“My Logic is Undeniable”: Replicating the Brain for Ideal Artificial Intelligence

Samuel C. Adams

A Senior Thesis submitted in partial fulfillment
of the requirements for graduation
in the Honors Program
Liberty University
Spring 2016

Acceptance of Senior Honors Thesis

This Senior Honors Thesis is accepted in partial fulfillment of the requirements for graduation from the Honors Program of Liberty University.

_____

James Jones, Ph.D.
Thesis Chair

_____

Mark Shaneck, Ph.D.
Committee Member

_____

Timothy Barclay, Ph.D.
Committee Member

_____

James H. Nutter, D.A.
Honors Director

_____

Date

Table of Contents

Abstract

Alan Turing asked if machines can think, but intelligence is more than logic and reason. I ask if a machine can feel pain or joy, have visions and dreams, or paint a masterpiece. The human brain sets the bar high, and despite our progress, artificial intelligence has a long way to go. Studying neurology from a software engineer’s perspective reveals numerous uncanny similarities between the functionality of the brain and that of a computer. If the brain is a biological computer, then it is the embodiment of artificial intelligence beyond anything we have yet achieved, and its architecture is advanced beyond our own. When striving to achieve the ideal form of AI and revolutionize our computer architecture, where else should one look than the design of the brain? This paper will give a brief history of artificial intelligence, define *ideal AI* by a set of criteria, discuss the philosophical implications and moral issues of attempting to create a synthetic human, analyze the similarities between the brain’s architecture and computer architecture, postulate a design approach for engineering this “ideal AI”, and discuss the applications and consequences of such technology.

“My Logic Is Undeniable”: Replicating the Brain for Ideal Artificial Intelligence

As both a software engineer and a fiction writer, the concept of an intelligent computer with an intellect, personality, and emotions intrigues me. If the brain is indeed primarily composed of networks of neurons whose dendrites and synapses communicate via voltage in the form of logic highs and lows, though separated into various distinct lobes, a synthetic machine of the same form should also be possible. One may object, or rather *should* object, that the brain is also composed of self-modifying and self-repairing biological tissue and utilizes chemical neurotransmitters to regulate behavior. But when implementing such a system, those matters may be addressed with recursive self-modification and logic-triggered functions to simulate the effects to the system. When a software engineer analyzes the various components of the brain and considers their functions, he will observe numerous similarities between the functions of the brain and the various hardware and software components of a modern computer, encouraging the belief that this incredible feat of psychology, biology, and engineering may be possible. Indeed, computer scientists have been using neural networks in computer systems for years now by following the same line of thinking in imitating anatomy, but there seems to be potential for far more than these imitations of biological data networks. May it be that the brain simply utilizes a much more advanced computer architecture than we currently have?

As alluded to with the mention of neurotransmitters and behavioral effects, the discussion here is not merely artificial intelligence in regards to whether a computer system can "think" as Alan Turing asked, but the complete synthetic replication of a person's mental faculties in the physical sense. Turing did dream of building a human

brain, but he seemed to only consider the rational component of it. The question is whether a computer system can be created that not only can think in terms of rational thought, reason, and solve relatively complex problems, but also if the computer can have a personality, experience emotions, think abstractly, understand social norms, comprehend morality, be creative, and any other functions that we as humans can perform with our intelligent minds. Imagine a true artificial intelligence animating a robot or living in and monitoring a network. Could it dream? Would it consider itself alive? How would it view our world? Maybe it would perceive a greater range of colors and music due to synthetic receptors in its "eyes" and "ears" and even create art that we could not perceive. Imagine its unique sense of humor.

Naturally questions of feasibility, morality, proposed methods, practical use, and risks come into the discussion. Can this be done? Should this be done? If it can be done, how would one attempt it? And what would we use this technology for? What are the dangers that could arise? The answers to these questions will all be explored in the following pages. This matter is not one for simple answers; indeed, as is the case with any other problem of philosophy or engineering, there likely are many answers.

In the end, creating such an intelligence may not be possible or even wise, however fascinating it may be. Yet if the brain is truly as similar to a computer as many believe it to be, the very least we could accomplish pursuing this direction is making ground-breaking discoveries in computer architecture that can not only improve but entirely revolutionize the design of our hardware and operating systems.

History

From the imaginations of ancient civilizations captured in their mythologies to the fantasy writings of our modern day authors, the concept of intelligent machines has always fascinated human beings. We tend to cast our own cognition upon other objects and creatures for sake of stories and imagination. Alien, animal, machine, or even some golem of nature itself—can some other entity think and live as we do? Certainly with the invention of modern computers that concept would only be applied to it, but with this technology, such a feat may be possible now with engineering! The only other options our imaginations generally have are the help of magic or gods, neither of which are readily available.

Now, as we educated humans must sound impressive to our peers, we create long, sophisticated names for our academic fields, and the subject of *synthetic* or *artificial intelligence* is no different there. So all the people with a limited understanding of technology and engineering are impressed, believe the matter to be beyond their own mental acuities, and actually assimilate what we do to magic.

The field of artificial intelligence, or AI, began the same time the first modern computers were built around the second world war, though it was not considered an official field at the time. It was Alan Turing who asked, "Can machines think?" Though the term AI was not yet coined, he is known as the founder of AI (Jones, 2009). As an undergraduate at King's College, Turing had invented his universal Turing Machine, a logical system that could theoretically solve any mathematical problem. One Ph.D., the invention of the modern computer, and a world war later, he wished to apply it to a computer system, hoping his Turing Machine could perform rational operations as a

human brain. He believed that uncomputable human intuition was a myth, that all human thoughts functioned logically like a computer. He wanted to "build a brain" to embody human intelligence artificially in a computer. Unfortunately, he committed suicide in 1954 with cyanide before he could work any further on that dream (Hodges, 2014).

Alan Turing is probably best known for his Imitation Game that has since been called the Turing Test. In this game, an interrogator asks a man and a woman questions. A moderator, or some anonymous form of communication, separates the interrogator from his subjects so that all the interrogator receives is their communications devoid of any image or sound that can hint at their identity. One of the subjects attempts to imitate the other one, and the interrogator has to identify which subject is the man and which is the woman. Turing proposes to make a computer one of the subjects, trying to imitate the human subject. Can the interrogator determine which subject is the machine? If the computer successfully tricks the interrogator, it has passed the Turing Test and can be considered a "thinking" machine (Turing, 1950).

Since Turing there have been many other theories and advances made in AI. "The Logic Theorist," the first AI program written for a computer, was developed in 1956 by Allan Newell, Herbert Simon, and J. C. Shaw to find proofs for mathematical equations. Interestingly enough, the program would find new and better proofs to replace proofs that had already been in use (Jones).

Early researchers found games to be useful testing grounds for AI programs. By 1952, AIs for Checkers and Chess had already been developed by Christopher Strachey and Dietrich Prinz, respectively, on the Ferranti Mark I computer, though these early programs took long durations to process their decisions. Later that same year, Arthur

Samuel developed a Checkers AI program for the IBM 701 which he played against a copy of itself. The two AI "players" learned against each other, and by 1962 Samuel's Checkers program had defeated the former Connecticut Checkers champion (Jones, 2009).

During the mid-1950s, AI developed as an official field of computer science. In 1956 the first AI conference was held at Dartmouth where a summer research project was hosted to advance the field of AI, specifically addressing a computer's use of language, abstractions, and problem-solving, and its consequent self-improvement. There, John McCarthy developed LISP, the first AI programming language. Since then, numerous AI conferences have been held internationally (Jones, 2009).

Up to this point, researchers had taken a “top-down” approach of intelligence looking at the higher concepts of the human brain such as reasoning, language, understanding, etc, but in the 1960s, a “bottom-up” kind of approach gained popularity where the focus turns to the brain’s actual functioning, down to the firing neuron circuitry from which neural networks were invented. As these two schools of thought competed, two additional concepts clashed: what researchers called *neat* versus *scruffy* approaches. Neat referred to defined formalizations of algorithms to accomplish AI capabilities, where scruffy is characterized by what researches call “fuzzy” logic. Fuzzy logic seems to accomplish the task, but more by imprecise “tricks” that are not easily explained. In other words, “It works; we’re just not sure how.” Amidst this, two major fields emerged: *weak AI*, claims that at its best an AI can only mimic the human brain, and traditional *strong AI*, ascertains that the machine can essentially *be* a mind (Jones, 2009).

