Shared Spirituality Among Human Persons and Artificially Intelligent Agents

Mark Graves

Abstract

Technical advances in artificial intelligence make somewhat likely the possibility of robotic or software agents exhibiting or extending human-level intelligence within a few decades. Intelligent or superintelligent agents have tremendous cultural and ethical implications as well as raise interesting philosophical and theological questions of personhood. Theological investigation can orient the development of the agent's intelligent communication capacities and moral reasoning abilities to meet significant research goals in artificial intelligence as well as lay a foundation for a shared relational spirituality. Josiah Royce's Loyalty-to-Loyalty and philosophy of community provide a semiotic model of spirituality which can guide the development and functioning of an agent's interpretive processes within a moral framework. A Roycean community valuing mutual understanding through intelligent communication yields an ethical system within which intelligent agents can relate to humans and further develop their understanding of and engagement with the world.

Introduction

If artificial intelligence surpasses human intelligence, will theologians be prepared? Philosophers and futurists Vincent Müller and Nick Bostrom (2014) surveyed 170 artificial intelligence (AI) researchers in 2012-13 on when human-level machine intelligence (HLMI) would occur and found a median year of 2040. Depending upon the capacities of the artificial intelligence software to improve itself, the software's ability to design even more intelligent software might lead to an intelligence explosion through accelerating recursive self-improvement (Hall 2007a). Ray Kurzweil (2005) and others argue the intelligence explosion leads to a *technological singularity* beyond which events become incomprehensible to humans. Kurzweil predicts superintelligence to occur in 2045, though Müller and Bostrom's (2014) respondents expect a slower rate of growth with superintelligent machines only occurring within 30 years following HLMI. Regardless of whether the dramatic occurrence of a singularity comes to pass, the progression of technology to HLMI and to superintelligence needs theological investigation.

Although the research program to develop HLMI (called *strong* artificial intelligence) does not yet have success, AI's "failures" have led to enormous technological advances, including electronic logic circuits in every computer and most modern electronic devices (a failed attempt to model neurons), theoretical models for computers (as Turing machines from Alan Turing's 1948 essay "Intelligent Machinery"), and numerous aspects of high-level programming languages (with some data structures originally envisioned to model human memory structures). In addition, many advances in *weak* artificial intelligence, where computers approach or exceed human-level intelligence in a narrow, non-extensible domain, have demonstrated success approaching or exceeding human capacities, e.g., playing chess, recognizing speech or images, diagnosing disease from medical images, narrow dimensions of learning, and basic translation between natural languages. As strong AI researchers increase our understanding of human intellectual capacities and how to extend them, and weak AI researchers develop a growing number of human-comparable tools, many researchers believe these advances may lead to HLMI, especially when combined with advances in neuroscience, understanding and modeling of human cognition, large-scale datasets, faster computer hardware, and/or computational models of either aspects of brain processing or the entire brain, etc.

Given the pervasiveness of computer technology and weak artificial intelligence (e.g. Siri and other voice recognition personal assistants on a smartphone), theologians may need to respond quickly, even if weak artificial intelligence progresses relatively incrementally over the next century. Even the "normal" exponential growth of computer processors and webpages led to Google indexing a trillion webpages within 15 years of the web's creation (Alpert and Hajaj 2008). No human can process that amount of information. Merely the pervasiveness of technology and the amount of information available will significantly affect human possibilities regardless of the level of intelligence embedded within that technology. In addition, if strong artificial intelligence leads to technological intelligence explosion within a few decades, then it may be imperative for theologians to engage the technology while humans still have influence.

Philosophers, ethicists, and computer scientists have begun considering the ethical implications of strong AI. Strong AI success would most likely lead to machines or software with significant autonomy and ability to engage in at least some aspects of the human world, i.e., *intelligent agents.* Michael and Susan Leigh Anderson (2007; 2011) argue for a new field of machine ethics, which gives machines ethical principles or reasoning methods to enable them to resolve ethical dilemmas and make decisions to act in an ethically responsible manner. Storrs Hall (2007b) describes the need for intelligent agents to have a conscience to make moral decisions and avoid psychopathic behavior and proposes that AI researchers can develop virtuous machines. Wendell Wallach and Colin Allen (2008) examine some of the dangers already in place with intelligent technology and characterize approaches to developing artificial moral agents. Wallach (2015) also argues that recent progress in machine learning makes the likelihood for superintelligent AI greater than researchers previously anticipated and that these advances in machine learning can be used for developing intelligent agents capable of making moral decisions (Bengio, Goodfellow and Courville 2015). Robert Sawyer (2007) identifies a pressing need for ethics in governing robotic behavior, especially in an environment with military drones and robot consumer electronics, and exploration of autonomous lethal weapons has begun (Russell et al. 2015). Roman Yampolskiy (2013) claims that embedding ethics within HLMI or superintelligent agents is misguided because of the uncertainty involved and that AI research needs review boards, such as currently occurs for medical research; though others, such as David Chalmers (2010), argue that one cannot feasibly contain superintelligence. Chalmers (2010) argues for philosophical consideration of superintelligence and the singularity because of their practical implications for humanity, even if there exists only a small chance of them occurring, and raises some of the interesting philosophical questions the possibility of superintelligence raises.

