Augustine, Aquinas, and Asimov: How traditional Christian thought applies in the creation of artificial moral agents

Gene Rohrbaugh
Messiah College

Follow this and additional works at: https://mosaic.messiah.edu/facscholar

Part of the Christianity Commons, and the Higher Education Commons

Recommended Citation
https://mosaic.messiah.edu/facscholar/31
Augustine, Aquinas, and Asimov
How traditional Christian thought applies in the creation of artificial moral agents
Gene Rohrbaugh, PhD
Associate Professor of Computer Science
Messiah College

1. Introduction

"Obey the rulers who have authority over you."
Romans 13:1 (Contemporary English Version)

"A robot must obey any orders given to it by human beings..."
Asimov's second law of robotics

"I'm sorry, Dave. I'm afraid I can't do that."
HAL 9000, in 2001 Space Odyssey

One of the most exciting yet controversial trends in computing today is the deployment of autonomous agents—computational systems that act independently and intelligently to achieve specific goals. Agents already exist that can carry out complex tasks such as piloting aircraft (see for example, Matthews, 2008; Vijaya, Suresh, Omkar, Ganguli, & Sampath, 2009), driving cars (Seetharaman, Lakhotia, & Blasch, 2006; Bohren et al., 2008), and trading stocks (Bao & Yang, 2008). In the near future they may carry out even more consequential tasks, such as delivering medical treatment (Jayender, Azizian, & Patel, 2008; Jayender, Patel, & Nikumb, 2009; Rogers, Light, Allmen, & Smith, 2009; Liang, Rogers, Light, von Allmen, & Smith, 2010; Merritt, 2010), and even selecting targets and firing weapons (Kumagai, 2007; Sparrow, 2007; Arkin, 2008; Arkin, 2009). The feature that makes these systems incredibly powerful is their autonomy—their ability to choose a goal or course of action without human intervention.

Nevertheless, imbuing machines with autonomy gives rise to numerous valid concerns. What happens if such an autonomous agent makes wrong choices? What if the autonomous medic administers the wrong treatment, or the autonomous military drone targets the wrong building? Ordered to rush into a burning building (or fire a weapon), do we want a machine to have the right to refuse? Such scenarios are no longer the sole purview of science fiction authors, but are being seriously addressed through scholarly research in Philosophy, Engineering, and Computer Science (see, for example Allen, Varner, & Zinser, 2000; Stahl, 2004; Peterson, 2007; Anderson, 2008; Wallach, 2008; Wallach, Allen, & Smit, 2008).

Given the rush to develop autonomous agent systems and the obvious potential for harm, how is a Christian in the computing sciences to respond? Is the development of such autonomous agents compatible with Christian faith? How can ethical principles inform our response? The juxtaposition of revered Christian thinkers Augustine and Aquinas with the atheist Asimov in the title is intended to reflect the inherent tension in trying to apply historically-grounded Christian thinking to a topic that is only now on the cusp of possibility. Is there a common ground for discourse between an explicitly Christian philosophical and theological tradition dating back to the early church, as represented by Augustine and Aquinas,
and the contemporary philosophical debate on the development of autonomous agents, as explored in the fiction of the Asimov?

In this paper, I will show that the thinking of Augustine and Aquinas can be applied to the development of artificial moral agents by addressing two fundamental questions regarding the implementation of such systems:

1. Is the development of artificial moral agents compatible with traditional Christian thought (the ethics of their implementation)?

2. If so, how can Christian ethical principles guide us in their creation (ethics in their implementation)?

To set the stage for the discussion of these two questions, this paper is ordered as follows: In section 2, I present an overview of the field of machine ethics, including some key definitions, commonly-used frameworks, and methodological approaches. In section 3, I present a brief summary of traditional Christian thinking from relevant portions of the writings of Augustine and Aquinas. Finally in section 4, I return to the two fundamental questions above and apply the ideas of Augustine and Aquinas in the context of the development of artificial moral agents.

2. Problem Context for Artificial Moral Agents

The motivations for developing autonomous machines are numerous. They carry out tasks that are deemed tedious (e.g., vacuuming the floor), undesirable (collecting garbage), unsafe (searching for survivors in collapsed buildings), or even impossible for humans (piloting severely damaged aircraft). They work in environments where people cannot, as in deep sea or space exploration. They perceive and process massive amounts of data that would be inaccessible or overwhelming to humans, for example detecting patterns of credit card fraud or identifying individual genes amidst massive DNA datasets. The development of autonomous machines seems especially desirable in cases where they might replace humans in dangerous contexts (e.g., military missions or search and rescue) and where they can achieve safety and efficiency otherwise unattainable (governing complex systems). Nevertheless the idea of autonomous machines carries concomitant concerns about their reliability and trustworthiness, leading to a desire to develop a sort of constrained autonomy, which allows sufficient freedom for the performance of the required tasks without granting absolute freedom.

Definitions

In the field of artificial intelligence, the term agent refers, in its most basic sense, to any entity that perceives and acts in an environment (Russell & Norvig, 2003). This definition covers both natural agents such as humans, who perceive sights and sounds, and simultaneously act upon their environment using hands and feet; and artificial agents such as a Roomba® vacuum cleaning robot, which perceives obstacles and dirt in its environment and acts by navigating around the room and using suction to clean the floor. The term percept (from perception) is
used to refer to data that an agent acquires about its operating environment through its sensors.