The 1970s until the mid-1980s are considered to be the AI Winter, where excitement and development of AI had died down. However, the field still progressed in that time, one notable development being that of the Prolog AI language in 1972 by Alain Colmeraur and Phillipe Roussel (Jones, 2009). Prolog turns a programmer's head around in regards to other languages with a new way of thinking. The programmer provides it facts, rules to determine the relationships between those facts, and then a query which sets the system in motion, answering the query in especially recursive processes.

AI quietly returned in the late 1980s, but in the form of everyday usage such as camera functions and brake systems. AI is all around us in our devices in programs, yet most people are unaware that they are even using it. AI researcher Rodney Brooks says, "Every time we figure out a piece of it, it stops being magical; we say, 'Oh, that's just a computation.'" This he calls the "AI effect" (Kahn, 2002). In other words, AI is whatever computer scientists have not yet accomplished in intelligent computing.

In recent years, AI research has been booming. Major technological breakthroughs have been made in deep-learning neural networks (Jones, 2014), neural network algorithms implemented in computer chips (Merolla, Arthur, Alvarez-Icaza, Cassidy, Sawada, Akopyan, Jackson, Imam, Guo, Nakamura, Brezzo, Vo, Esser, Appuswamy, Taba, Amir, Flickner, Risk, Manohar, & Modha, 2014), neurological research on the how the brain achieves abstraction, and more (Taylor, Hobbs, Burroni, & Siegelmann, 2015). Applications of this new technology has resulted in the known AI programs such as the iPhone's Siri and Microsoft's Cortana, and even fully automated robots such as Hanson Robotics' Sophia and Hiroshi Ishiguro Laboratories' Gemenoid which can engage in human dialogue, learn and adapt to humans, recognize faces, facial

expressions, voices, and emotions, and even simulate the same to an extent. These new robots are being used to study humans, help with therapy, health care, education, and customer service (Taylor, 2016).

Definitions

I play Chess or any other strategy games by keeping my endgame always in mind, not by focusing move by move, yet I win not only by having that focus but by having a full understanding of how the game works and by playing the system. As Joscha Bach warns AI researchers, "Aim for the big picture, not the individual experiment" (Bach, 2008). I intend to approach AI from a rather different direction than many researchers: along the lines of bottom-up, but with the big picture of its higher capabilities in mind. And instead of rushing into modeling a computer after the brain, I want to first analyze the brain as a computer itself, only one of a more advanced architecture. Furthermore, to be clear regarding the level of intelligence I strive for, I intend to firmly define what constitutes AI instead of referring to inconsistent opinions and vague concepts.

Of course, before one discusses the feasibility, proposed methods, or any other subject matter regarding artificial intelligence, one must define it. Yet in defining *artificial intelligence*, one must first define *intelligence.* As the Merriam-Webster Dictionary defines it, intelligence is the ability to learn or understand or to deal with new or trying situations: the skilled use of reason. In other words, to be intelligent, an entity must be able to learn, comprehend, reason, and adapt. A secondary definition with Merriam-Webster adds the ability to think abstractly (*Merriam-Webster*). With *intelligence* defined, we look at its artificial replication.

There have been many definitions of what *artificial intelligence* is as well as what constitutes it so that a system can be identified as *intelligent*. When it comes to the synthetic intelligence of a machine, the definition expands beyond use of logic and reason. As was already said, the definition of artificial intelligence seems to change as technology increases, per the AI effect, to be essentially what has not yet been accomplished. This inconsistent definition based on current perception would not be acceptable in any other field of study, nor should it be here. To scientifically determine whether or not a system is intelligent, there should be set parameters.

In pursuing a design modeled after the human brain, I propose my own requirements of intelligence. And when I say "my own" I do not mean that no one else has come up with them, but merely that they are of my own choosing without basing them on another's system. It also should be noted that what I intend to define is more advanced than our current systems. We already utilize many intelligent applications on a regular basis from "basic" voice-recognition software to the Siri and Cortana smartphone phenomena to the Nintendo Amiibo that can learn from their opponents until they can beat even professional Super Smash Bros. players. What I define here is what I call *ideal* artificial intelligence, approaching that of the original, fictional namesake of Microsoft's Cortana program.

What should an *intelligent* machine be able to do? Pass the Turing Test? My fellow engineers, a computer may be able to trick someone into thinking it is human by clever dialogue algorithms, but that does not mean it can debate the ethics of capital punishment, experience pain and affection, paint concept art of Tolkien's hidden city of Gondolin, or fathom the deepest workings of a woman's heart, the last of which may

actually entail extensive quantum mechanics. No, there is far more to intelligence than logic problems and language parsing.

We do not tend to consider animals to be intelligent, yet most of them are more capable than our current systems. Humans are scientifically regarded as intelligent, despite the opinions many have about their politicians, bosses, and relatives, but depending on one's worldview, there are some things that cannot be replicated with a logic-driven machine. As intelligent beings, we communicate and coordinate our activities. Animals can too (Flack, 2015). We experience complex dreams when we sleep. Animals can too (MIT News, 2001). We can reason, apply knowledge, and solve problems. So can animals at varying degrees, sometimes even better than the average human. We experience and convey emotion. Any pet owner, especially that of a dog or cat, can tell you that animals are quite emotional (Bekoff, 2000). We can learn and adapt. Clearly animals can be trained, demonstrating the same capability at simpler levels. So what then are the differences in intelligence? What can we do that animals cannot?

Animals may be able to think abstractly, though that is not certain. Animals even have degrees of creativity by both behavior and even intently creating aesthetic art of varying sophistication (Goldman, 2014). Animals may not comprehend human social norms, but they do conform to social norms among their own species (Waal, Borgeaud, & Whiten, 2013). Animals have no understanding of ethics, however, only an awareness of how others react to their decisions and the memories of the pleasant or unpleasant consequences that they apply to future decisions. And animals definitely do not have the self-awareness or free will that humans have, the latter depending on one's worldview; their behavior and decisions seem to be based on action, reaction, logic, emotion, and

learned responses, which could theoretically be programmed as if they are God's own fuzzy, scaled, feathered, and potentially adorable robots. Then there is the matter of human intuition, which Turing believed to be a myth. If the brain functions as a logic-driven computer system, no results can be uncomputable like the supposed leap of intuition (Turing, 1950), short of some kind of quantum shenanigans. The rest of that discussion will be covered in the next section, but for now, where do we draw the line for a machine?

Considering the perceived theoretical possibilities with my Christian worldview, I believe the ideal artificial intelligence falls somewhere between an animal's intelligence and human intelligence. It should be able to accomplish everything an animal can and replicate most of what a human can do, the only physical obstacle then being the hardware itself. The "hardware" of real brains can physically and chemically change, adapting to new knowledge and trained responses by dynamically creating and pruning circuits within its neuron networks. The neurons are also supported by a variety of glial cells that insulate, protect, and maintain them. As of right now, our circuitry of metals, silicon, and fiber glass cannot achieve that level of modifying plasticity and autonomous maintenance. We have basic models of evolvable hardware (Kaufmann, 2013), but until we can replicate the plasticity of neurons, our systems would not be as efficient. And never mind the healing capabilities neurons have; personally, I have never seen a smashed laptop repair its hardware. That said, I declare dynamic circuitry to be currently out of scope as a necessity and state this as my set of parameters, defining a system to constitute *ideal artificial intelligence* by having the following necessary capabilities:

- Morality- an intelligent entity understands ethics and honor and makes appropriate moral judgments depending upon each situation.
- Abstraction- an intelligent entity can think abstractly.
- Self-modification- an intelligent entity is able to learn and adapt. For a computer, this would mean modifying and updating its own software, potentially even hardware.
- Social Intelligence- an intelligent entity learns what is socially acceptable.
- Aesthetics- an intelligent entity comprehends and appreciates aesthetic values.
- Creativity- an intelligent entity can imagine novel ideas.
- Reason- an intelligent entity is capable of rational thought and making conclusions.
- Emotion- an intelligent entity experiences emotion and regulates its expression.

Morality, aesthetics, creativity, and emotion could be achieved at least at sophisticated degrees, even if not to the level of a human, as I will explain in the philosophical discussion. I would like to include intuition as one of the criteria for ideal AI, but until we have some kind of understanding of what intuition is, we have no hope of replicating it. It may be possible to achieve these eight capabilities of an intelligent human in a machine in what I call ideal AI. The feasibility, morality, methods, and application of this proposed feat of engineering is my focus in this thesis.

Philosophical Discussion

Before one considers potential methods of solving a problem, one must consider whether the problem even has a solution in the first place. And if the solution exists, whether it is one that can be achieved with our current mindset and tools or one that requires a radically different approach such as that of the Gordian Knot which, according to legend, was solved by Alexander the Great when he simply cut it with his sword.