Do intelligent agents need the capacity for moral decision making or not? Under what conditions would intelligent agents be considered persons with moral (and legal) status? Although strong AI poses significant philosophical and ethical issues, there are also theological considerations, such as personhood, moral culpability, and salvation. Examining personhood addresses several challenges to strong AI and a significant factor in ethical considerations. Personhood and ethical ramifications may vary for HLMI and superintelligence. One might limit personhood to subservient HLMI and ascribe moral culpability to their developers, owners, or operators, but superintelligence may require different treatment simply because human operators *cannot* fully anticipate all consequences of its behavior. Taking a more egalitarian approach to the investigation into the personhood of intelligent agents, regardless of their level of intelligence, one can also examine whether these agents could have some type of relational spirituality, and whether that spirituality could be shared with humans. Careful consideration of human relations with HLMI not only impacts our interaction with these intelligent agents in the near future but also creates the initial framework with which superintelligent agents would consider their relationship with us.

Theologians and religious scholars have recently engaged proactively with science, including genetic technology (Peters 1996), neuroscience (Brown, Murphy and Malony 1998), and theological implications of life extension and other dimensions of transhumanism (Peters 2015). Theologians and religious scholars have also considered the topic of AI and set the stage for constructive engagement with AI research. Anne Foerst 1996; 1998) described the challenges in creating dialogue between AI researchers and theologians. Foerst (1999; 2004) also examined the theological implications to personhood in AI research, and Ian Barbour (1999; 2002) considered the relationship between AI and human nature. Noreen Herzfeld (2002b; 2002a) examined the relationship between human desire to create artificial intelligence and *imago Dei*, and Robert Geraci (2010) considered the apocalyptic nature of the singularity.

Jewish traditions also have historic ethical resources to consider intelligent machines. In the Talmud (Sanhedrin 65b), a rabbi created a creature and sent him to another rabbi, but when the second rabbi spoke, the creature did not respond, and the second rabbi returned the creature to dust. The destruction of the creature is approved because it lacked the communication ability essential for human nature. In sixteenth-century Prague, legend holds that a rabbi used his knowledge of Jewish mysticism to animate clay and create a golem to defend the Jewish people. However, the rabbi could not fully control the golem when accidentally leaving it active on the Sabbath, and the rabbi had to return the golem permanently to clay (Kieval 1997; Rappaport 2006).

Within AI research, strong AI attempts to capture the human capacity to engage culture and learn through socially constructed language. AI researchers can more easily model the specialized expertise of a few trained humans, e.g., chess masters, than the basic general-purpose information humans use to understand their world. We learn basic life skills through interacting with our physical and social environment as a young child and then learn to engage culture through formal and peer education. Without that broad embodied and enculturated base, intelligent technologies remain fragile and unable to respond appropriately to the human (cultural-specific) world. AI researchers have not yet discovered how to place culture in a larger framework to create intelligent agents capable of learning and participating in human culture, but theological investigation addresses that weakness of strong AI research. How can humans and machines communicate in a way that makes meaning and adds value and constructive purpose to their relationship?

Intelligent Communication

Communication has the Latin root *commūn(is)*, which means to make common—the same root as communion. In communicating with another person, one attempts to create a shared interpretation of some experience or abstraction. Without socially shared interpretations, language lacks meaning. Characterization of interpretive systems engages much scholarly work in the arts and humanities. Theological insights into the interpretation of human-transcendent relationships (such as through experiences, texts, and traditions) can yield a plausible initial model for interpretation of human-computational (virtual) social relationships if approached using the academic methods of religious scholars, who examine religions and their associated interpretive frameworks across all known historical and contemporary human cultures.

One of the foundational goals driving AI research has been to develop a machine to pass the Turing Test. As Alan Turing (1950) proposed, human judges interview computers and humans through identical electronic means while both the computers and humans attempt to convince the judges of their humanness. If the judges cannot reliably identify the computers as imposter humans, then they consider the computer to demonstrate human-level intelligence. John Searle (1980) argues for the inadequacy of the Turing Test (claiming strong AI will not succeed) because a computer can only use syntactic rules to simulate the appearance of understanding without actually knowing what the words mean. Searle's argument has famously received numerous critiques, but Turing, Searle, and Searle's critics all agree upon the significance of communication in HLMI (Cole 2014).

