BETWEEN

MAC, MAN, & MANET:

an interdisciplinary exploration of artificial intelligence, art, and personhood

Jacquelyn Truong is a member of the class of Davenport 2010 at Yale University. This project, her senior thesis for the program in Humanities, is part scholarly essay and part op-ed. It represents the culmination of her undergraduate academic pursuits in philosophy, cognitive science, and art theory. Its intended audience is one of thoughtful non-experts who are, perhaps, largely unfamiliar with topics in philosophy of mind, artificial intelligence, and the philosophy of art.

This essay explores the concepts of mind and thought, and the role of creativity in personhood, through the lenses of artificial intelligence and art philosophy. Part I offers an introduction to concepts in artificial intelligence, under the precept that it is valuable to pursue explanations of thought in terms of computational models because, even though humans are not computers, if programs are capable of thought then that says something significant about the nature of human minds. The argument proposes a framework for cognition as a spectrum, with pure cognition on one end and human cognition on another; it also proposes and explains two special landmarks in the spectrum of intelligence, embodiment and self-awareness. I go on to argue that both landmarks are prerequisites for artistic creativity. Part II further explores the role of art and creativity in personhood and thought. It discusses the hypothesis, "Only humans can create art," and looks at several interesting examples in hopes that they will illuminate some truths about the nature of art and creativity as they relate to artificial intelligence.

TABLE OF CONTENTS

PAGE

3 **PART I: COULD A COMPUTER EVER THINK?**

4 **Intelligence... Apparently: Turing and Searle**

5 solipsism and the problem of other minds

9 that special something: the problem of consciousness

10 **Subjective Experience: Consciousness, Qualia, and Mental states**

13 **Pure Cognition/Human Cognition: A Spectrum of Intelligence**

14 feelings, emotions, and motivation in cognition

15 perception: information reception and retention

16 understanding: a coherent internal model

18 from pure to human cognition

22 **Between the Extremes: Embodiment, Sentience, & Animal Cognition**

22 embodiment: animal cognition

24 personhood and self-awareness

28 **PART II: THE "ART" IN ARTIFICIAL INTELLIGENCE**

31 **Art Must Articulate: the Role of Emotion in Artistic Creativity**

aaron & brutus: natural beauty vs. art

33 **The "I" in Creative: Personhood & Agency in Authorship**

35 **CONCLUSIONS**

Part I:

COULD A COMPUTER EVER THINK?

This section compares the ideas of some of the key figures in artificial intelligence theory; on one hand from the *computationalists* like Alan Turing, who believe that software is capable of thought, and on the other hand from *anti-computationalists* like John Searle and Paul Ziff. It continues to offer a hypothesis concerning the nature of cognition in support of the computationalist argument: that intelligence exists on a spectrum between pure cognition and human cognition, and that many of the capacities which some have deemed necessary for intelligence, but which may be out of reach for digital computers (such as feelings and emotions), may be required for human cognition, but not for pure cognition.

What is thought? What does it mean to have a mind? In exploring this question, many have turned to explorations from the philosopher's armchair; others, to the intricate disciplines of biology, psychology, and neuroscience; and many, since the invention of the digital computer in the 1940s, have tried to generate explanations in terms of software programs and artificial intelligence. The use of computational models in the exploration of human cognition takes for granted the idea that even though humans and computers are very different, there must be something fundamental about human intelligence that is not reliant upon the structure and physicality of the human brain – something that might be, if not precisely duplicable, effectively recreated through software.

INTELLIGENCE... APPARENTLY:

ALAN TURING & JOHN SEARLE

There are essentially two schools of thought in AI: computationalism and anti-computationalism. Computationalists believe it is possible to create a mind from software. One of the most important thinkers in the field, Alan Turing, wrote an article in 1950 entitled “Computing Machinery and Intelligence” establishing a thought experiment called the Imitation Game, which would later give rise to the famous Turing Test, a measure which many still consider to be the litmus test for artificial intelligence.

The Turing Test pits a human judge up against two interlocutors: one is human, and the other is a computer program. By way of written word alone, both must convince the judge that the other is the computer, and that they are the human; any question or statement is fair game.

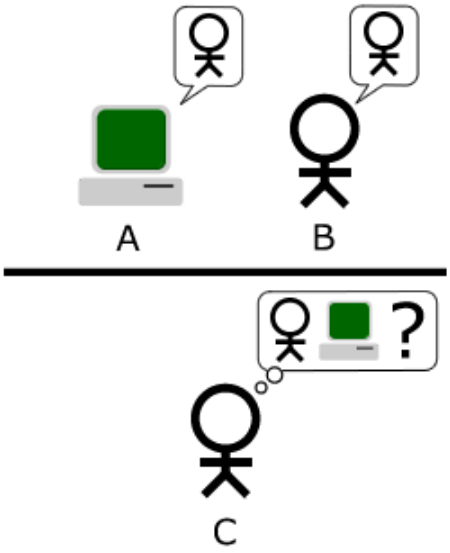


Illustration 1: The Turing test. Image from the public domain.