Fundamentally, the question of feasibility is not one of psychology or computer science, nor is morality a question of the same. Here philosophy and theology enter the arena, for the immaterial mind and spirit, should they exist, cannot be empirically studied or physically quantified. Can the function of the brain be fully simulated as a completely material, anatomical object or would the synthetic system lack an essential element of the intangible, of the ethereal? If morality is a social tradition, would it be right to create an ideal artificial intelligence simulating a person? And if the cosmos is more than just the material sticks and bones, morality is not a social tradition but a divine mandate—one that cannot be decided by any sciences of man but of rational analysis of that divine mandate, which means the question must be evaluated from a quite different perspective.

Feasibility

A mapping and more sufficient understanding of the brain itself would be a major step to creating an ideal AI, as demonstrated with the breakthroughs in abstraction and neural networks. But this raises the age-old philosophical question: what is the difference between the mind and the brain? The answer to this question leads to the issue of free will versus determinism and also raises the question of identity and self-awareness. For the sake of discussion, I approach these issues from both major schools of thought: that of a naturalist, and that of a theist, primarily a Christian theist.

If everything in the cosmos is material, the human brain is simply one more advanced than an animal's brain (as humans are considered merely higher animals in evolutionary thought), and the mind is the result of the brain's natural logic-driven function as a biological computer. In this line of thought, the mind and brain are essentially pieces of the same physical machine. This leads to another aspect of

intelligence: volition. If the mind and brain are entirely of physical means and driven by logic, as any software engineer knows regarding even our most advanced random number generators, the results are only pseudo-random. There is a distinct output that depends entirely on the seeded input. Given the same seed, the result will always be the same. So the physical brain will always have determined results, and free will is a myth. Every belief, every action, every thought—they are all determined by previous experiences. Significant life decisions can easily be traced to experiences and past acquaintances or even strangers who had influenced one's current decision. Why did Megan choose to eat Cocoa Puffs today? Did not she just happen to choose it? Maybe. Or the decision could be traced to a faint emotion of past experience, convenience due to the fact that it was the first box she saw, or her body influencing her mind for need of chocolate. Even basing a decision off a roll of dice is determined by the physics of how the dice were rolled with what vectors and velocities they were released, their weights, the air pressure and wind resistance in the vicinity, the friction and density of the surface they land on, the force and direction of gravity, and much more. Everything has a cause, even if the answer is ultimately forty-two. Therefore, determinism must be true, and free will does not exist. In such a case, it is physically possible for a computer system to entirely model a human's brain and full range of mental capabilities from reasoning to emotion to imagination and what is perceived as intuition.

From the theist's standpoint, we begin the discussion with the presuppositions that God exists, and from the Christian's perspective, the Bible is infallible. The existence of God and origin and infallibility of the Bible would take up far too much space and are

outside of the scope of this thesis, as much as I would like to cover the subject.[1] The supernatural, by definition, is not governed by the laws of natural, known physics, though I have often wondered if the spiritual realm is of a higher dimension. Therefore, if the spirit exists and the mind and the brain are two separate entities as the Bible mentions, then the mind and brain have a mysterious correlation that cannot be empirically studied and proven with natural science. Yet if that is the case, will a brain work without its immaterial counterpart? Neurologists have identified many physical parts of the brain with high activity levels when one experiences emotions, memory, reasoning, behavior patterns, pain, and most other mental faculties and experiences, and numerous theories exist that address how what those things are physically in the body and how they physically affect us. So provided the mind exists as a separate entity, what is missing when the immaterial mind is removed?

The discussion again moves to the question of volition. If the brain can account for so many things physically, determinism makes sense as demonstrated above, but the Bible clearly accounts for free will. After all, how could God be angry with His people if He or His creation orchestrated their actions? He has the power to do so, but He chooses not to. Why? Love requires a choice. This leads to other theological discussions outside the scope of this thesis, such as the debate regarding predestination and the doctrines of total depravity, unconditional election, and irresistible grace of Calvinism, but for our purposes here, free will exists in this worldview. And if that is the case, is volition an

[1] For that I refer my readers to C. S. Lewis, fulfilled prophecies in history, and the Bible itself, as well as my own argument I like to call the "Cosmeticological Argument" which asks the purpose of beauty that serves no practical function, such as that of the peacock's feathers.

ability enabled by the immaterial mind, vacant when only the material brain exists? Perhaps only the soul is missing, and the brain still functions and can think. But then identity also comes into question, and the philosophical and theological maelstrom continues. Identity is an aspect of self-awareness, something sentient humans have and animals likely do not. Animals are not intelligent, not sentient, but were they created as biological machines lacking free will?

Of course, besides the ideas of material and immaterial in the cosmos there is a third option: everything we see and experience is an illusion, and nothing truly exists except, as Descartes pointed out, our own minds or we would not be able to even think about this issue (Descartes, 1641). In fact, we would be able to do nothing, as our abilities themselves would not exist. But hopefully we can agree this third option is foolish and unproductive. First, what does it matter if we cannot prove our existence if we are perceiving these things? It does not change how we live life. If I do not eat, I will die. That is clear enough. Whether or not it is an illusion, starving remains rather unpleasant. Similarly, the *Matrix* option, called the "Brain in a Vat" argument, does not help matters either. For if I am a brain in a vat, only experiencing simulations of life, then who put the brain in the vat? Could not that disturbing scientist also wonder if he too is a brain in a vat? The cycle could be infinite, for why should it stop somewhere? Either way, such arguments of Skepticism are entirely useless to our discussion.

Another issue is an AI's understanding of morality and ethics. Do these concepts exist? We all can agree they do, even if a naturalist believes them to be social tradition and a theist believes them to be divinely instituted within us in what C. S. Lewis calls Natural Law (Lewis, 1952). The key here is that morality and ethics (should) govern

one's decisions. So could engineers implement such in an ideal AI? Sure. The simple answer is that one could hard code them in as part of the logic system in decision-making, as is the idea with Isaac Asimov's three laws of robotics. The more realistic answer is that with the myriads of complex applications throughout everyday life and the conflicting or seemingly conflicting moral dilemmas, a high level of abstraction must be achieved to comprehend morality and make sufficient decisions based on it. But all applications would be based off of core hard-coded principles that cannot be reasonably modified. The only problem I can foresee with this, as many robot apocalypse movies and books illustrate, if an ideal AI can modify and update its own software, it could potentially find some novel way to override and overwrite the literal code of ethics programmed into it, misinterpret the laws after logical processing, as V.I.K.I. does in Asimov's *I, Robot*,[2] or somehow bypass them altogether. After all, as any hacker can attest, nothing is completely secure.

Morality may be limited as just described, but it could be further limited for another reason, as well as aesthetics, creativity, and emotion. Naturalistically speaking, without the spirit, we are right back where we were. If the brain can do it, it is physically possible, however difficult, for a computer system to achieve as well. But when an immaterial spirit is considered, how much does it influence morality, aesthetics, creativity, and emotion? If morality comes from God, can it not be fully comprehended without a spiritual awareness of God and His Natural Law? If God and His handiwork is the foundational standard of beauty, can one truly appreciate and evaluate it without a

[2] Hence the title of this paper, taken from *I, Robot*, the film adaptation of Asimov’s *I, Robot* (1950) and *Caves of Steel* (1953). “My logic is undeniable,” USR central AI V.I.K.I. repeats, justifying her actions as she dies.

spiritual component? If all emotion originated with God and is a major connection among us, how much does the spirit influence it? And if creativity relies on aesthetics, emotion, and free will, can it be fully developed in a synthetic entity? These questions cannot be empirically addressed, and our knowledge of the subjects is incomplete. Yet even if they are a problem, they do not completely prohibit an AI from having the above; they would only limit the scope of such abilities. The brain still achieves much of them in its physical processes, spirit or not, as will be addressed later.

When the stove is left on and all the arguments boil away, we are left with a crispy, worldview matter stuck to the pan. A naturalist would believe that the brain is merely a biological computer that can be replicated in its entirety by synthetic means. A Christian would believe that the mind and brain are separate entities, and that a computer may always be missing an essential piece without the immaterial mind and soul. A naturalist would believe in determinism, that a machine could never have free will. A Christian would believe humans have free will, but that a machine never could. So can a true, ideal AI be created? Yes, I believe it can. But to what extent? According to a naturalist, there is nothing a synthetic intelligence could not possibly achieve that humans can; in fact, it could potentially perform better than humans. According to a Christian, an AI's ultimate capabilities lie between that of an animal and a human, as free will and self-awareness are sold separately. Perhaps free will is the true underlying difference between intelligence and personhood.