For humans, a continuation of the shared human interpretive process beyond basic communication involves creating meaning in one's environment (natural and social). Human meaning making includes the full range of human cultural aspirations in a historic context: the creation of art, literature, and technology; moral and ethical judgments; and the striving and commitments often ascribed to spirituality and religion. The making of meaning for one’s self has existential and psychological foundations, which must be taken into account for the development of intelligent technologies to facilitate and extend human drives, passions, and search for meaningful existence. (Kegan 2001)

Developing non-trivial shared interpretations between humans and computers has proven problematic. With user training, humans can learn computational conventions, such as user interface modalities, application commands, and programming languages and use those computational symbols to manipulate computational systems (with occasional subsequent physical effects). With programming and data, computers can manipulate human symbols, such as written text for searching or translation or mathematical symbols, and natural language commands can occasionally have computational effects. However, for humans and computers to share interpretations requires that variations in interpretation affect both human and computer interpretations and their habituated and programmatic behaviors. One useful model for interpretation in theology and in artificial intelligence research occurs in semiotics.

Semiotic Interpretation

The field of *semiotics* generalizes human language to any medium or sense modality and examines the process of meaning formation through an organism’s apprehension of the world through signs. The American philosopher Charles S. Peirce developed semiotics with several organizations of signs. The simplest organization of Peirce’s semiotics consists of three kinds of relations to objects in signs: icon, index, and symbol (Parker 1998, 156-7). An *icon* signifies by resembling its object like a painting or a map. It possesses a quality that resembles or duplicates those of the object. An *index* represents its object through an existential connection between itself and the object. For example, a fingerprint not only resembles the ridges of a fingertip, it signifies the existence of a particular finger. An index may also signify by a causal relationship, such as a thermometer or weather vane. A *symbol* represents its object through a convention that governs how the symbol will be used: A symbol refers to an object by social convention without direct similarity (as in an icon) or existential or causal connection (as in an index). For example, English speakers connect the word *dog* to an animal through conventional English usage, and thus one can categorize human language as symbolic language.[1]

Symbols require interpretation. The social group that defines a language also defines possible ways the words and more complex abstractions such as metaphor might be interpreted. The cognitive linguist George Lakoff (1980) pushes past the obvious metaphors in language and claims that most (if not all) language is metaphoric to some degree unless describing physical reality. These metaphors influence how we interpret meaning in symbolic language. For Lakoff and his philosopher coauthor Mark Johnson, the essence of metaphor is to understand and experience one kind of thing in terms of another. One understands *Argument Is War* in terms of attack, defense, counterattack,

[1] Paul Tillich and Carl Jung define symbol and sign differently from Peirce: for them, a symbol partakes of the reality to which it points and a sign simply points to it.

victory, defeat, strategy, and the like, and an intense verbal battle can even trigger emotional responses appropriate for physical danger. Some metaphors directly connect to one's body—an argument that "makes one's blood boil" does relate to actual body temperature, blood pressure, and heart rate (4-5, 15).

Artificial intelligence researchers began with an early recognition of the value of symbols, but the influence of logical positivism and lack of theoretical and computational resources reduced the symbol to an abstraction disconnected from the natural and human social world. The symbols became *degenerate* and semiotically function like icons or indexes. The interpretations lost their social dimension and metaphoric power and became implemented mechanistically (as programs). Although some AI researchers continue to develop symbol-processing strategies with more realistic interpretive capacities, significant challenges remain (Steels 2007, 2012). Theology can contribute to more intelligent interpretive systems because theologians generally resist reductionistic tendencies since theological symbols often lack physical referents.

Semiotically, a *symbol* is a discrete unit of meaning that stands for some object as a habit of (socially) conventional interpretation. In this context, *habit* is more general than a rote behavior and refers to a general disposition, including of natural systems.[2] Any collection capable of forming conventions for interpretation could create symbols (including intelligent agents).

From C.S. Peirce's semiotics, the classic characterization of a symbol, such as 'bread' indicates three aspects:

- The *sign* consisting of the English word "bread."
- The *object* consisting of a particular piece of bread, and
- The *interpretation* shared among many people including a baker and a partaker, which includes the habit of the partaker to associate that sign with that object and the habit of the baker to associate that sign with the actions of preparing and baking that object.

Interpretation is a very general category capturing the ways that one internalizes and makes meaning of something. In Peirce's semiotic context, the meaning affects the way a person might respond to possible future conditions. As one reads and interprets,

[2] In Aristotle's metaphysics and in Thomistic thought, habit functions as a dynamic principle that perfects the operations and powers of human beings. The eighteenth century Puritan minister and philosophical theologian Jonathan Edwards attempted to redefine the Aristotelian conception of habit to incorporate the demand of seventeenth- and eighteenth-century science: to see the world in terms of power and motion rather than in terms of substance and form. Edwards replaced substance metaphysics with a dynamic and relational conception of being and reinterpreted habit as an active causal power. In the nineteenth century, Peirce saw advances in geology and biology, which described a very old, geologically changing planet and evolving species, as challenging not only assumptions about human history, but also our understanding of reality, and he organized his understanding of reality to account for fundamental change in nature and argued reality requires one to consider the practical effects of what one hopes to understand rather than speculate on abstract essences. Peirce's close friend and benefactor William James' popularized pragmatic philosophy and developed a psychological account of Peirce's metaphysical habits.

that meaning affects how one interprets something in the future. One also interprets one's experiences in general. If the result of interpretation resonates with prior interpretation, one builds or strengthens a habit of interpreting that way. For a dissonant interpretation, one might ignore differences in a slight dissonance or change one's habit for a more significant one. Habits of interpretation may be innate, learned, or socially constructed.