Turing suggests that because we can never truly know anything about minds other than our own, we must be satisfied with results in terms of performance and percentages, measuring artificial intelligence as a computer's capacity to fool the Turing Test judge a certain percentage of the time. (Turing 1950)

.....solipsism and the problem of other minds

The problem of other minds is an old epistemological challenge. It is a skeptical claim that because we can only be sure of our own consciousness, other people who appear to be conscious could be mere “automata” or “philosophical zombies”, apparently conscious and intelligent but in fact devoid of thought, mere puppets in a thoughtless, though convincing, simulation. In his essay, Turing addresses an “argument from consciousness” that is antithetical to his hypothesis. According to this argument, an intelligent being does not merely simulate having experiences, it actually has experiences. The argument accuses the Turing Test of failing to distinguish between truly intelligent machines and convincing pretenders. Turing responds that since it is impossible to prove the presence of consciousness in others (whether machine or human), then to doubt that a seemingly intelligent machine is conscious is also to doubt that seemingly intelligent people are conscious; in other words, he argues that if apparent intelligence is not enough to convince someone that a machine has “true intelligence,” then that person is not justified in taking for granted the intelligence of other human beings, who also may only ever *appear* intelligent by virtue of the problem of other minds. For these reasons, Turing accuses his opponent of solipsism, the much-frowned upon belief that one is the only intelligent entity in the universe, and that all other apparent intelligence is fraudulent, as in the philosophical zombies. By Turing's logic, anyone doubting

that *anything* possesses consciousness could be called a solipsist:

> INTERLOCUTOR: I'm just saying that *apparent* intelligence, in the form of a computer that really seems smart, is not enough to prove the existence of real, conscious intelligence! There could be some really clever deception going on!
>
> TURING*: Well, good sir, are you an expert on which kinds of things have minds, and which don't?
>
> INTERLOCUTOR: Well, I wouldn't say that, but there are some things I can say with certainty; for instance, I really doubt that this chair I'm sitting in has a mind of its own.
>
> TURING: Well, you could never really know that for sure, since you can only ever perceive your own mind, and no one (or nothing) else's! If you doubt that the chair has a mind, and you doubt that an intelligent-seeming computer has a mind, what's to stop you from doubting that I have a mind, even though I also *seem* intelligent? You must be a solipsist!
>
> *this dialogue is of my invention and should not be understood as quoting Turing.

Turing misses the point. Whether or not skepticism about the presence of intelligence in computers, staplers, or any other non-human things is philosophically hypocritical, ultimately a thinking computer can either exist, or it cannot. I believe that toaster ovens don’t have minds, and - whether I am a solipsist or not - I am either right or wrong. The alleged “sin” of solipsism cannot affect the truth of his opponent's claim, and by not admitting this Turing skirts the issue.

This failure to distinguish between real intelligence and the simulation of intelligence constitutes the greatest weakness in Turing's argument, and it is this very issue that John Searle, one of the key anti-computationalist thinkers, tackles in his essay, “Minds, Brains, and Programs.” In the essay, Searle creates what has become one of the most well-known thought experiments in modern philosophy - the Chinese Room. (Searle 1982)

Imagine that John Searle is isolated in a room where he receives slips of paper with what appear to be “meaningless squiggles” written on them. His job is, with the help of a catalog of rules (written in English without any indication of what any squiggle “means”), to match the input squiggles to corresponding output squiggles, and thus to write “responses” to the person outside the room. The catch is that the “meaningless squiggles” are actually Chinese characters written by a native Chinese speaker, and the responses that Searle is responsible for outputting are formulated in correct Chinese.

Illustration 2: The Chinese Room.

Searle uses this thought experiment to illustrate what he believes occurs in a computer program, in order to convince the reader that it is impossible for a program ever to truly understand when all it is capable of is blind “formal symbol manipulation” - in other words, dumb, mechanical processing. In distinguishing

between formal symbol manipulation and "really understanding information," Searle argues that the former can create the illusion of understanding while missing a crucial point: the symbol manipulator never has to deal with any symbol as a representation – it needs only to have a complete set of instructions on what to do with each formal symbol itself. The fact that it is symbolic at all is irrelevant to the manipulator. Thus, Searle appears to suggest that understanding implies a higher level of association between symbols and their referents. Ultimately, I think that the Chinese Room thought experiment successfully refutes the Turing Test by proving that it is possible for a system to appear as intelligent as a human without actually being as intelligent as a human. However, I think Searle fails to prove that it is impossible for software to think. Proving that X without Y is possible is not the same as proving that X with Y is impossible; in other words, proving that a computer with apparently human-like intelligence can be unintelligent does not prove that a computer with apparent *and* real intelligence is impossible.