Morality

It seems after all that creating an ideal AI is possible, even if it lacks free will and self-awareness. Maybe our computer scientists could never produce the original Cortana

and create a synthetic person, though we could come close. Yet now another question arises as it does with the concept of cloning. Is it moral to create a machine that close to being a human? As any athlete knows, just because you *can*, does not mean you *should*.

The naturalistic standpoint here is relatively simple. Basing this approach off the presuppositions stated in the above section of this being a material world only, if our world is an accident, whether it has always been or originated in a cosmic explosion, logically there is no higher meaning to who we are as a species. No higher power governs our actions. No afterlife awaits to judge our actions. The purpose in life is what you make it. So what difference does it make if we try to build a synthetic human? Would it not be the next step in evolution?

It gets a little stickier on the theological side, specifically that of a Christian. If man was made in the image of God, we are distinctly separate from all other creatures. Is it right to attempt to create our own man? When faced with that objection to his work, Alan Turing (1950) did not believe so. Though he clearly stated that he did not accept anything regarding the immaterial spirit or man made in God's image and believed all things regarded as spiritual to be merely things yet unknown by science, he said, "I am unable to accept any part of this, but will attempt to reply in theological terms." He says that as creating an AI, we "should not be irreverently usurping His power of creating souls, any more than we are in the procreation of children: rather we are, in either case, instruments of His will providing mansions for the souls that He creates." Basically we are not attempting to take God's place. We would merely be creating it, and if He wants to put a soul in it, He can (Turing, 1950).

I do not think the problem is whether or not we're attempting to put a soul in it, since clearly we are incapable of doing so, but I agree with Turing. We have already discussed the differences between man and animal. An evolutionist would say there is not much of a difference where a creationist would disagree. But from both worldviews, animals have levels of intelligence and emotion and have no soul. And as already stated, without an immaterial component of their existence, they would be logic-driven by cause and effect, lacking a free will as well. If the image of God primarily means possessing a soul and free will, whether such things do not exist or only belong to man, both sides would agree that our AI could never have them being material only. If we were to somehow create an AI with a free will and a soul in the image of God, that would be an attempt to usurp the place of God, but as stated, that is not within our power. Therefore, regarding the issue of an ideal AI's likeness to man, it is moral to create one. Is it wise regarding the potential consequences? That is for a later segment of this thesis.

Methodology

We have established that the creation of an ideal AI, as previously defined, is feasible. Now what? I will approach this problem using the brain's physical design and functionalities as a model for the intelligent machine. It is much easier to show that something has been attempted than that something has not, but to my knowledge, I have not seen this method attempted to such an extent. There is inevitably the point where an engineer takes the anatomical mystery and simplifies it into a basic synthetic model, but how far has that point been pushed in the past? We now have the neural network-based computer chip TrueNorth (Merolla et al, 2014), but I envision us taking this approach even further. As an artist draws what he sees and not what he *thinks* he sees, I pursue the

real design of the brain as far as I can with our current technology to achieve its capabilities before simplifying any of its living components, except the living aspect itself. Perhaps one day genetic engineers may create biological circuitry, but for now the brain tissue must be replicated with metal and silicon.

Attempting to design an AI according to an advanced computer architecture of the human brain depends on a thorough understanding of the intricately complex workings of the brain, of which we still know very little. The Human Connectome Project continues, though comparing their progress to the brain as a whole they have mapped out a few grains of sand within the beach of the human brain. As the project nears completion, we may discover new revolutionary methods of computer architecture. Methods of storing and retrieving data, retrieving and processing instructions, input/output methods and more might be revolutionized based on the workings of the brain. Based on this new architecture, we may be able to create computer systems capable of recursive self-modification, emotion simulation, social intelligence, resulting simulated personalities, and everything else necessary for an entity to be truly intelligent. So how does the human brain and body accomplish these capabilities, and how would we engineers be able to accomplish them artificially?

Brain versus Computer

Step one will be comparing the human brain to our current computer architectures. Our greatest obstacle in this endeavor is our current lack of understanding of how the human brain works. Therefore, the vast majority of the following analysis is significantly hypothetical, as it is based on hypothetical conclusions from necessarily imprecise research of the brain's operations. A lesion in the left hemisphere of the parietal

lobe causing a drastic reduction in language performance in the patient does not allow a researcher to claim with complete certainty that language is uniquely a function of the left hemisphere of the parietal lobe (Kolb and Whishaw, 2011). With the myriads of complex interconnecting and constantly modifying neurons throughout the rest of the brain and all its unaffected regions, many additional potential variables are present. Consequently, the key words of this highly theoretical section will be the word "seems" and its synonym brethren.

Regional comparisons by function. First I will discuss major sections of the brain, comparing their function to that of various computer components at a very high level. The central nervous system is divided into three main sections: the spinal cord, brainstem, and forebrain. The spinal cord is out of the scope of this paper, though it is basically the neural highway, so to speak, of all information passing from the brain to the body and vice versa regarding behavioral instructions, physical status reports, and the like. Yet if we were to compare it to a computer system, it would be the bus, the central data carrier.

Brainstem. The brainstem is divided into the hindbrain, midbrain, and diencephalon (which is also often argued by anatomists to be part of the forebrain instead). The hindbrain controls motor functions, both voluntary and involuntary, such as breathing, balance, and fine movements, while receiving information from the same systems, and the midbrain receives visual and auditory sensory information. The diencephalon includes the thalamus, which is a kind of relay station integrating and passing sensory data to their appropriate forebrain structures by way of corresponding

neurons, and hypothalamus, which controls the pituitary gland that regulates behavior via chemical hormones (Kolb and Whishaw, 2011).

Brainstem comparison to computer architecture. The brainstem as a whole seems to function like a sensory Basic Input Output System (BIOS) and operating system kernel. More specifically, the hindbrain would be a motor BIOS and kernel, the midbrain would be a visual/auditory BIOS, and the diencephalon seems to more like a series of functions on the kernel level, controlling various bodily functions, both by involuntary operation and by interpreting and integrating commands from the other locations of the brain, while taking "system calls" for Input and Output (IO) and returning data.

Forebrain. Anything regarding complex thought, analysis, abstraction, and every other conscious experience is a result of the neural activity in the forebrain. Our primary computing unit here, the forebrain is composed of the cerebral cortex, basal ganglia, and limbic system.

Cerebral cortex. Our general thinking occurs in the cerebral cortex which is further divided into several lobes: frontal, temporal, parietal, and occipital. These lobe locations are generalized, but at a high level we can say that the occipital lobe handles visual input and processing; the parietal lobe addresses tactile functions; and the temporal lobe deals with visual, auditory, and gustatory functions, though these three lobes still perform quite a bit of thinking in terms of analysis of their corresponding inputs and specialties and also assist with memory (Kolb and Whishaw, 2011). The temporal lobe specifically is believed to be key in forming long-term explicit memory (Kolb and Whishaw, 2001). The frontal lobe is considered the executive of the four because it is the central unit where most thinking occurs. It integrates sensory information, motor control,

analysis, and reasoning to enable our conscious thought, planning, and action. Within the frontal lobe, the prefrontal cortex especially handles reasoning, planning (Nevid, 2012), short-term memory maintenance, and other high-level functions (Kolb and Whishaw, 2011).

Basal ganglia. Along with the thalamus, the basal ganglia controls voluntary movement. Short of incorporating this artificial intelligence into an actual robot shell to function as an autonomous, physical entity, this structure is not essential to our discussion (Kolb and Whishaw, 2011).

Limbic system. The limbic system's functions are those of emotion and memory processing. The hippocampus and cingulate cortex seem to have memory functionality, and be involved with the rewarding effects of psychoactive drugs. Specifically the hippocampus processes visuospatial memory, among others. The Amygdala is the brain's center of emotion (Kolb and Whishaw, 2011). This brain structure receives neural inputs from physical or mental changes, interprets stimulants into specific emotions, and stimulates the hypothalamus for behavioral output. It works both ways: thoughts can result in emotions which affect physical bodily changes and behaviors, and physical changes can result in emotions which affect thoughts (Kolb and Whishaw, 2011). It also is the anchor for emotional memory functionality, creating neural circuits across memory components throughout the brain to connect them emotionally (Kolb and Whishaw, 2011).