Though even simple devices can be portrayed as agents (a thermostat ‘perceives’ the ambient temperature and ‘acts’ by turning the heat on or off), the true usefulness of the agent paradigm is in the development of more complex intelligent agents. An intelligent agent, then, is an agent that acts intelligently, meaning that it does the right thing(s). As the determination of the “right thing” to do becomes increasingly complex, so too does the intelligence required on the part of the agent. For example, in a simple game such as tic-tac-toe, the correct move at any point in the game (one that will lead to a win or draw) is much easier to find than in a more complex game such as chess or Go. Thus an agent that plays chess well exhibits greater intelligence in doing so than one that plays tic-tac-toe well. Further expansions on the notion of an agent presume a goal of intelligence, and focus on specifying the means whereby the agent makes decisions.

Types of Agents

Beyond this basic definition of an intelligent agent, a variety of agent taxonomies have been developed (Franklin & Graesser 1996; Sánchez, 1997; Russell & Norvig, 2003; Moya & Tolk 2007). Relevant expansions include reflex agents, knowledge-based agents, learning agents, goal-based agents, and utility-based agents.

A reflex agent is an agent that bases its decisions entirely on its percepts. For example, existing robotic vacuum cleaners operate as reflex agents. They select their actions (turn, move forward, engage suction vacuum) based on their current percepts. “If dirt is detected, engage suction; if obstructed, turn a random direction; otherwise proceed forward.”

A knowledge-based (or model-based) agent is an agent that, in addition to percepts considers internalized knowledge in its decision making. A checkers-playing agent, for example, may store a database of strategies, which it can then consult in order to find appropriate responses to the opponent’s moves.

A learning agent is a knowledge-based agent that acquires and stores newly-gained knowledge while operating in its environment. A search-and-rescue robot, for example, may know nothing about its operating environment when it is deployed; but as it encounters objects of interest such as doors or walls it can record such information, building a mental map of its environment.

A goal-based agent is a knowledge-based agent that has an internal representation of one or more goal states and uses reasoning to predict which action or sequence of actions is likely to lead to a goal state. For example, a robotic taxi could pick up a passenger and then adopt the goal of arriving at JFK airport. Consulting stored maps, it can determine which sequence of roads will allow it to achieve its goal.

Finally, a utility-based agent is a goal-based agent that adds a utility evaluation metric to select among alternative action sequences that all lead to the same goal state. The robotic taxi, for example, might discover numerous paths that lead to its destination; it might then rank
them as shortest, quickest, most scenic, etc., and choose a path based on whichever utility metric is in play.

**Autonomy**

In a broad philosophical sense, autonomy is freedom to act. Kant’s conception of autonomy includes both negative and positive components:

- A moral agent is free in a *negative* sense insofar as no foreign causal forces dictate what she, as a rational agent, ought to do. Moral agents are free in a *positive* sense insofar as reason is freely able to give to itself and follow laws of its own fabrication—free will as subject only to its own laws. (Tonkens, 2009, p. 427)

Within the field of AI, the term *autonomous* is typically interpreted in a narrower sense, corresponding to Kant’s negative conception of autonomy. Lin, Bekey, and Abney (2008), for example, define autonomy in machines as “the capacity to operate in the real-world environment without any form of external control, once the machine is activated...” (p. 4).

Consider a maze-traversing robot with a well-informed designer. If the designer knows the details of the maze to be traversed, she could simply specify the appropriate sequence of steps (10 inches forward; turn 90 degrees right; 3 inches forward; etc); the agent would not even need percepts. Such an agent lacks autonomy in the relevant sense. It may solve that one particular maze in the most efficient way; however, placed in another maze, it would fail miserably.

Rather than a simple binary feature, autonomy is viewed as a scale, such that the autonomy of an agent is inversely related to the degree to which its specific actions are determined by its designers or users. A maze-traversing robot with greater autonomy could make navigational decisions in situ; sensing its environment; noting where it detects walls; and adjusting its course accordingly. Thus the term *autonomous agent* as used in AI does not imply absolute autonomy. An artificial agent’s choices will be inspired, directed, and constrained by its programming; for now, at least, more so than a human’s choices are inspired, directed, and constrained by her natural inclinations, predispositions, and neuro-chemical makeup. But the clear trend is toward greater autonomy, and it is not at all unimaginable that artificial agents may someday rival humans in the complexity and mystery of their decision-making processes.

**Artificiality**

To distinguish the entities under consideration from humans and other ‘natural’ agents, the term *artificial agent* is used. Though the applicability of the term *artificial* may be clear enough in a case such as a stereotypical robot, whose body is composed of artificial (i.e., non-biological) parts. But the same philosophical questions would arise for any constructed agent, regardless of the details of its constituent parts. Thus, if a robot could someday be built using chemical, biological, or any other sort of components, one might still wonder about the nature of its autonomy and its moral status. Thus the term *artificial* as applied to agents should be interpreted as referring to their origin (artificially constructed) rather than their constitution.
Machine Ethics

Even apart from the scholarly literature, the corpus of science fiction has clearly portrayed the dangers of turning over to machines the power of decision-making, for example, movies such as Terminator or 2001 Space Odyssey. Any attempt to build autonomous machines, then, must take steps to protect the interests of humans by limiting the potential actions of the machines. The consensus view in the literature is that the field of ethics, which in its broadest sense has dealt with principles that govern the choices and behavior of natural autonomous agents, is the appropriate place to look in the quest to constrain the actions of artificial autonomous agents. This field has adopted the term Machine Ethics, which first appeared in Waldrop (1987, p. 38), in a discussion that includes Asimov’s vision of robotics.