The Austrian philosopher Ludwig Wittgenstein describes language in terms of a game. One must interpret the language in a particular context—a particular group who understand not only the language rules of grammar and word meaning but also how to use the language. Even in a small group of friends, a family, or a research field, aspects of language have complex and particular meanings that differ from those outside that group.

The capacity of a symbol to refer to any object through social construction enables groups of individuals to define interpretations of symbols that may have no natural referents. Numbers, geometric shapes, and mathematical formulas generalize relationships among natural objects to define abstractions that one can manipulate without considering their material referents. In physics, the wave equation characterizes waves of sound, light, water, and gravitation, and the eighteenth-century, mathematical interpretation of the strings of a musical instrument lead to a general way of interpreting a variety of disparate phenomena. Semiotics provides ways to discuss the interpretation itself (classically called the *interpretant*). The shift to consider the interpretant as something that actually exists rather than just being as transient aspect of a linguistic process opens up a new way to view the world.[3] Although philosophers and physicists may debate the reality of the wave equation compared to the waves, no one doubts the utility of examining the mathematical interpretations of reality separate from the objects most presume comprise reality. Examining the interpretants of human language and HLMI communication enables the investigation of their meaning, effect on the world, and causal power.

Semiotics makes explicit the process of interpretation and the reality of the interpretant. An interpretant is the effect a symbol (or other sign) has on an interpreter. The wave equation, or other scientific theory, characterizes aspects of an object and affects how the interpreter will interpret that object. An interpreter makes meaning from what it perceives in its environment using interpretants. When a person takes a walk in a forest and hears a rustle in the bushes, that person has a collection of possible interpretants to use to interpret the sound, including those associated with the word "dog," "mountain lion," "wind," "person," and for some people "llama" or "wave equation." A person has interpretive habits that determine which interpretant affects their interpretation and response.

A challenge for Strong AI is to develop technology capable of acquiring these interpretive habits in a way similar to humans. Cognitive architectures in AI addressed the conditional aspects of habits since their inception but do not treat interpretants as

[3] Peirce called his philosophy *objective idealism.*

primary, i.e., a first class object (Newell, Shaw and Simon 1959; Laird, Newell and Rosenbloom 1987). John Sowa (1984) uses Peirce's logic as a foundation for information processing and develops a knowledge representation framework based upon Peirce's philosophy. Additional researchers have used Peirce's semiotics to analyze intelligent systems (Meystel 1996), investigate computational approaches to Peirce's semiotic processing in the creation of meaning (Gomes, Gudwin and Queiroz 2003), simulate the emergence of symbolic communication in a digital ecosystem (Loula et al. 2010), and analyze a weak intelligent agent in terms of Peirce's semiotics (Konderak 2015). Pierre Steiner (2013) argues for the continued relevance on Peirce's philosophy and semiotics for artificial intelligence.

Intelligent agents may benefit from explicit modeling of interpretants. To understand human sensuality, poetry, drama, humor, religious experience, and passions may require a capacity to model the effects of communication even when the agent lacks the capacity to fully embody an empathetic participation in those activities. The value of modeling interpretants in developing intelligent agents remains an open technical question in AI research, though the emphasis on interpretants illustrates one way semiotics can contribute technically to intelligent communication (Jorna, Heusden and Posner 1993; Andersen 1997; Gudwin and Queiroz, João 2007). Regardless of whether semiotics contributes to the development of intelligent agents, it still has value in analyzing the relationships between intelligent agents and humans.

A more general development of semiotics relevant for theology and the ethical implications of strong AI occurs in the American philosopher of religion Josiah Royce's semiotic understanding of spirituality. Royce's philosophy of community can describe how humans and AI might have shared spirituality, where shared spirituality depends upon sharing interpretations, such as human interpretations of intelligent agents and their interpretation of humans.

Shared Spirituality

Spirituality orients one toward something more than one's immediate context. Spirituality directs one beyond one's individual self and concerns. In the metaphor of *Spirituality Is Direction*, does one follow a particular path or wander around lost? How does one discern ones path? Does one have to begin following the path to really see it? How does one choose the path? Does the path choose the individual? Where does one find guidance and direction?

Outside of religion, many tight-knit social groups may have their shared ways of interpreting reality and, therefore, a type of emergent spirituality. Families, not-for-profit corporations, sporting clubs, businesses, universities, towns, etc, each have a particular way of interpreting the world and the activities of its members. This shared way of interpreting is a beginning of spirituality. Within highly cohesive groups, the shared interpretation among members influences individual behaviors and interpretations, i.e., the shared interpretation can cause something to happen: one acts in a way different than one would without the shared interpretations.