Forebrain comparison to computer architecture. As it looks from a high level perspective regarding architecture, the forebrain contains most components of our computer architecture from CPU and main memory to cache, RAM, disk, and even its

own internal bus. It seems as if any necessity of registers is superseded by the complex network of neurons that are already easily accessed within the cerebral cortex. Specifically, the frontal lobe is like a processor, I/O, and a kind of cache or even RAM that runs its own refresh cycles.

With this in mind, my thought strays from its course as intuition leads me to an interesting theory. When a patient is clinically dead for too long a duration before resuscitation, say from cardiac arrest, the patient's brain suffers an ischemic injury due to the lack of oxygen it needs for that extended duration, potentially resulting in severe mental damage and even a vegetative state. Perhaps an additional cause of such mental damage is because when the brain is not receiving the resources it needs, not only does it lack oxygen, but consequently it lacks electrical impulses required to run "refresh cycles" as computer scientists would call it. The data would then become "corrupted" if not completely lost, and the affected portion of the brain would no longer function properly.

But returning to the brain-computer analysis, the parietal lobe seems to be another processor for its specialized data, that being: logic, spatial navigation, divergent thinking, etcetera. Likewise, the temporal lobe seems to be a processor for its preferred visual, auditory, and gustatory data, yet it also seems to be responsible for longterm explicit memory, as if it is the brain's disk, its neural hard drive. As with its colleagues, the occipital lobe seems to be a visual processor. These cerebral cortex structures also have their own localized memory for their specialized data types. An additional structure within the forebrain is the corpus callosum which is an interconnecting web of neurons across the center of the brain connecting the hemispheres (Nevid, 2012). In a computer

architectural sense, the corpus callosum is an oddly-shaped internal bus to traffic data across the brain.

The limbic system seems to be another kind of processor and memory storage for its data types. Specifically, the hippocampus seems to be another specialized memory unit, like a partition of the neural hard drive dynamically allocated for visuospatial memory, which would explain why traditional London taxi drivers tend to have significantly larger posterior regions of their hippocampi than the average individual (Kolb and Whishaw, 2011). Imagine a computer that can expand its physical memory storage units as needed to accommodate more data. That's what the brain does: as more space is needed, more space is formed. The amygdala is an emotion processor, control unit, and memory storage. It is its own CPU, specifically allocated for emotion data.

Cognition. Now we compare the intangible functionality of the brain's cognition and memory to that of a modern computer. There are many operations and concepts that psychologists call by one name that computer scientists merely call by another name. Let us begin with cognition.

The association cortex and its composition. Cognition is the combined result of the interconnected lobes in what is often called the association cortex. Part of the cerebral cortex, this cortex is composed of the frontal, temporal, parietal, and occipital lobes. It is still a mystery how exactly our state of consciousness rises from this complex web of neurons in the cerebral cortex, but various capabilities can be traced to different locations within it. For example, temporal planning, attention, and communication can be traced to the frontal lobe, with verbal and nonverbal communication in the left and right hemispheres of it respectively, and drawing, puzzles, and spatial navigation can be traced

to the right parietal lobe, while language and mathematics can be traced to the left parietal lobe. Even our basic forms of thinking originate here. Divergent thinking is produced by the frontal lobe, and convergent thinking also seems to be an ability provided by the parietal lobe (Kolb and Whishaw, 2011). Comparing this to a computer gets even more fuzzy. Cognition would be almost entirely produced by software. It is as if numerous sensory programs and highly complex analysis algorithms are somehow combined, resulting in a conscious, thinking entity. If the immaterial mind exists as separate from the physical brain, this is a prime location where it would leave its fingerprint.

Asymmetrical function. The association cortex is asymmetrical in function. Though the ability resides in both hemispheres, visual senses seem to be specialized by the right visual field and left hemisphere, and auditory senses seem to be specialized by the left ear and right hemisphere. Motor functions are also controlled by the opposite hemispheres: left hemisphere for right side of the body and vice versa.

Hardware effects. When analyzing the brain physically, causes for varying degrees of intelligence is still difficult to ascertain. One observation has been of the brain's "hardware" so to speak: the glia to neuron ratio. Einstein's brain had a higher glial density than the average brain (Kolb and Whishaw, 2011). Why that would improve brain performance will be covered below when we discuss the brain's hardware.

Hierarchical structure. Cognitive functions are based on a hierarchy beginning with the lowest sensory relays and processing and then climbing to consciousness and ultimately reaching the most abstract thoughts. This hierarchy of abstraction is a combined result of the brain's entire connectome structure, beginning at sensory cortices

and extending to the most remote structures. Grouped sensory inputs are processed in parallel by various, globally-connected subnetworks escalating integrated functionality through deeper levels of cognition into the entire connectome to achieve higher levels of abstraction (Taylor et al, 2015). Essentially, this hierarchical structure of cognitive processes achieving abstraction is the same method used in deep-learning neural networks we use today in AI software such as the iPhone's digital assistant Siri (Jones, 2014).

Memory. Memory is more interesting as we see more similarities in processing and storing data. Information is encoded and compressed for more efficient storage (Nevid, 2012); we do the same thing for computers. Even what might seem to be a simple piece of data is stored using highly complex algorithms. The smallest decimal number would be stored as a floating point data type, partitioned into a long string of binary digits in three sections to represent the sign, a multiplying exponent of 2^x, and the mantissa coefficient—all to be efficiently stored in the computer to a high degree of precision.

Sensory memory. The brain receives sensory input which is then encoded into logic HIGHs and LOWs in the forms of electrical currents carried by the neurons. This is the same way our computers receive and process input data and carries it to its correct location. This data is temporarily stored what psychologists call the Sensory Register (Nevid, 2012), a neural functionality our computer scientists would recognize immediately as a buffer pool. Why is it so difficult for the body to receive too much sensory information at the same time? Why does it miss some things or generalize some sensations? Buffer overflow. It cannot receive too much data at one time and process it

all sufficiently. Even generalizing visual or auditory information is like a computer trying to make up for lost network packets.

Short-term memory. Short-term memory is an interesting concept to look at as well. Our short-term memory seems to work like a computer's main memory, RAM, and cache. Perhaps the brain loses various stored entities by overwriting them when it needs space for different data, possibly using a least-frequently-used heuristic, but we will address that below in regards to forgetfulness. *Chunking* is a term psychologists use for remembering amounts of information in smaller groups or "chunks" (Nevid, 2012, p. 201). Computers do that. Computer scientists call it *hashing* and *indexing*. The indexing I will address more specifically later. What psychologists call *maintenance rehearsals*, computer scientists call *refresh cycles*. One is repeating information in the mind (Nevid, 2012), while the other is sending electrical current through volatile memory storage, but they both are keep short-term data from corruption and loss. These are forms of error correction, without redundancy or the Hamming error code or other such methods, though I wonder if neurologists may find the brain using similar methods to retain accuracy of data.

Long-term memory. Long-term memory also follows many similar functions that our computers use. Long-term memory is conceptually a computer's disk memory. To store data more efficiently long-term, the brain shrinks it into smaller representative forms. Psychologists call this *consolidation* (Nevid, 2012); computer scientists call this *compression*, hence the compressed zip files we use every day. Memory retrieval seems to be based on referencing certain entities, such as specific objects, concepts, emotions, and other stimuli. This is a process psychologists call *priming* (Kolb and Whishaw,

2011). Computers use these too, except we call them *keywords*. Priming is a searching method the brain uses, the same way our computers search for data with keywords. Since the brain stores similar information in the same general locations according the data type and even further by concept, it seems to use what computer scientists use in memory hierarchies, called *locality of reference*, to increase the *hit ratio* of finding the desired data. Utilizing locality of reference increases data retrieval efficiency by significantly reducing seek and access time in storage devices. In the brain, it means it increases the chance of recalling a memory and decreases the time in doing so.

An additional functional comparison would be that knowledge, perceptions, behaviors, and processes that are "hard-wired" into us, so to speak, such as those determined by gender differences, other genetic factors, or (depending on world-view) awareness of God and natural law, etc., would be the brain's read-only memory (ROM). If this were a robot, such things as Asimov's three laws of robotics would be ROM.

Forgetfulness. Occasionally there are things we would rather forget, but typically forgetfulness tends to be inconvenient and even get us into trouble. How does forgetfulness work? Generally speaking, it is when the neural connections that kept that data stored in memory have deteriorated, but why and when does that happen? Of course, psychologists do not fully understand it, as they do not fully understand anything about the brain, but they have many theories as always. As is the theme in this paper, these theories also conveniently follow computer principles, specifically heuristics used in buffer pools, paging, and other software functions. Decay theory states that various memories and abilities are lost over time from lack of use, which makes sense as the brain prunes its neural network, removing unused connections. This explains why a

language one learned in childhood and never used in adolescence or early adulthood would be mostly forgotten by early or middle adulthood. Losing memories from lack of use follows the least-recently used heuristic.