Artificial Moral Agents

Allen et al. (2000) first introduced the term artificial moral agent (AMA) to refer to any autonomous agent produced by artificial intelligence technology that “...takes into consideration the interests of others rather than acting solely to advance... [its own] self-interest.” Moor (2006, 2009) describes a hierarchy of four sorts of moral agents: (1) ethical-impact agents, (2) implicit ethical agents, (3) explicit ethical agents, and (4) full ethical agents.

An ethical impact agent, according to Moor, is simply one “...whose actions have ethical consequences, whether intended or not” (Moor, 2009, p. 12) He gives the example of robotic camel jockeys, whose adoption in Qatar has reduced the exploitation of children previously used as jockeys in camel races. Any artificial agent will almost certainly be an ethical impact agent, as it seems unlikely that it could completely avoid having at least some ethical consequences, whether positive or negative. The Code of Ethics and Professional Conduct of the Association for Computing Machinery requires computing professionals to consider and take into account the impact of their creations (Code, 1992); thus I take it as given that any agent created must be at least an ethical impact agent.

An implicit ethical agent is one whose design has incorporated ethical considerations (Moor, 2009, p. 12).

For example, automated teller machines and Web banking software are agents for banks and can perform many of the tasks of human tellers and sometimes more. Transactions involving money are ethically important. Machines must be carefully constructed to give out or transfer the correct amount of money every time a banking transaction occurs. A line of code telling the computer to be honest won’t accomplish this. (Moor, 2006, p. 19)

Thus an implicit ethical agent acts in accordance with ethical principles, but need not represent them explicitly or even recognize them as such.

An explicit ethical agent is one that can “identify and process ethical information about a variety of situations and make sensitive determinations about what should be done. When ethical principles are in conflict, these robots can work out reasonable resolutions” (Moor, 2006, p. 12).
Finally, full ethical agents, according to Moor, have all of the characteristics and abilities of explicit ethical agents, and in addition “have those central metaphysical features that we usually attribute to ethical agents like us—features such as consciousness, intentionality, and free will.” Although these are certainly interesting philosophical issues worthy of scholarly discussion, they are also notoriously difficult; and they need not be resolved in order to have a productive discussion about artificial agents. Therefore the remainder of this paper will focus on implicit and explicit ethical agents—which could collectively be called ethically acting agents. Whether their actions are merely in accordance with ethical principles, or whether they are the result of ethical deliberation, it is clear that the result is preferable to the alternative: autonomous agents who act without concern for ethical principles.

With the necessary definitions of key terms in place, I can now specify precisely the class of agents that the remainder of this paper will focus on: artificial, autonomous, learning, goal-based, ethically-acting agents. These are constructed entities that perceive and act in an environment, make operational decisions rather than relying on outside agents to do so, choosing their course of actions to achieve explicitly represented goals, using both stored knowledge and knowledge acquired through perception, while applying explicit ethical reasoning, or at least acting in accordance with ethical principles, in their decision making.

**Two issues of moral status**

In the discussion that follows, there are two recurring issues related to the moral status of an entity. The first is that of moral agency. A moral agent is simply a being who is capable of acting with reference to right and wrong. Of particular interest is the question of who (what sorts of beings) can have moral obligations. In the discussion that follows, an additional distinction will be useful: weak moral agency versus strong moral agency. I will use these terms to distinguish the nature or degree of agency exhibited by the being in question. A strong moral agent is autonomous in both the positive and the negative senses. The choices and behavior of a strong moral agent are motivated and governed by internalized explicit ethical reasoning. They are agents in a strong sense—able to freely choose which principle(s) to follow and which to disregard. A weak moral agent, in contrast, exhibits autonomy only in its negative conception. It does not freely choose which principles it will follow; but it can nevertheless exhibit complex behavior in accordance with ethical principles imposed on it, reason about and even justify complex moral decisions.

The second issue of moral status is that of moral considerability. Philosophers and ethicists use the term morally considerable to refer to beings whose interests are to be considered in relevant moral deliberations. Tragically, humans all too often have denied the moral considerability of others; devaluing children for example, or women, or people of a different racial or ethnic heritage than one’s own. Progress has been made (and one hopes, will continue to be made) in expanding the circle of morally considerable beings to embrace (at least) all of humankind; but the very search begs the question of what principled basis might be applied. Isaiah Berlin (1969) summarized as follows:

> Unless men are held to possess some attribute over and above those which they have in common with other natural objects—animals, plants, things, etc.—(whether this difference is itself called natural or not), the moral
command not to treat men as animals or things has no rational foundation.
(as cited in Watson, 1992, p. 1)

Numerous authors have further explored the extension of moral considerability to animals (see for example Hayward, 1994; Steiner 2005), all living things (Schönfeld, 1992), ecosystems (Callcott, 1984; Watson, 1992), extraterrestrial life (Callcott, 1992), robots (Petersen, 2007), and other artificial life (Thiel, Bergmann, & Grey 2003).

Parallel to the notions described above of strong and weak moral agency, it will be similarly useful to distinguish two sorts of moral considerability. Strong moral considerability is roughly equivalent to the traditional usage of moral considerability: it applies to beings who have standing in a moral milieu; beings to whom debts or obligations are owed directly. In contrast weak moral considerability would apply to entities that must be considered in moral deliberations, but not for their own sake. As an example, consider the obligations one incurs upon borrowing a car from a friend. The borrower should certainly take certain steps to ensure the “well-being” of the car (e.g., checking the oil; driving safely); but not because the car has standing or is owed a direct debt. Instead there is a debt owed to the owner of the car, who is morally considerable in the strong sense. The car in this example is merely an object with weak moral considerability.