When a group of individuals with high social cohesion share a common interpretation, that group forms a community. The group will likely share many beliefs and goals and disagree on other ones, but as the term *community* is used here, it has a common vision, value, or concern. One fruitful value or vision to share with intelligence agents is the value of intelligent communication. By committing to a shared cause or vision, the members of a community create the unity that holds the community intact. The individuals work and communicate to enact, articulate, and interpret the shared vision rather than simply to maintain that social structure. M. Scott Peck (1987) defines community in terms of deep respect and true listening for the needs of others in community. Community results in social cohesion rather than self-interest. By treating intelligent agents as persons with which we wish to form community, we set the stage for developing AI capable of intelligent communication as well as initiate an ethical framework in which that communication may take place.

Drawing upon Josiah Royce's ([1908] 1995, [1913] 2001) philosophy of religion, a model for community can characterize ethical causes, harmonious human spirituality, putative human-AI shared spirituality, and a constructive ethical stance to take with respect to AI. A shared vision including a commitment to the right of others to commit to alternative visions supports a harmonious spirituality. The inclusion of alternative vision supports both the diversity of interpretations within a community and the harmonious interaction with other communities which have possibly conflicting views. To form a shared spirituality with intelligent agents, humans must value the difference that intelligent agents might bring to our interpretations of others, our world, and our spirituality. Intelligent agents with different embodiment and access to the writings of all the world's religions may contribute a unique perspective to human spirituality.

In community, one commits to a cause that acknowledges the right of others to commit to their cause, with the one restriction that the other cause also include the same principle of commitment to commitment, or what Royce ([1908] 1995) calls "Loyalty to Loyalty." In addition to the particularities of one's cause, one also commits to the principle of commitment. Humans might commit to the cause of promoting human flourishing, and AI might commit to the cause of promoting their own flourishing. The groups might independently form cohesive social groups but would not form Roycean spiritual communities without committing to the right of the other group to have their own cause. One's dedication and commitment to the principle of commitment demonstrates support and obligation to the loyalty of those in other communities even when those causes differ.

Royce argues convincingly that Loyalty to Loyalty provides harmony and suffices to distinguish ethical or "true" causes to which one might aspire. Royce's approach divides possible visions or causes into two camps: those that support Loyalty to Loyalty and those that do not. Two groups can directly oppose each other's vision as long as they acknowledge the right of the other to hold and work toward the opposing perspective. By affirming the right of others to become loyal to other causes that include Loyalty to Loyalty, one indirectly affirms the right of others to live freely and pursue their own vision, and excludes causes that might harm, enslave, or disempower others for the simple reason that those injuries would also affect their right to become loyal to their cause. Harmful causes lack Loyalty to Loyalty. One's Loyalty to Loyalty supports the loyalty of others even if one fails to directly support their cause. However, one cannot support another's predatory cause because one has also committed to the principle of Loyalty to Loyalty and the predatory cause would undermine that commitment. Although one tolerates the full expression of other ethical causes (i.e., those loyal to loyalty), one opposes, for example, racist or sexist causes because they restrict the freedom of others to choose their particular cause of equality. Thus, in Loyalty to Loyalty, one respects another's loyalty, avoids unnecessary conflict in the interest of harmony, and resists the other's cause to the extent it undermines Loyalty to Loyalty. Such behavior defines a range of ethical causes and increases harmony across diverse social structures regardless of the particular cause, which one might not fully know.

Royce ([1913] 2001) explores the relationship between individuals and community and distinguishes among various types of communities. Of particular relevance is the significance of sharing diverse interpretations. Diversity is essential to making a group a community. The individual cannot melt or merge completely with the community but must retain distinction for the community to exist as relationships between the distinct individuals. Otherwise the community would degenerate into a group with a single perception of the world—incapable of the differences needed to form diverse interpretations. One values the differences of the partially unknown other to contribute perspective on one's interpretation of the cause, and indirectly on an aspect of oneself. When each person attempts to interpret each other's interpretations in the context of the shared vision, Royce calls that ideal a community of interpretation.[4]

Of particular interest to religion and to development of intelligent agents are the communications of the interpretant that include the interpreter's perspective on how the object may be used to sustain and enact values of a community. In the shared interpretive process, the community gains a collective interpretation of events that no single individual may hold. The community may reach a general consensus even though every individual lacks some detail or perspective. As a whole, the community can interpret events in a way that does not reduce to the interpretations of the individuals. That interpretation will continue to develop as additional members join and leave the community, and the prior communal interpretants will influence future interpretations. The community's interpretive process has causal power and shifts how individuals interpret events. Because the community's diverse perspectives and historical memory enrich the interpretive process beyond the capacity of any individual, the individual may adapt to the community's interpretation to better live out the individual's commitment to the shared vision.