In his essay, Searle also discusses what he calls the "causal powers of the brain". This term, along with "intentionality", are what Searle uses to describe a somewhat ill-defined quality that humans, and other biological creatures, allegedly have that programs alone lack. Searle believes that somewhere in the complicated web of neurons and molecules that make up a living brain, there is a recipe (though not the only recipe) for true understanding that is simply absent

when a machine merely crunches numbers. He claims to have no qualms about the idea of a thinking machine, but believes that the process of symbol manipulation alone can never generate intentionality, and that his own “causal powers” give him the uniquely intentional capacities for “perception, action, understanding, and learning”. Later in my argument, I hope to isolate the concepts of understanding and intentionality, because I think they belong to fundamentally different categories as requirements for intelligence.

.....that special something: the problem of consciousness

Searle's critics might argue that his idea of intentionality is too vague; he does not offer a satisfactory explanation of what exactly “causal powers” are and how they emerge. These qualities are members of a subset of characteristics which are of special interest to me; they might be described as those ineffable qualities of human intelligence that software cannot seem to duplicate, including intentionality, point of view, qualia, and mental states. These can all be roughly boiled down to the “hard problem of consciousness,” another well-known philosophical dilemma. Its basic formulation is, “How does the purely physical system of my brain give rise to an entity with subjective experience, when other physical systems like mountain ranges or tree branches apparently do not?” My position is that the anti-computationalists who would insist that qualia are essential to intelligence either have an overly anthropocentric concept of

cognition, or they are imposing requirements on intelligence that are too strict. In response to these arguments, I will discuss what I call *pure cognition*, and will defend the claim that there is a kind of minimum requirement for intelligence that is possible in an unemotional, unfeeling digital computer.

SUBJECTIVE EXPERIENCE:
CONSCIOUSNESS, QUALIA, AND MENTAL STATES

Though they are not precisely the same thing, most would agree that cognition and consciousness go hand in hand, and that if something is not conscious, it is not capable of thought. Many opponents of computationalism argue that digital computers are inherently incapable of consciousness, and one of the concepts that they use to illustrate their point is “qualia”. Qualia are, in simple terms, the particular ways that things seem to us when we experience them; the way it feels to smell cookies, the way it feels when we touch velvet, the way it feels when we hear the sound of boiling pasta. Many believe that, in order to be truly conscious, the mind must have subjective experience with qualia, and they argue that there is something about experience that is more than quantifying or listing physical information can amount to. A well-known thought experiment that is used to illustrate this point is called Mary the Super Scientist (Jackson 1982).

Imagine that there is a scientist named Mary who specializes in the neurophysiology of vision, but who has spent her entire life in a room where it is only possible to see in black and white. Even though Mary knows all of the facts about what it means to see the color red – for instance, how the wavelength of light at 650 nanometers stimulates the retina, and how this information travels from the eye to the brain, etc. - since she's never actually seen it, when she finally breaks free from the black and white room and sees a red rose, she will have learned something that she didn't know before; namely, the "qualia" of what it's like to see the color red.

Illustration 3: Mary the Super Scientist.

I agree that it "is like something" to see red. However, I disagree with the idea that qualia are some kind of transcendental, inexplicable phenomenon; Daniel Dennett argues that experience can be completely described as collections of data that the mind processes, and that if Mary really knew everything there was to know about what it would mean for her to see red, this would imply a very deep knowledge about the experience of color, and with it she would be able to properly imagine seeing red even if she had technically never seen it before. (Dennett 1991, 398)

This idea that there is something about human experience that cannot be reduced to "mere" mechanical processes and collections of data is called "the

explanatory gap," because its proponents believe that something special is needed to fill the gap between our mental experience and the physical mechanisms from which they apparently spring. While I understand the feeling that there must be something more to conscious experience than just information processing, I believe that this insistence on some kind of transcendent or metaphysical solution to "the mystery of experience" is grounded in a delusion similar to the one at play for supporters of the teleological argument for the existence of God:

> INTELLIGENT DESIGN SUPPORTER: The universe is beautiful, complex, and amazing; for such beauty to come about by chance is so unlikely even to the point of being statistically impossible; thus, it can only be explained by the presence of a cosmic designer: God.
>
> EXPLANATORY GAP PROPONENT: Human subjective experience is unique, complex, and amazing; there must be something beyond physical mechanisms and information that explains it.

