz SH, Cieciuch J, Vecchione M, Davidov E, Fischer R, Beierlein C, Ramos A, Verkasalo M, Lönnqvist JE, Demirutku K, Dirilen-Gumus O (2012) Refining the theory of basic individual values. J Pers Soc Psychol 103:663–688
Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, Hubert T, Baker L, Lai M, Bolton A, Chen Y (2017) Mastering the game of go without human knowledge. Nature 550(7676):354–359
Soares N, Fallenstein B, Armstrong S, Yudkowsky E (2015) Corrigibility. In Artificial Intelligence and Ethics, ed. T. Walsh, AAAI Technical Report WS-15-02. Palo Alto, CA: AAAI Press.
Sotala K (2016) Defining human values for value learners. In: Proceedings of the Workshops of the 30th AAAI Conference on Artificial Intelligence: AI, Ethics, and Society. AAAI Press, Phoenix, pp 113–123
Sotala K, Yampolskiy R (2017) Responses to the journey to the singularity. In: Callaghan V et al (eds) The technological singularity, the frontiers collection. Springer-Verlag GmbH, Germany, pp 25–83
Tegmark M (2017) Life 3.0: Being human in the age of artificial intelligence. Knopf, New York
Trabesinger A (2017) Quantum computing: towards reality. Nature 543(7646):S1
Turchin A (2019) AI alignment problem: "human values" don't actually exist. Available at: https://www.lesswrong.com/posts/ngqvnWGsvTEiTASih/ai-alignment-problem-human-values-don-t-actually-exist. Accessed 14 July 2021
Ulam S (1958) Tribute to John von Neumann. Bull Am Math Soc 64:1–49
Verbeek PP (2003) Material hermeneutics. Tech Res Philos Technol 6:181–184
Verbeek PP (2011) Moralizing technology: understanding and designing the morality of things. University of Chicago Press, Chicago
Walsham G (2012) Are we making a better world with ICTs? Reflections on a future agenda for the IS field. J Inf Technol 27:87–93
Worley GG III (2019) Robustness to fundamental uncertainty in AGI alignment. J Conscious Stud 27:225–241
Yudkowsky E (2011) Complex value systems are required to realize valuable futures. The Singularity Institute, San Francisco, CA. Available at http://intelligence.org/files/ComplexValues.pdf. Accessed 14 July 2021