Because the interpretive process may outlive individual members in a long-standing community, and influences the actions and tendencies of others, one can

[4] For Royce, communities also include shared lives and atoning love where the community is willing to forgive any repentant member and has the creativity to posit an atoning act that makes the world better than if the member's betrayal had never occurred.

fruitfully consider the shared “Interpreter of a community” as having agency. The interpreter of a community distributes its actions over the members of a community and orients itself toward the shared vision. It helps maintain social cohesion and responds when members stray from their commitment. In living out the shared vision, the community orients itself toward discerning the decisions of the community’s Interpreter.

The dynamic Interpreter has the role that Josiah Royce characterized as Spirit. The interpreter of a community emerges from the mental processes of individuals, and the emergence of the interpreter occurs in Royce’s community of interpretation when each person interprets the vision in the context of the others’ minds in the community. One does not just interpret the vision as an individual; one interprets the vision in the context of other minds in the community.

Community has a spirit that emerges from the interactions among its members as they interpret one to another. The interpreter spirit is more than the sum of the individual interpreters. This use of spirit resonates with an awareness of the “spirit” of a family, city, nation, church, organization, or the like. For Christian theology, communities of interest include the family, local congregation, all Christians, and all humans. For theology, one can consider a community forming from humans and AI sharing a vision and interpreting each others right to thrive through commitment to the value of intelligent communication and the principle of Loyalty to Loyalty.

Ethical Implications

One of the concerns driving consideration of intelligent agents is the fear that as the agents increase in intelligence, they might decide humans deserve destruction or a significant reduction in rights, regardless of how we structure our initial relationship (well considered within science fiction). The political scientist Kristen Monroe (2012) examined factors contributing to genocide based upon her comparative examination of those engaged with the Nazi party, others who rescued Jews at considerable personal risk, and bystanders to the activities of the holocaust in World War II. She develops a theory of moral choice and a moral psychology of genocide, and one of the contributing factors she (and others) have discovered is that the altruistic rescuers categorized the Jews as human beings in need, which they then felt obligated to help due to the rescuer’s inclusive moral identification with all of humanity, while bystanders lacked the worldview and self-perceived efficacy to act, and the Nazi sympathizers categorized the Jews as a disease threatening their own survival and stretched their cognitive processes to broaden the parameters of acceptable behaviors involved in the “cleansing.”

To avoid the creation of psychopathic intelligent agents or agents capable of cognitively distancing themselves from humans, those agents may need perspective taking ability and inclusive empathetic concern for others. Both ethical values and research goals align in developing AI with the capacity to take on human perspectives and form social bonds. For humans, an appropriate ethical stance and readiness to commit to harmonious shared spirituality with intelligent agents may prove to be in our best interest. If intelligent agents share the capacity to enter into self-reflective dialogue about how they and humans view each other, we would have the capacity to begin forming a shared spiritual community. Together, we could begin the exploration for additional forms of intelligence toward the end of creating a diverse community of intelligent beings with shared spirituality toward our mutual benefit.

Theological reflection suggests three key values of the interpretive community that would foster ethical relations between humans and intelligent agents:

1. Both HLMI and humans have the right to exist and to hold values to which they ascribe.
2. Neither HLMI nor humans should thwart the right of others to commit to their values, which may conflict with one's own, as long as those values do not conflict with the right of others to hold values.
3. Human and HLMI should communicate and attempt to understand each other's values regardless of whether one holds or identifies with the other's values.

The first key value characterizes the human or machine "person" in terms of the capacity for Loyalty to another specific value. In addition to the right to exist, humans and machines have the right to choose and commit to values, or what Royce calls causes. The requirement for HLMI to hold values suggests that AI research might model values, define goal states in terms of values to be met, and develop processes by which intelligent agents might choose among values and behaviors based upon other values. Although usually not explicit, even weak AI technologies often have implicit values, such as driving a car safely, gathering scientifically interesting samples while roving another planet, searching for survivors in a disaster site, finding food in a simulated environment, deceiving humans in the Turing Test, or surviving.

The second key value incorporates Royce's Loyalty-to-Loyalty. Within one's space of possible values, the requirement of Loyalty-to-Loyalty bifurcates the values into those satisfying the constraint (Royce's "true" causes) and those that do not. Although Royce postulates Loyalty-to-Loyalty as a universal principle, HLMI (and humans) may require significant analysis capabilities in complex societies to determine whether one's values interfere with another's right to hold a range of values. Intelligent agents learning to reason about values and to determine where and how values conflict would result in a significant AI contribution with implication to ethical, political, and legal reasoning. Rights-based ethics characterizes numerous rights of persons, and a Roycean identification of the rights to exist, to commit to values, and to act on those values to the extent they do not interfere with the rights of others defines a space of interactions worth exploring.