**Ethical Frameworks for Machine Ethics**

As mentioned above, scholars in the field of machine ethics have explored the existing literature on human ethics to find applicable tools, concepts, and approaches. Two common frameworks considered as possible foundations for machine ethics are deontology and consequentialism.

**Deontology**

Deontological approaches to ethics hold that the rightness or wrongness of an act is an inherent quality of the act itself, apart from its consequences. Kant (1998) develops a deontological moral framework summarized in a general universal law known as the categorical imperative:

“Act only in accordance with that maxim through with you can at the same time will that it become a universal law.” (p. 31)

Imagine a robotic taxi driver, running late as he approaches a red light. He considers the question of whether he should run the red light in order to make up lost time. Application of Kant’s categorical imperative would yield the answer that he should only do so if he would like running red lights to become a universal requirement. Since the likely result would be frequent collisions, slowing the taxi even more, the conclusion is that he should not run the red light.

Kant expands on the categorical imperative with a second formulation or maxim:
“So act that you use humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means.”
(Kant 1998, p. 38)

As an example of applying this second formulation, consider the question of whether it is ethical to have a child only in order to harvest its organs for transplant. By doing so, the parent treats the child as a means to an end (prolonging or enhancing the parent’s own life), rather than as an end in and of herself. Thus, according to the second formulation of Kant’s categorical imperative, having a child for such purposes is clearly unethical.

How would Kant address the two issues of moral status? As to moral agency, note that in his first formulation, the phrase ‘universal law’ presupposes some agreed-upon scope of applicability (presumably ‘universal’ does not mean that the law should apply to mice or stalks of wheat). That scope would define the extent of the class of moral agents. As to moral considerability, in his second formulation, Kant explicitly references humanity, implying that only humans are morally considerable. For Kant, the class of moral agents is coextensive with the class of morally considerable beings (a principle known as agent-patient parity): both being exactly the class of human beings.

**Consequentialism**

Contrasting with deontological ethical frameworks, consequentialist approaches hold that the rightness or wrongness of an action depend solely on its consequences. Actions such as lying or telling the truth, healing or killing, are not right or wrong in and of themselves; rather individual acts of lying or killing would be judged right or wrong depending on their particular consequences. One of the better-known consequentialist theories is the utilitarianism developed first by Jeremy Bentham (1789) and later by John Stuart Mill (1864). In utilitarian models of ethics, an action is judged right or wrong based on the impact it has on the parties involved. An act is right if the total benefit (or pleasure) outweighs the total harm (or pain).

Just as it was for deontological approaches, the question of moral considerability is a central issue for utilitarian ethical theories. Whose interests should be considered in calculating total benefit and total harm? Should any one agent’s interests be weighted higher than those of other agents? Should the interests of animals be considered? Note that Bentham and Mill’s consequentialism is viewed as more generous in its treatment of animals, since they can experience pleasure and pain and thus can be considered in the calculus of moral virtue.

**Methodological Approaches**

Apart from considerations about the structure and content of the ethical principles that are to govern the behavior of an agent are the questions about how to implement them. Methodological approaches to machine ethics mirror the larger field of artificial intelligence in that they can be classified as either top-down or bottom-up. Top-down approaches to AI, also known as symbolic approaches or knowledge-based approaches, rely on explicit representation of facts and knowledge combined with logic-based reasoning (see, for example Nilsson, 1980; Lenat, 1995). Wallach, et al. (2008) combine this general engineering sense of “top-down” with a slightly different meaning as applied in ethics, to arrive at the following definition:
"...a top-down approach to the design of AMAs is any approach that takes the antecedently specified ethical theory and analyzes its computational requirements to guide the design of algorithms and subsystems capable of implementing that theory." (Wallach, et al., 2008, p. 569)

One advantage of such an approach is the availability of reasoning traces—in effect the agent can explain its chain of reasoning in terms of explicit rules and principles that people can understand. As applied within the field of machine ethics, top-down approaches would make use of explicitly represented specific knowledge (“Fred is human”), general knowledge (“humans are damaged when run over by cars”), contextual knowledge provided by percepts (“Fred is standing in the middle of the road in front of me”), and explicit ethical principles (“Don’t harm people”) to arrive at specific decisions (“I should apply the brakes”).

Contrasting with traditional top-down approaches are bottom-up approaches to AI, which rely on statistical methods, neuronal simulations, evolutionary programming, and learning algorithms. In bottom-up approaches, tasks are “…specified atheoretically using some sort of performance measure (such as winning chess games, passing the Turing test, walking across a room without stumbling…” (Wallach, et al., 2008, p. 569). Builders of such systems then apply the various techniques to progressively refine the performance of the system until the specified performance criteria are met. Bottom-up approaches benefit from the rapid growth of available computing resources, as compared to the relatively stable or declining availability of human resources required to implement top-down systems; however, Lin et al. (2008) note that

“…computer scientists who wish to build robots with higher-order faculties out of discrete subsystems are confronted with a difficult, and perhaps insurmountable, challenge of assembling components into a functional whole. Whether the aggregation of discrete skill sets will lead to the emergence of higher-order cognitive faculties—including emotional intelligence, moral judgment, and consciousness—can only be known once roboticists go through the exercise of building the systems.” (Lin et al., 2008, p. 35)

Wallach (2008) is much more optimistic about bottom-up approaches, suggesting that they have much in common with natural processes that may have given rise to hardwired behavioral constraints in living beings:

In bottom-up approaches to the development of moral acumen the emphasis is placed on creating an environment where an agent explores courses of action, learns and is rewarded for behavior that is morally praiseworthy. Evolution provides such a bottom-up model for the adaptation, mutation, and selection of those agents best able to meet some criteria for fitness. Game theorists and evolutionary psychologists theorize that at least some moral principles, such as cooperation and fairness, might be naturally selected during evolution and hardwired into the genes of even simple organisms. (Wallach, 2008, p. 467)

The key distinguishing feature between top-down and bottom-up approaches that will be relevant in the discussion below is the former’s explicit reasoning—the internal representation
of facts, ideas, hypotheses, and principles, along with the processes of reasoning. Inherent in top-down approaches, these features also appear in hybrid approaches that combine the two. Such hybrid approaches are recommended by Lin et al. (2008, p. 41) and Anderson & Anderson (2007). Nevertheless, even though bottom-up approaches do not start with such explicit symbolic manipulation, it is possible that such could arise as an emergent property in an agent built through a bottom-up approach.

3. Asimov, Augustine, and Aquinas

Isaac Asimov was just a teen when Konrad Zuse invented the first modern programmable computer in 1936, and consequently he lived his adult life and wrote in a world where computers grew ever faster, more powerful, and more ubiquitous. In contrast, Augustine (345-430 A.D) and Aquinas (1225-1274) predated modern computing by many centuries. One can only expect, then, that the writings of each focused on the realities of his day. The portions of their writings most relevant to the topic at hand are those that deal with ethics. In particular, this section will explore their ideas on the two issues of moral agency and moral considerability.

Asimov

Though known primarily for his contributions in science fiction, where he has explored many of the salient issues, Asimov is frequently cited in scholarly research on Machine Ethics (Lin, et al., 2008 pp. 29-33; Anderson, 2008; Wallach 2008; Allen, et al., 2000; Wallach, et al., 2008; Peterson, 2007). A self-identified atheist (Free Inquiry, 1982) and open critic of evangelical Christianity (The Religion, 2005), Asimov might be expected to have little in common with Augustine and Aquinas.

One of Asimov’s greatest contributions, both by his own admission (Asimov & Asimov, 1995) and by frequency of citation in scholarly literature, is his articulation of the “Three Laws of Robotics” (Asimov, 1950). These three laws were intended as constraints on the choices and actions of robots, hard-wired into their “positronic circuits” as a measure of protection for human owners.

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.

2. A robot must obey any orders given to it by human beings, except where such orders would conflict with the First Law.

3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

Asimov’s Approach to Machine Ethics

Asimov’s three laws of robotics are clearly deontological in approach—no provisions are given for special cases where some other good may result from the violation of one of the laws.
However, much of Asimov’s fiction is dedicated to exploring the unforeseen and often undesirable consequences of the laws and their interactions.

_Moral Status of Asimov’s Robots_

Asimov portrays his robots as explicit moral agents, able to reason about morality, resolve conflicting ethical principles, and articulate clear justifications for their actions. They are, however, moral agents only in the weak sense, as they cannot freely choose whether or not to follow the three laws. They are ascribed ethical duties, act in accordance with ethical principles, and can reason about moral dilemmas; but they cannot choose which to follow and which to disregard.

In Asimov’s fictional world, robots are at best weakly morally considerable. The focus in his stories is on the obligations of robots toward humans, as illustrated in Laws 1-2 (thereby demonstrating the moral considerability of humans). And whereas Law 3 might be viewed as protecting the interest of robots, note that it is not a prohibition of harming other robots; only against a robot harming himself or herself. One of Asimov’s stories—“Evidence”—features a politician who is suspected of being a robot. The politician publicly strikes another human, apparent proof of his own humanity as Law 1 would prevent a robot from doing so. Later, a robotics expert who is not convinced speculates that the person struck was actually a robot as well, and a part of the ruse. This line of thinking seems to indicate that Asimov’s robots are not due the same consideration as humans.

_Augustine_

Saint Augustine Bishop of Hippo (345-430 A.D.) was an early church father, a philosopher, and a theologian who wrote extensively on Christian doctrine. His ideas were widely influential in both the Catholic Church and in the thinking of key figures in the protestant reformation, such as Calvin (McMahon, 2003).

_On moral agency._

Regarding the moral status of earthly beings, Augustine carefully delineates a hierarchy of beings based on likeness to God:

“Among living things, the sentient are placed above those which do not have sensation: animals above trees for instance. And, among the sentient, the intelligent are placed above those which do not have intelligence; men, for example, are above cattle.” (Augustine, as cited in Steiner, 2005, p. 116).

According to Augustine’s thinking, “the more rational and less dominated by bodily impulses a being is, the closer it is to the divine” (Steiner, 2005, p. 117). This rationality that serves to distinguish humans from all other earthly beings, for Augustine, is equated with the “ability to perceive the truth, which is ever unchanging” (Steiner 2005, p. 117). By following this inner light of rationality, humans can rule over their natural impulse to pursue physical pleasure and avoid pain. Animals, being purely physical in nature and thus by definition lacking rationality, cannot do so. Thus, for Augustine, there can be no weak moral agents as defined above; if a being is not
a strong moral agent, ruled over by the spiritual inner light of reason, then it cannot act ethically.