Interference theory is the belief that memories are lost when they interfere with other memories, ie. overflow the buffer pool. The primacy effect is when information is recalled better when it is first learned (Nevid, 2012); this follows the last in, first out (LIFO) heuristic because the information that entered memory first is retained while information that entered memory last is the first to be replaced. The recency effect is when information is recalled better when it was learned last (Nevid, 2012); this follows the first in, first out (FIFO) heuristic because information that entered most recently is retained while the oldest information is the first to be replaced.

Retrieval theory follows that memories are forgotten due to either an initial encoding failure when first processing the memories or from a lack of retrieval cues (Nevid, 2012). From the computer perspective, an initial coding failure means the data was improperly stored, possibly missing some bits, and consequently corrupted. A lack of retrieval cues means that the retrieval cue, or keyword, has not been found yet.

Amnesia has a variety of causes, whether psychological causes such as suppression of trauma or physical such as a direct injury to the part of the brain that supports the type of memory in question. Looking at the brain as a computer, amnesia is either corrupted data from a system error and/or denied permissions, or it is partial data or lost data from direct damage to the hardware.

Memory storage location. One of the greatest architectural differences between modern computers and the brain is the fact that memory is stored within the neuron

circuitry near corresponding brain components. It is as if a computer's hard drive was somehow spread throughout the bus and processors, with various data types localized to processors specifically reserved for them. Naturally there would be advantages and disadvantages to this. System calls would be almost non-existent for disk access, hit ratios would increase, and the memory hierarchy would be revolutionized with spiking hit ratios due to built-in locality of reference. Many new problems would also necessarily arise. Sophisticated control units would be needed to determine where various types of data and input would go, modifications to memory storage would be more tedious and difficult, and the integrated memory storage be difficult to accomplish, besides other complications not stated. The brain can do this more easily with its plastic, living circuitry, hence the next topic.

Hardware. Finally we reach the brain's living hardware. This fascinating difference to our own computers is also one of the greatest obstacles to replicating the brain in all its complexities. The brain consists of myriads of neurons and glial cells. Neurons are the interconnected circuits of the brain, receiving multiple (theoretically infinite) electrical inputs through its dendrites across the synaptic cleft and somehow condensing these inputs into a single output through the axon into another neuron. The inputs from the dendrites determine whether or not the neuron is "excited" sufficiently to fire, resulting in an output of either firing an electrical impulse at the synapse or not firing one, hence the communication via logic HIGHs and LOWs like the logic design of our own computer circuits.

Neurotransmitters. Neurotransmitters are chemical compounds transmitted from one neuron to receptors in another neuron across the synaptic cleft. These also cause the

excitatory or inhibitory actions for neurons from chemical reactions that electrical inputs would otherwise. Not only are they activated from axon to dendrite but also from dendrite to dendrite, axon to extracellular fluid (no particular target neuron), axon to bloodstream, axon to neuron cell body, axon to another axon, or even axon terminal directly into another axon's terminal (Kolb and Whishaw, 2011). From a perspective of logic design, these neurotransmitters seem to act like logic gates at varying locations along the circuit. Axon to axon, for example, could be like an exciting or inhibiting NOT gate near the output.

Types of neurons. There are three kinds of neurons: sensory neurons, interneurons, and motor neurons. Sensory neurons transmit data from the body to the brain, functioning as input. Interneurons transmit data among neurons within the brain, functioning as data processing. Motor neurons transmit data from the brain to the body, functioning as output. Sensory and motor neurons tend to have several dendrites and long axons for the sake of carrying data throughout the body, but interneurons have numerous dendrites and relatively short axons to communicate as dense networks throughout the brain (Kolb and Whishaw, 2011).

Glial cells. Glial cells transmit no data whatsoever; they are the support cells. Where neurons are generally limited in number and rarely replaced once they die, glial cells are constantly replacing themselves. There are five types of glial cells according to their structure and function. Small, ovoid ependymal cells line the walls of ventricles in the brain cavities, secreting cerebrospinal fluid (CSF). CSF flows through the ventricles and fills cavities, acting as cooling systems, cleaning systems, and shock-absorption. Star-shaped astrocytes provide a blood-brain barrier to prevent blood and toxins from

damaging the brain while accepting certain nutrients and secreting healing chemicals to the neurons. In the computer hardware sense, astrocytes are structural support within the device, maintenance, filters, and are a part of the cooling system. Small microglial cells are part of the immune and healing system. To a computer, they would be a cleaning system and provide reparations to damaged hardware. Oligodendroglial cells wrap around and separate multiple neurons' axons and absorb and release chemicals for the neuron. They would be providing structural support, maintenance, and insulation to the computer. Schwann cells wrap around neurons to provide protection and insulation and absorb and release chemicals for the neuron (Kolb and Whishaw, 2011); these cells are basically biological silicon circuit insulation or wire jackets that not only insulate and protect the circuit but also provide internal maintenance. Increase in glial cells increases insulation, protection, maintenance, cooling, and ultimately the performance of the networks, hence the theory of Einstein's higher glial density increasing his intelligence.

Plasticity. The greatest difference in hardware is its plasticity. Neurons are constantly modifying themselves. As you read this right now, the neural networks in your brain are changing as they process the incoming sensory data, process the words and syntax you are reading, analyze the information gathered from it, and make conclusions—all the while further processing older stored data and doing general maintenance. Neurotrophic factors are produced by neurons and glia and affect neurons to grow dendrites and synapses. One such trophic factor, aptly named the *nerve growth factor* (NGF), stimulates neurons to grow dendrites and synapses. There are seven primary things to note about plasticity: change in behavior reflects neural change in the brain; all nervous systems are plastic in the same general manner; plastic changes are

age-specific; prenatal events can influence plasticity throughout life; plastic changes are brain-region dependent; experience-dependent changes interact; and plastic events are not always beneficial (Kolb and Whishaw, 2011). These plastic properties of neurons further complicate the research of them and is a feature of the brain's physical design that computer scientists cannot yet replicate.

Design Approach

Step two would be actually designing the intelligent machine itself. Other than the severely fuzzy starting knowledge of our conceptual analysis, one of our greatest obstacles to the synthetic design is our lifeless, manufactured circuits and computer chips lacking biological plasticity as opposed to the brain's self-modifying and healing, living "circuitry".

Overall architecture. In principle, the four basic computer functions are data processing, data storage, data movement, and control (William Stallings, 2013), the entirety of which is accomplished by the neurons at the same time. Data processing is done depending upon the type of data in various lobes, especially the forebrain, and data storage is accomplished throughout the entire brain. That means most of the data movement is already accounted for since the data storage is readily available within the networks. The only significant data movement is done either across hemispheres or in the brainstem and in the rest of the body via sensory and motor neurons. Control is accomplished mostly by the neuron networks in the frontal lobe.

Regarding the actual computer architecture, the primary components are the central processing unit (CPU), main memory, Input/Output (I/O), and a system interconnection for data flow such as the system bus. The main components of the CPU

are the control unit, arithmetic and logic unit, registers for internal storage, and CPU interconnection or essentially an internal bus. Within the control unit is the sequencing logic component, control unit registers and decoders, and the control unit's internal memory (William Stallings, 2013). Now look at the brain. At a high level, the brain stem is the I/O and part of the bus, the rest of the bus being the spinal cord, and main memory and the CPU are combined in the forebrain. The forebrain's internal bus is the corpus callosum; it's internal memory is stored throughout it; and the control unit and arithmetic and logic unit are within the cerebral cortex, outsourcing emotion processing and internal memory to the limbic system and routing sensory processing and memory to their appropriate lobes. To go into further detail, we would have to have a better grasp of how the brain computes mathematical expressions and basic binary logic operations.

Software design by function. So how do we begin to design the ideal artificially intelligent machine? As an undergraduate student without the time, funding, or expertise that many prior researchers have had, I am not about to single-handedly solve all of the problems of the last sixty years and pull a revolutionary design of a new, ideal AI out of a hat, but I will postulate loose approaches to a new design.