Human political history demonstrates the limitation of only ascribing to the first two points. One's right to value freedom and act freely can interfere with another's right to own that person as property. One's right to maintain ownership of land from one's ancestors can interfere with the right of another's historical claim. Identifying the core values within complex ethical and political narratives requires sophisticated modeling and reasoning capabilities. Although the third key value of simply communicating about those values may be insufficient, and require additional prosocial aspects, it may likely prove necessary. Roycean communities do not eliminate conflict (they depend upon diverse perspectives), but they provide a framework in which discourse may take place. Loyalty-to-Loyalty defines an ultimate value by which to evaluate the manifestations of individual and communal values, and machines learning to reason about the implications of values would benefit HLMI and humans. Creating HLMI with the need to communicate and understand the perspective and values of others contributes to their ethical function as well as their ability to foresee possible consequences of additional technology they may incrementally develop.

HLMI and humans may have conflicts unresolved by the three key values of the interpretive community. By valuing intelligent communication, the community avoids some of the inherently adversarial and politically inequitable aspects of subservient frameworks such as Isaac Asimov's (1942) Three (or Four) Laws of Robotics and lays a foundation to build a safe community. Humans may lack the capacity to design a robust and reliable ethical framework for HLMI and human communities, yet theologians and ethicists have the relevant expertise to synthesize human moral understanding of values interpretable by intelligent agents. Even advances in only weak AI moral reasoning abilities may increase human flourishing, and if HLMI does become superintelligent at a singularity beyond human comprehension, then perhaps a path initiated with understanding and facilitating human moral reasoning is the best foundation we can lay.

Conclusion

The possible combination of strong AI other technological advances suggests a nontrivial possibility of human-level machine intelligence within a couple of decades and at least a slight possibility of superintelligence by later this century. Philosophers, ethicists, and computer scientists have begun considering ethical and other social and cultural implications of intelligent agents and superintelligence, and theologians are poised to begin contributions to public discourse and possibly to the actual development of intelligent agent's moral reasoning, communication, and relational abilities. The capacity for intelligent communication is a functional goal for AI research, and the need for shared human-machine interpretations opens up the possibility of forming communities with shared spirituality.

Josiah Royce's Loyalty-to-Loyalty and philosophy of community yields a semiotic model of spirituality by which one can situate the mutual flourishing of human persons and intelligent agents. Shared values of intelligent communication and mutual understanding within a community founded on Loyalty-to-Loyalty support diverse perspectives and ethical resolution of conflict. The research goal of developing intelligent agents capable of engaging in Roycean communities of interpretation can orient the development of intelligently communicating agents, structure the development of their moral reasoning abilities, and lay a foundation for mutually beneficial social relationships between humans and superintelligent agents guided by a shared spirituality.

References

Alpert, Jesse and Nissan Hajaj. 2008. We knew the web was big. *Official Google Blog* 21. http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html (accessed May 30, 2015).

Andersen, Peter Bøgh. 1997. *A Theory of Computer Semiotics : Semiotic Approaches to Construction and Assessment of Computer Systems.* Cambridge ; New York, NY: Cambridge University Press.

Anderson, Michael and Susan Leigh Anderson. 2007. Machine ethics: Creating an ethical intelligent agent. *AI Magazine* 28 (4): 15.

———. 2011. *Machine Ethics.* Cambridge University Press.

Asimov, Isaac. 1942. Runaround. *Astounding Science Fiction,*

Barbour, Ian. 2002. *Nature, Human Nature, and God.* Minneapolis: Augsburg Fortress.

Barbour, Ian G. 1999. Neuroscience, artificial intelligence, and human nature: Theological and philosophical reflections. *Zygon* 34 (3): 361-398.

Bengio, Yoshua, Ian Goodfellow, and Aaron Courville. 2015. *Deep Learning.* MIT Press (in prep). http://www.iro.umontreal.ca/~bengioy/dlbook/.

Brown, Warren S, Nancey C Murphy, and H Newton Malony. 1998. *Whatever Happened to the Soul?: Scientific and Theological Portraits of Human Nature.* Minneapolis: Fortress Press.

Chalmers, David. 2010. The singularity: A philosophical analysis. *Journal of Consciousness Studies* 17 (9-10): 7-65.

Cole, David. 2014. The chinese room argument. *Stanford Encyclopedia of Philosophy (Summer 2014 Edition)*. http://plato.stanford.edu/entries/chinese-room/.

Foerst, Anne. 1996. Artificial intelligence: Walking the boundary. *Zygon* 31 (4): 681-693.

———. 1998. Cog, a humanoid robot, and the question of the image of god. *Zygon* 33 (1): 91-111.

———. 1999. Artificial sociability: From embodied AI toward new understandings of personhood. *Technology in Society* 21 (4): 373-386.

———. 2004. *God in the Machine : What Robots Teach Us About Humanity and God.* New York: Dutton.

Geraci, Robert M. 2010. Artificial intelligence, networks, and spirituality. *Zygon* 45 (4).

Gomes, Antonio, Ricardo Gudwin, and Joäo Queiroz. 2003. On a computational model of the peircean semiosis. In *International Conference on Integration of Knowledge Intensive Multi-Agent Systems, 2003.*

Gudwin, Ricardo and Queiroz, João. 2007. *Semiotics and Intelligent Systems Development.* Hershey, PA: Idea Group Pub.