**On moral considerability**

In reference to moral considerability, Augustine follows the Stoic tradition in the view that the material world exists for the benefit of humans. Thus humans, as moral agents, owe no obligations directly to lower beings such as animals. This does not mean, however, that the well-being of animals need not be considered. For Augustine, though all of creation exists for the use of humans, we are to use them in our quest for redemption (Steiner, 2005, p. 117). Exploitation of nature for the sake of self-gratification goes against that ultimate purpose. Steiner (2005) summarizes Augustine's position as follows:

"... it is important to note that limits on our exploitation of animals are intended not to benefit animals but rather human beings: The problem with the lust for mastery in earthly affairs is not that it does violence to animals and other subluminary beings, but that he who 'is avid for mastery surpasses even the beasts in the vices of cruelty and luxury.' Augustine's concern here is with the corruption of the human soul, not with the fortunes of beings such as animals." (p. 118)

So for Augustine, our duties concerning the treatment of animals are duties owed to humans (both to oneself and to one's fellow humankind), not to the animals themselves. Our moral obligations are to other humans (and higher beings); not to lower beasts; "every soul is better than every body." (Augustine, as cited in Steiner, 2005 117). Thus, for Augustine, animals and lower beings are morally considerable, but only in the weak sense. Humans, on the other hand, are morally considerable in the strong sense.

**Aquinas**

Writing nearly a thousand years after Augustine, Aquinas (1225-1274 A.D.) nevertheless articulates a philosophy of ethics remarkably consistent with—and in fact often drawing from—his precursor. A Catholic priest of the Dominican order, Aquinas was influential not only within the church, but throughout much of Western thought and modern philosophy.

**On moral agency**

Like Augustine, Aquinas distinguishes humans from all non-human animals on the basis of rationality; but for Aquinas, the nature of rationality is less mystical than Augustine's "inner light." For Aquinas, rationality is associated with "universal reason" or conceptual abstraction—an ability to reason abstractly (Steiner, 2005, p. 129). Nevertheless, though his conception of rationality is different, like Augustine, Aquinas views rationality as the mechanism of human autonomy.

"The intellectual agent acts for an end, as determining for itself its end; whereas the natural agent, though it acts for an end... does not determine its end for itself, since it knows not the nature of the end, but is moved to
the end determined for it by another.’” (Aquinas, as cited in Steiner, 2005, p. 127)

For Aquinas, animals’ lack autonomy as a consequence of their inability to recognize and reason about universals. “Beings that are ‘more noble’ stand in closer proximity to God, in virtue of their rational capacity for self-determination.” (Steiner, 2005, p. 127). Thus while Aquinas follows Augustine in attributing strong moral agency to humans alone, he does not deny that animals may exhibit a sort of weak moral agency: “In the actions of animals ‘we notice certain marks of sagacity,’ but what we are referring to as ‘sagacity’ is merely a simulacrum the like of which ‘may be seen in the movements of clocks’” (Steiner, 2005, p. 128)

**On moral considerability**

Much like Augustine, Aquinas holds that though we do not owe moral consideration directly to animals (and thus they do not enjoy strong moral considerability), we do have obligations to our fellow-humans with regard to our treatment of animals. Thus they are morally considerable in the weak sense:

“...since it happens that even irrational animals are sensible to pain, it is possible for the affection of pity to arise in a man with regard to the sufferings of animals.’ The value of such an affection is simply that ‘if a man practice a pitying affection for animals, he is all the more disposed to take pity on his fellow-men’” (Steiner, 2005, p. 131).

**Summary**

On the issues of moral agency and moral considerability, Augustine and Aquinas agreed that human beings (and not non-human animals) carried moral responsibilities; and that human beings (and not non-human animals) were owed moral consideration. Moral agents may need to give thought to the well-being of animals, but never for their own sake. How, though, can their thinking be applied to artificial moral agents—entities of a sort that were not a part of the world or the thinking of their respective eras? To answer this question, it is necessary to look beyond the de facto classification of humans, to the principles upon which Augustine and Aquinas argue for the distinction between moral agents and beings who are not moral agents; between those who are morally considerable and those who are not. With respect to the relevant features, are artificial moral agents more like humans or more like animals?

**4. Application**

Having set the stage in the previous sections, I now return to the two motivating questions that were introduced at the beginning:

1. Is the development of autonomous moral agents compatible with traditional Christian thought (the ethics of their implementation)?
2. If so, how can Christian ethical principles guide us in their creation (ethics in their implementation)?

It bears repeating that both Augustine and Aquinas argued that humans stood alone in the natural world as moral agents and as morally considerable beings, at least in the strong sense of the latter. Both carefully distinguish humans from non-human animals in their writings on ethics. But as each lived in a time when artificial agents were not part of the conceptual framework, one must look beyond what is explicitly stated in their writings to find the principles that they use to determine or justify moral status.

Both Augustine and Aquinas develop lines of thinking according to which the development of AMAs is ethically desirable (or even ethically mandatory), with certain caveats. Autonomy should be extended only in conjunction with ethical ability. As autonomy increases, so will our obligations toward AMAs, requiring a cautious approach. Along with autonomy and rationality, certain other features must be present in order to support ethical action.

**Ethics of the development of AMAs**

To answer the question of whether we should develop AMAs—whether doing so is compatible with Christian faith—this section will address obligations to fellow humans, obligations for responsible forethought, and finally ethical obligations that might be owed directly to AMAs.