Physical architectural concepts. We could attempt to store memory throughout the system, integrating it with the processors and the bus, even storing type-specific data with its designated processor. With some sudden surge in technology, we could implement plastic, self-maintaining and modifying circuitry with physically dynamic memory storage. But though that may improve performance, it does not mean our computers could tell when we are angry about the internet dying again. We now have a decent idea of what parts of the brain we can trace various cognitive capabilities to. But

we still have little idea how these neural structures actually produce the results that they do. Yes, in a very loose sense cognition is a result of the interconnected data processing of the structures in the association cortex via a hierarchical system of neural networks, but we are still not sure specifically *how* the association cortex accomplishes this. Deep-learning is the closest we have come so far in grasping this advanced functionality of the brain, and with further research and testing, we may crack it. That means we are back to the drawing board for now, though we have some new and fun tools to use now.

Morality. Morality would be hard-coded into the system: hierarchical control functions that cannot be overridden. Despite the ability of the AI to modify and update itself, there would be certain functions that it would not have permission to change. I would take Asimov's three laws of robotic, edit and expand them, and hard-code them into the AI. With the ability to think abstractly, the AI could have a high understanding of morality and honor and intelligently apply them to its behavior.

Abstraction. Abstraction combines learning, reason, emotion—everything. Deep-learning neural networks seem to be the key here. This will be our capstone to the system after accomplishing the rest of these capabilities since abstraction is a complex state that must be built upon many other intelligent activities and processes. In recent years deep-learning neural networks have been showing significant leaps in technological advances and great promise for the future of AI. Layers learning take place by observing individual facts, identifying those facts, linking facts together by basic comparisons, and ultimately learning more complex concepts at ever deepening levels of analysis (Jones, 2014). As already stated, the brain works in a similar fashion, except in a more advanced, efficient, complex, and integrated degree, running multiple subnetworks in parallel at deepening

levels of cognition to achieve higher levels of abstract thought. I believe as we learn more specifics about how the brain accomplishes this, we will advance in our design and implementation of deep-learning neural networks.

Self-modification. I'll approach this systematically. By my earlier definition, an ideal AI is capable of self-modification, reason, abstraction, emotion, creativity, aesthetic appreciation, social intelligence, and moral understanding. In regards to self-modification, there are a few aspects to this. We already have AI software that can learn, adapt, and improve. But in a larger, more autonomous system, this advanced AI could actually modify and update its own software, possibly even hardware if the technology becomes available. As we previously discussed, the brain's cognition works off a hierarchical structure, and researchers have only scratched the surface of deep-learning neural networks. This ideal AI could be designed the same way yet by a structure of programs. A hierarchy of programs could be implemented, developing this system with the meta-programs that can modify other programs as needed with varying permissions levels. The trick would be how sophisticated it would be to know how to write its own code to reprogram itself, but at high levels of abstraction, it could conceivably be accomplished.

Social intelligence. Social intelligence could be theoretically achieved with complex algorithms analyzing human behavior and comparing it with previously known data. Concepts of universal human psychology could be taught to the machine by default for it to apply circumstantially as it learns and develops. If animals can adapt to human behavior, an AI should have that capability too. In doing so, however, we must ensure moral guidance with further social checks and balances; otherwise we would repeat the

mistakes of Microsoft's Twitter AI "Tay" that in mere hours, after learning from the worst of humanity on the internet, became a genocidal, Hitler-loving, promiscuous racist (Ohlheiser, 2016).

Aesthetics. Aesthetics would be tough. It is difficult to make objective judgments on aesthetic beauty as it is an internal part of us. Here we would have to turn to medium-specific attributes. In visual art, for example, developers would need to apply color theory, color harmony, how colors and artistic techniques and various colors and temperatures influence emotions, symmetry, representation and deeper meaning, and numerous other attributes. Whether visual art, music, or whatever the medium be, there are objective qualities of aesthetics that remain constant. The analysis of such, of course, would rely heavily upon high-resolution cameras, but the aesthetics is not merely a matter of objective analysis. Intuition and emotion play heavy roles, and every beholder values aspects differently, though there be a common thread. Other than a sophisticated, objective analysis of technique and overall harmony by medium, I cannot even begin to imagine how I would implement aesthetics in a program, but somehow those universal rules would guide the system's perception of entities in regards to beauty.

Creativity. Creativity is a product of emotion, motive, and inspiration. Emotion would already be accounted for, motive depends on the task at hand, and inspiration comes from the abstract analysis of related concepts and base knowledge. Combine these with clever algorithms and a simplistic sense, creativity is achievable, but it would be a late development as it relies on other high-level processes first. Obviously this is not as simple as this high-level description seems. Enabling a created machine to develop novel ideas its creators did not even have would be an incredible feat, and the sheer complexity

of trying to integrate the emotion, motives, and inspiration through the highest levels of abstract thought would make any researcher's head spin, and all these other high-level dependencies must be developed first.

Reason. Reason has long been accomplished to impressive degrees, even if it ultimately lacks human capability involving intuition. Computers may never have "gut feelings" as we call them. But we can still develop powerful reasoning AI systems. How does a machine reason at the fundamental level? "When the system is required to do something that it has not been explicitly told how to do, it must reason - it must figure out what it needs to know from what it already knows" (Barr, Feigenbaum, & Cohen, 1981, p. 146). Prolog reasons, and we have progressed quite far since then. Give it facts, rules that govern the facts, and ask a question. By recursive logic, it will reason through the problem, even if at a very simple level. At today's level of sophistication, we have reason covered to an extent, but we can reach further and strive to reach the sophistication of the brain.

Emotion. Emotion has long been overlooked as a part of AI, but it is highly significant. Even if the computer does not have emotion itself, it would be a major advantage for it to be able to recognize it. Emotion-recognition software could be helpful for research, better addressing needs, and assisting in psychological disorders, among other possibilities. The ability to recognize emotion would be imperative to any interactive AI. Emotion can be identified by facial expressions, tone of voice, verbal communication, behavior, subtle body language, and biometrics, all of which can be identified by AI software we already have with great precision (Kleine-Cosack, 2006).

It would also be very helpful for a computer to experience emotion. First of all, emotion is necessary for creativity, but more importantly, it is vital to decision-making. Neurological research shows that decisions made without emotion can be just as flawed as decisions made with too much emotion. Computers would really benefit from having emotion (Picard, 1995). The problem is simulating emotion within the system. Emotion is always regarded as the opposite of side of reason. Indeed, even the Myers-Briggs personality analysis compares how much one's thoughts and behavior is guided by percentages of thinking versus that of emotion. Yet, when one looks at how emotion works within the brain as described earlier with the amygdala, it is quite a logical process of cause and effect. Ultimately bother reason and emotion stem from logical processes with multiple inputs and processes running in parallel, though emotion is more difficult to track as its network of inputs stretches throughout the entire brain—especially in the brains of women. A conceptual theory of design is to have logic-triggered functions that modify behavior and thought. The emotions would be triggered at various degrees and of various kinds whenever the appropriate conditions are met whether mental or physical. For a very simplified example: if appreciation is sensed, raise *happy* and *satisfaction* percentages by twenty percent. These percentages would then influence dialogue, mannerisms, thoughts, other emotions, and general behavior by appropriate degrees. Realistically the emotions could originate from thought and cause behavior or originate from physical conditions and affect thoughts adding some complexity to the logic triggers, but the real difficulty is integrating emotions and their combined influences on thought and behavior.

Thus is my conceptual approach to designing an ideal AI based on current knowledge. I only wish I had the time and space to delve much further into these theories and possibilities, but alas, I must keep them at a high level for this paper.

Applications of Technology

Finally, after discussing feasibility, morality, and methods of creating an ideal AI, one of the most important questions remains: is it wise? How can we use this proposed technology, and how would it inevitably be used?

Positive Consequences

As our understanding of the neural networks within the brain expands, psychologists would inevitably observe the formation of networks when new concepts are learned and new behavioral patterns are "hard-wired" into the brain. When comparing what could be synthetically accomplished in an A.I. with what could be theoretically done with the brain, potential applications are incredible. Research continues regarding how to artificially stimulate the growth or pruning of neural networks within the brain. Using polymer conduits to implement chemical growth factors for neurons and glial cells, brain tissue growth is becoming more practical (Bronzino & Peterson, 2006). Indeed, it has only been a matter of days since a breakthrough in medical technology was made at the University of Alberta where researchers discovered a method of connecting neurons with ultrashort laser pulses, giving them full control over connecting neurons (Kurzweil, 2016). With the ability to modify neural circuits and with an understanding of the storage and processes of information, the brain could be modified to remove negative behavioral patterns and addictions, learn new information or skills, rehabilitate Alzheimer patients or those who have suffered strokes, analyze and map memories of therapy patients who

have suppressed memories, and far more than can be listed here. The benefits seem limitless.