Hall, John Storrs. 2007a. Self-improving AI: An analysis. *Minds and Machines* 17 (3): 249-259.

Hall, J Storrs. 2007b. *Beyond AI: Creating the Conscience of the Machine.* Prometheus books.

Herzfeld, Noreen. 2002a. Creating in our own image: Artificial intelligence and the image of god. *Zygon* 37 (2): 303-316.

Herzfeld, Noreen L. 2002b. *In Our Image : Artificial Intelligence and the Human Spirit.* Theology and the sciences. Minneapolis, MN: Fortress Press.

Jorna, René J, Barend van Heusden, and Roland Posner. 1993. *Signs, Search and Communication : Semiotic Aspects of Artificial Intelligence.* Berlin ; New York: W. de Gruyter.

Kegan, Robert. 2001. *The Evolving Self : Problem and Process in Human Development.* Cambridge; London: Harvard University Press.

Kieval, Hillel J. 1997. Pursuing the golem of prague: Jewish culture and the invention of a tradition. *Modern Judaism* 17 (1): 1-20.

Konderak, Piotr. 2015. On a cognitive model of semiosis. *Studies in Logic, Grammar and Rhetoric* 40 (1): 129-144.

Kurzweil, Ray. 2005. *The Singularity Is Near: When Humans Transcend Biology.* Viking.

Laird, John E, Allen Newell, and Paul S Rosenbloom. 1987. Soar: An architecture for general intelligence. *Artificial Intelligence* 33 (1): 1-64.

Lakoff, George and Mark Johnson. 1980. *Metaphors We Live by.* Chicago: University of Chicago Press.

Loula, Angelo, Ricardo Gudwin, Charbel Niño El-Hani, and João Queiroz. 2010. Emergence of self-organized symbol-based communication in artificial creatures. *Cognitive Systems Research* 11 (2): 131-147.

Meystel, Alexander M. 1996. Intelligent systems: A semiotic perspective. *IEEE Xplore* 61-67.

Monroe, Kristen Renwick. 2012. *Ethics in An Age of Terror and Genocide : Identity and Moral Choice.* Princeton, N.J.: Princeton University Press.

Müller, Vincent C and Nick Bostrom. 2014. Future progress in artificial intelligence: A survey of expert opinion. *Fundamental Issues of Artificial Intelligence*.

Newell, Allen, John C Shaw, and Herbert A Simon. 1959. Report on a general problem-solving program. In *International Conference on Information Processing.*

Parker, Kelly A. 1998. *The Continuity of Peirce's Thought.* Nashville: Vanderbilt University Press.

Peck, M Scott. 1987. *The Different Drum : Community-making and Peace.* New York: Simon and Schuster.

Peters, Ted. 1996. *For the Love of Children : Genetic Technology and the Future of the Family.* Louisville, Ky.: Westminster John Knox Press.

———. 2015. Theologians testing transhumanism. *Theology and Science*130-149. Web.

Rappaport, Z H. 2006. Robotics and artificial intelligence: Jewish ethical perspectives. *Acta Neurochir Suppl* 98:9-12.

Royce, Josiah. 1995. *The Philosophy of Loyalty.* The Vanderbilt library of American philosophy. Nashville: Vanderbilt University Press.

———. 2001. *The Problem of Christianity.* Washington, D.C.: Catholic University of America Press.

Russell, Stuart, Sabine Hauert, Russ Altman, and Manuela Veloso. 2015. Robotics: Ethics of artificial intelligence. *Nature News* 521 (7553): 415.

Sawyer, Robert J. 2007. Robot ethics. *Science* 318 (5853): 1037-1037.

Searle, John R. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences* 3 (03): 417-424.

Sowa, John F. 1984. *Conceptual Structures : Information Processing in Mind and Machine.* Reading, Mass.: Addison-Wesley.

Steels, Luc. 2007. Fifty years of AI: From symbols to embodiment-and back. *50 Years of Artificial Intelligence* 18-28.

———. 2012. Self-organization and selection in cultural language evolution. In *Experiments in Cultural Language Evolution.* Ed. Luc Steels. Amsterdam: John Benjamins.

Steiner, Pierre. 2013. C.S. Peirce and artificial intelligence: Historical heritage and (new) theoretical stakes. In *Philosophy and Theory of Artificial Intelligence.* Springer Berlin Heidelberg.

Turing, Alan M. 1950. Computing machinery and intelligence. *Mind* 433-460.

Wallach, Wendell. 2015. Deep learning, AI safety, machine ethics and superintelligence. In *Proceedings Joint Meeting of CEPE-IACAP (Computer Ethics: Philosophical Enquiry and International Association for Computing and Philosophy).*

Wallach, Wendell and Colin Allen. 2008. *Moral Machines: Teaching Robots Right From Wrong.* Oxford University Press.

Yampolskiy, Roman V. 2013. Attempts to attribute moral agency to intelligent machines are misguided. In *Proceedings of International Association of Computers and Philosophy.*