**Obligations to fellow-humans**

Augustine and Aquinas never equivocate about the moral considerability of humans—the well-being of other people must be considered in our ethical deliberations. The Bible clearly teaches that Christians have certain moral obligations to our fellow humans: to feed the hungry (Matthew 25:35; Isaiah 58:10); to clothe the naked (Isaiah 58:7); to care for the fatherless and the widow (James 1:27). Though we cannot expect to relieve all human suffering (and some would argue that it is not desirable to do so), we should not refuse help when it is in our power to help. Consider, then, how AMAs might be used:

- A search and rescue robot enters collapsed or burning buildings to find, protect, and retrieve those trapped inside
- An autonomous medic retrieves wounded soldiers from an active battlefield and initiates emergency treatment
- A robotic microbiology lab independently plans and carries out experiments to contribute to the corpus of scientific knowledge
- An intelligent web-bot takes the place of human content reviewers, searching for instances of graphic violence, abuse, or pornography

The first two agents described serve not only the victims that they rescue, but also the rescuers in whose stead they go. The robotic lab extends our power of discovery, by allowing productive
work to continue even when no human is present to direct it. The fourth protects human content reviewers who may otherwise suffer psychological damage viewing a constant stream of atrocities (Stone, 2010). For Augustine, the material world exists for the good of humanity. To the extent that we are duty-bound to relieve human suffering, we should use our material resources—including even complex tools such as AMAs—to do so. Nevertheless, we should remember that the chief end of our dominion over the material world is salvation—not merely comfort or luxury. Augustine cautions against “the lust for mastery in earthly affairs” since “he who is avid for mastery surpasses even the beasts in the vices of cruelty and luxury.” (Augustine as cited in Steiner, 2005, p. 118)

**Obligations for responsible forethought**

The line of reasoning outlined above is valid only if we can realistically expect that the AMAs will perform as designed, and that their actual behavior will be in accordance with ethical principles. In other words, we should build autonomous agents only if we can ensure that they will be at least weak moral agents (if strong moral agents, so much the better).

Augustine’s view would appear to be pessimistic on the question of whether we can expect ethical behavior from AMAs. Recall that for Augustine, the source of ethical behavior is rationality—which for him means the ability to see and follow that inner light of truth which is imperceptible to the physical senses. Without it, one would be consigned to act according to the merely physical whims of pleasure and pain. I say “pessimistic,” rather than an outright rejection, because I don’t see any principled reason why a constructed entity couldn’t have that connection to the immaterial and eternal truth. It surely seems far-fetched today to claim that an artificial being could experience genuine spirituality; however less than a hundred years ago, the notion that such a being could exhibit intelligence may have seemed equally unlikely. However, even if the development of “spiritual machines” is possible, it is by no means imminent, as current approaches to developing AMAs largely ignore any possible connection to the immaterial world.

In contrast, the views of Aquinas seem quite amenable the possibility, perhaps even likelihood, of developing ethical behavior in AMAs. For Aquinas, ethical behavior is founded on self-determination and universal reason (recall, “conceptual abstraction”); these allow us to be spiritual. Autonomy is inherent in the design of AMAs; and rational thought is the foundation of top-down approaches. On this basis, AMAs are entirely compatible with a Thomist conception of strong moral agents.

**Obligations to AMAs?**

A final issue of concern is whether we might owe any moral obligations to AMAs that would preclude their creation. From a Kantian perspective, one might ask if AMAs are beings that we would wrong by treating them as a means to an end. From a utilitarian point of view, must we consider the good and the harm that our actions would bring upon AMAs? Tonkens (2009) follows this line of thinking, arguing that any attempt to implement Kantian morality in an AMA would result in a paradox, as doing so would violate Kant’s categorical imperative. His conclusion only follows, though, if the AMAs are both strong moral agents and strongly morally
considerable. In a model such as Asimov’s, where robots are only weak moral agents and weakly considerable, no such violation occurs.

Petersen (2007) addresses the issue as well, asking whether it is permissible to engineer AMAs in such a way that they desire to do the things that humans want them to do. He concludes that it is not unethical to do so, with two provisions. First, he excludes the case of robot slavery, drawing a parallel to the case of specialized breeds of dog. Retrievers, for example, desire to fetch, so it is no violation of their well-being to call on them to do so. Second, he insists that the AMAs in question be designed ‘from scratch,’ to rule out the case of manipulating an existing person’s desires in order to serve one’s own purposes. With these two provisos, he avoids the objection raised by Tonkens (2009). There is no sense of unethically using such an AMA as a means to an end, as “just the fact that [a creature’s] actions are a means to others’ ends is itself no Kantian transgression” (Petersen 2007, p. 48).

From an Augustinian perspective, since AMAs are not moral agents, and hence not morally considerable in the strong sense, they are owed no ethical debt. We are free to build AMAs without fear of violating any rights on their part. For Aquinas, in contrast, since AMAs conform to the defining characteristics of moral agency, we would indeed owe them moral consideration, and could not, therefore, develop and deploy them without regard for their well-being.

**Ethics in the development of AMAs**

In the previous section, an account was given for the permissibility of building AMAs, along with some motivations for why it is desirable to do so. How, then, can ethical principles further guide us in our approach?

**Some prerequisites for moral judgment**

Just as questions of moral status dominated the discussion of whether we should build AMAs, those same questions will weigh heavily in their implementation. Before an AMA can reach decisions that conform to ethical obligations, it must recognize to whom those obligations are due. Thus the percepts of an agent must be sufficient to detect and distinguish morally considerable beings in its environment. A search and rescue robot should have percepts sufficient to distinguish human victims from random debris. An autonomous navigation system should be able to track pedestrians and children at play. A robotic soldier will have to make even more fine-grained distinctions: which humans are civilians and which are soldiers? Which are allies and which are enemies? Which are active threats, and which are incapacitated? The appropriate response will depend crucially on accurate classification.