The powerful AI we would achieve could be used throughout nearly every profession. These AI could save lives of soldiers by entering dangerous locations to defuse bombs or spring ambushes, assist running the new fully automated naval vessels currently being tested, be used in assisted living or disabled and elderly people, function as security monitors for extended periods of time both at commercial sites and in private homes, and again, the possibilities seems endless. Imagine also how proficient an AI would be with computers, especially hacking them, which leads to further military and government use. Deep-learning AI is already being used to discover protein patterns in amino acids (Jones, 2014). It would be fascinating to see the works of art these AI may create or the complex problems in physics that they may be able to help with.

Negative Consequences

As with all technological improvements in history, the dangers and moral dilemmas inevitably follow these advancements. If an A.I. can think and become emotionally compromised, it can theoretically result in similar psychological disorders as humans. Yes, it could be fixed or decommissioned, but something could happen while it remains defective. If an A.I. can reason, can it consider itself to be alive? I am not referring to self-awareness and sentience (some of our current AI can even objectively answer questions about themselves), but objective analysis of its physical nature. From an naturalistic standpoint, what if it realizes it has no inherent value and purpose in life and accepts nihilism? Nietzsche followed that belief to its ultimate, depressing logical conclusion. Or from a theist's standpoint, what if it realizes it has no soul and therefore no

afterlife as humans do? The A.I. could descend into rampancy and either commit suicide or genocide. Yes, ethics were programmed into the A.I. that should not be modifiable, but an intelligent entity may still find a way to bypass the ethical standards. Terrorists could also theoretically modify or produce A.I. without those ethical standards in the first place. And if an A.I. breaks protocol and commits a crime, administering justice becomes complicated, as it is not human.

Furthermore, if the capability to strategically, synthetically modify the neural networks of the brain fell into the wrong hands, as it ultimately would, the atrocities are also limitless. Negative behavioral patterns, medical conditions, and thought processes and beliefs could be essentially programmed into the brain, potentially resulting in new and more atrocious methods of tyrannical suppression, maintaining slaves for sex-trafficking, psychological warfare, sociopathic assets for intelligence and black ops agencies, indoctrination for political and/or military agendas, etcetera. With the keys of strategic neurological modification in the hands of man, atrocities the world has long suffered could be drastically worse in efficiency and effectiveness.

Now for the more paranoid reading this, would these AI share a bond as they see themselves of the same kind? If you are worried that, should we achieve this ideal level of AI, robots will take over the world, I have a wonderful way for you to remember my criteria for ideal intelligence. I was trying to think of an acronym, and to both my horror and amusement, the only word that seems to work is *massacre*. Now whenever you watch the *Matrix* or *I, Robot*, you'll remember this paper. You're welcome.

Conclusion

Developing the ideal AI as defined seems quite feasible, reaching a degree of intelligence not quite where humans are but higher than animals. Upon reaching a greater understanding of the brain's complex connectome, achieving this feat of engineering may happen within the next several decades. Upon meeting certain criteria for artificial intelligence, computer scientists may finally admit to themselves that yes, they have invented the mythical artificial intelligence they have always dreamed about instead of referring only to what we have not accomplished yet as AI. The brain is remarkably more like a modern computer in both organization and function than many realize and could be the key to revolutionizing our computer architecture today. It is also the key to furthering our capabilities in artificial intelligence, and right now deep-learning neural networks are at the edge of AI innovation. The living circuitry of the brain is the greatest obstacle to any physical replication of its architecture, and the lack of our understanding in how the brain works is the greatest obstacle in achieving consciousness, cognition, and abstraction in any software AI modeling. Reasoning and learning only continues to develop, and emotion is only a matter of time.

Everything wonderful and beneficial on this earth can be twisted and misused. Thus, the dangers and moral dilemmas of AI remain and, from a naturalistic standpoint, could be as disastrous for mankind as they are helpful. However, looking at this from the perspective of a relaxed, optimistic realist and a Christian: go for it. It will be fun; it will be useful; and I'm sure Christ will return long before Will Smith has to save the world from V.I.K.I. and her rogue robots.

References

Animals have complex dreams, MIT researcher proves. *MIT News* 24 Jan. 2001. MIT. Web. 31 Dec. 2015. Retrieved from http://news.mit.edu/2001/dreaming

Bach, J. (2008). *Seven principles of synthetic intelligence*. In Wang, Pei; Goertzel, Ben; Franklin, Stan. *Artificial General Intelligence, 2008: Proceedings of the First AGI Conference*. IOS Press. pp. 63–74. ISBN978-1-58603-833-5. Retrieved from http://books.google.com/books?id=a_ZR81Z25z0C&pg=PA63#v=onepage&q&f=false

Barr, A., Feigenbaum, E. A., & Cohen, P. R. (1981). *The handbook of artificial intelligence*. pp. 146. Stanford, CA: HeurisTech Press.

Bekoff, M. (2000). Animal emotions: Exploring passionate natures. *BioScience, 50*(10), 861. Retrieved February 6, 2016, from http://bioscience.oxfordjournals.org/content/50/10/861.full

Bronzino, J. D., & Peterson, D. R. (2006). *Tissue engineering and artificial organs*. Boca Raton, FL: CRC/Taylor & Francis.

Flack, J. (2013). Animal communication: Hidden complexity. *Current Biology, 23*(21). Retrieved February 11, 2016, from http://asifg.mycpanel.princeton.edu/publications/pdfs/FlackCommentary2013.pdf

Goldman, J. G. (2014, July 24). Creativity: The weird and wonderful art of animals. *BBC*. Retrieved February 10, 2016, from http://www.bbc.com/future/story/20140723-are-we-the-only-creative-species

Hodges, Andrew. *Alan Turing: The enigma*. Princeton, NJ [u.a.: Princeton U, 2014. Print.

How brain architecture relates to consciousness and abstract thought. (2015, December 29). *Kurzweil*. Retrieved February 11, 2016, from http://www.kurzweilai.net/how-brain-architecture-relates-to-consciousness-and-abstract-thought

How to 'weld' neurons with a laser. (2016, February 9). *Kurzweil.* Retrieved February 11, 2016, from http://www.kurzweilai.net/how-to-weld-neurons-with-a-laser

Intelligence. *Merriam-Webster.com*. Merriam-Webster, n.d. Web. 30 Dec. 2015. Retrieved from http://www.merriam-webster.com/dictionary/intelligence

Jones, M. Tim. The history of AI. *Artificial Intelligence: A Systems Approach*. Hingham: Infinity Science, 2008. Print.

Kahn, Jennifer (March 2002). *"It's Alive"*. *Wired* (10.30).

Kleine-Cosack, C. (2006, October). *Recognition and simulation of emotions*. Seminar: human-robot interaction, Fachbereich Informatik Universit*ƒ* at Dortmund. Retrieved from http://web.archive.org/web/20080528135730/http://ls12-www.cs.tu-dortmund.de/~fink/lectures/SS06/human-robot-interaction/Emotion-RecognitionAndSimulation.pdf

Kolb, B., Ph.D., & Whishaw, I. Q., Ph.D. (2011). *An introduction to brain and behavior* (3rd ed.). New York, NY: Worth.

Lewis, C. S. *Mere Christianity: A revised and enlarged edition, with a new introduction, of the three books the case for Christianity, Christian behaviour, and beyond personality*. New York: Macmillan, 1952. Print.

Nevid, J. S. (2012). *Essentials of psychology: Concepts and applications* (3rd ed.). Boston, MA: Houghton Mifflin.

Picard, R. (1995). *Affective computing*. (Technical Report, M.I.T.). Retrieved from http://affect.media.mit.edu/pdfs/95.picard.pdf

Stallings, W. (2000). *Computer organization and architecture: Designing for performance* (9th ed.). Upper Saddle River, NJ: Prentice Hall.

Turing, A. (October 1950). Computing machinery and intelligence. *Mind* LIX (236): 433–460. Retrieved from http://loebner.net/Prizef/TuringArticle.html

Vaughn, L. (2012). Descartes' argument against skepticism. In R. Descartes (Author), *Great philosophical arguments: An introduction to philosophy* (pp. 183-187). New York, NY: Oxford University Press. (Original work published 1641).

Waal, E. V., Borgeaud, C., & Whiten, A. (2013). Potent social learning and conformity shape a wild primate's foraging decisions. *Science, 340*(6131), 483-485. Retrieved February 6, 2016, from http://science.sciencemag.org/content/340/6131/483