Effective implementation of moral judgment may also depend on the engineering approach taken. Top-down approaches would appear to more directly support the requirement that moral agents reason about complex moral decisions and justify them, but they also require a substantial investment of human effort to design, implement, and test the requisite large knowledge representation and reasoning systems. Bottom-up approaches are far less labor-intensive, but it remains to be seen whether complex ethical decision-making can arise through the interaction of discrete learning components.
Other aspects of moral judgment, according to Wallach, et al. include:

...emotional intelligence..., sociability, the ability to learn from experience and social interactions, consciousness, the capacity to understand the semantic content of symbols, a theory of mind, and the ability to ‘read the minds’ of others all contribute to moral intelligence. (Wallach, et al., 2008, p. 566)

Certain of these are being addressed in current approaches to machine intelligence, including emotional intelligence (Shibata, Inoue, & Irie, 1996; Picard, Vyzas, & Healey, 2001; Ebrahimijam & Rasamfard, 2009) and social learning (Shibata, Tashima, & Tanie, 1999; Acerbi, Marocco, & Vogt, 2008). Others, such as consciousness and a theory of mind are largely deferred.

**On increasing autonomy**

Artificial agents today are at most weak moral agents, exhibiting autonomy in its negative conception (freedom from external control in the operating environment). We have not yet developed artificial strong moral agents, free to choose which principles they will follow. Should we? Though the idea may seem dangerous, bringing to mind alarmist science fiction tales of machine uprisings, there may be nonetheless good reasons to do so. Ethical decisions are often informed by conflicting moral principles, requiring the decision-maker to choose which to follow. Who should be rescued from a burning building, and who left behind? If an AMA lacks the power to choose, the resulting impasse could mean that no one gets rescued. Of course, in simple cases, sufficient foresight may allow the designer to order the rules to break such an impasse; however it seems doubtful that such conflicts could be avoided entirely in any system complex enough to support robust ethical reasoning. In the rescue scenario, it would be better to rescue anybody than to continue in an endless loop of reasoning while victims perish.

Based on the principles from Aquinas, the push for increasing autonomy is desirable, since, together with increasing rationality, it will enable better ethical choices. In contrast, from an Augustinian perspective, increasing the autonomy of an AMA is undesirable or even impossible, in the absence of the spiritual inner light of rationality. On the one hand it is undesirable, since without reason, a being's impulses cannot be properly governed. It would therefore be irresponsible to turn over ethically consequential decisions to beings who cannot by their very nature act ethically. Furthermore it is likely impossible, since in the absence of rationality, there is no mechanism for autonomy. Irrational beings simply act based on impulses of earthly stimuli. Thus for Augustine, increasing the autonomy of an AMA would only be justified if that AMA had access to eternal truth through rationality. If, then, we do not know how to ‘create’ spirituality in such beings, does that preclude our building them? Not necessarily: we can and do build many things that work even though we don’t understand the mechanism of their working.

**On the treatment of AMAs**

Given that both Augustine and Aquinas made provision for the kind treatment of animals as an indirect duty, it seems likely that they would view AMAs as being at least weakly morally considerable. Several years ago there was a disturbing phenomenon involving people posting videos of burning “Tickle-me Elmo” dolls. The unsettling clips showed these robotic toys engulfed in flames while laughing and saying “Oh that tickles!” Many people found the videos
deeply disturbing—not, I think, because of any imagined pain suffered by the dolls, but rather because of the troubling insight that other people found it funny. The creators of such videos were certainly not “practicing a pitying affection,” and one is left to wonder how they would react to the suffering of true moral patients. As AMAs become more human-like in appearance and social interaction, our treatment of them will reveal, if not reinforce, whatever natural inclinations we have.

Summary

Even though Augustine and Aquinas had no notion of autonomous machines in their worldview, their principles surprisingly seem more generous in their treatment of AMAs than Asimov. Asimov’s robots are weak moral agents, and at most weakly morally considerable, and hence we can do with them as we please, to serve our every whim and pleasure. In contrast from the perspective of Augustine and Aquinas, our obligations to our fellow-humans might encourage the use AMAs—as the rest of the created world—for whatever uses we may have, provided they are in line with an ultimate goal of redemption. Both Augustine and Aquinas would argue that we should only give AMAs autonomy commensurate with their potential for ethical behavior. For Augustine, this depends on their ability to perceive and respond to the eternal truth of the non-material world, making it unclear how or if we can ever build ethical machines. For Aquinas, the potential for ethical behavior depends on autonomy and rational thought, both of which are common in existing approaches to the development of AMAs. Thus for Aquinas, increasing autonomy concurrent with increasing rationality is called for.

As the autonomy of AMAs increases, so will their moral considerability, resulting in a potential paradox: if they rise to the level of strong moral agents, it may be an infringement of their autonomy to engineer them. This is especially salient for an Augustinian view: if AMAs are not strong moral agents, they cannot behave ethically, and hence we should not develop them. But if they are strong moral agents, then they are due ethical consideration, which we may violate in the very act of creating them. At a minimum, both Augustine and Aquinas would indicate that if AMAs become strong moral agents, we would then owe them a moral obligation to consider their interests, both in their creation and thereafter.

If we do pursue the development of AMAs, it will be crucial to give them perceptual abilities sufficient to distinguish the relevant classes of morally considerable entities. Their participation in the community of moral agents will require certain abilities related to their ethical deliberation, such as the ability to justify their actions, as well as other general abilities needed to operate in a moral community. Among the latter are emotional intelligence, sociability, and social learning. Finally, if in the development of AMAs we continue to increase their autonomy, we must be prepared to grant them due consideration in our own moral deliberations.
Sources Cited