If the basis for the explanatory gap is really no stronger than the argument for intelligent design, it is in serious trouble. But all things considered, I do agree that qualia exist and that they are a requirement for what I will describe as "human cognition". In my understanding, a quale can be described as a combination of two things: first, it requires sensory information, like visual data from an eye or a digital camcorder, haptic data from touch sensors or fingers, or audio data from ears or a microphone; the other part of a quale is the subjective

experience part, which I interpret as a physical sensation or feeling, perhaps occurring as a function of the corresponding brain activity that a living body inevitably experiences which accompanies the reception of sensory information. In other words, there is all of the physical information – the sunlight hitting the rose with a particular brightness, the reflected light waves hitting Mary's eye at 650 nanometers, the rods and cones in her eye responding in a particular way, etc. - and then there's the collection of *feelings* that Mary gets as she is seeing the rose, which is ultimately just more physical information (though perhaps of a slightly different or more complex kind). So while both a computer and Mary can receive the same physical data about the red rose, Mary has an experience with qualia because she has a body with feelings, and the computer does not.

Consciousness, experience from a point of view, the subject-ness of perception – these things boil down to qualia, and qualia boils down to input data and the feelings that come with them. While necessary for (or inseparable from) human cognition, I believe that qualia are not necessary for pure cognition, and that they are fundamentally a by-product of perception occurring in a living organism.

PURE COGNITION / HUMAN COGNITION:
A SPECTRUM OF INTELLIGENCE

> I believe it is possible to build a thinking digital computer. To begin, I want to draw a distinction between pure cognition and human cognition.

> Then I wish to discuss how the distinction between human cognition and pure cognition is relevant to the ethical questions emergent in discussions of artificial intelligence, namely: would a thinking computer have any rights, and if any, which? Would a mind made of software be considered a person? Furthermore, the answers to these questions set the stage for a discussion about the relationship between human cognition, embodiment, and creativity.

We all know what software is. But what is a mind? Specifically, what is necessary for a mind, and what is not? I believe that it is useful to identify what I call *pure cognition* not only as an abstraction from human intelligence as we know it, but also as a distinct and real kind of intelligence. I will argue that there are only two things are necessary for pure cognition – perception and understanding. Moreover, I will argue that pure cognition can be attained through software, and that many of the other abilities that have been discussed as necessary components of a thinking mind (emotion, empathy, creativity, spontaneity, personality, ability to fool humans, et cetera) are unnecessary for pure cognition, though perhaps necessary for human cognition.

.....feelings, emotions, and motivation in cognition

Computationalists and anti-computationalists alike have defended the claim that emotions or feelings are necessary for thought. In a dialogue called "A Coffeehouse Conversation," Douglas Hofstadter argues that emotion in the form

of motivation or desire is at the heart of all action; he argues that conversations between thinking beings must necessarily be driven and framed by an emotional undercurrent (Hofstadter 2001); I think this is an anthropocentric idea, and ultimately reject the claim that emotions are necessary for bare-bones cognition. I'm not the only one with this idea, either; the idea of an unemotional but thinking being is not only imaginable, but is practically an archetype in creative portrayals of artificial or non-human intelligence. Hofstadter might respond that just because something is philosophically possible or imaginable doesn't mean that it is really possible. I agree, but propose an explanation of how it might be possible in the following sections.

.....perception : information reception & retention

Most people agree that perceptions or experiences are necessary for thought. After all, a thought occurs in time, so if a mind has no experience of thought, how can any be said to occur? If you think of perceptions as sensory input, it seems obvious that a program could perceive, technology allowing, any information from the external world, whether it be user input or audiovisual data collected from a camera. Software, I think, is inherently capable of perception because of the simple fact that it can receive and retain data; so as long as data are given to it and it has access to and can “remember” these data, it can be said to be capable of perceiving – regardless of whether or not it felt any kind

of qualia in conjunction with its perception. You could call this idea a “perception as accessible input hypothesis”. That said, I don't believe that perception is the only necessary component for pure cognition; not only are data reception (input) and retention (memory) critical to perception, but what must come hand in hand with them is the capacity to understand what is being perceived. Without understanding, perceptions might not be not perceptions at all, but just collections of meaningless data. But what does it mean to understand?

.....understanding: a coherent internal model

Is understanding possible in the context of software? Namely, as John Searle put it, can “meaningless symbol manipulation” alone result in, not just accurate responses and behavior, but true understanding (Searle 1982)? This problem, at first, seems to be an insurmountable obstacle; there does seem to be some elusive “spark” or “je ne sais quoi” about human understanding that any combination of ones and zeros (or, in Searle's terms, any number of rule books) could never attain. But this picture of understanding as some kind of transcendent or unexplainable power is unjustifiably narrow and, frankly, mystical. Douglas Hofstadter offers an explanation of consciousness as a “pattern of organization” in which information about the external world (and information about the self) is given representation internally, and responses to the external world are formulated on the basis of this internal model (Hofstadter

2001). As long as the mind has full and recursive meta-access to its internal representational model, I think it can be said to understand. For instance, if a program “knows” that its name is SAM, not only should it have the datum that “My name is SAM;” it should also be able to figure out that *it knows its name is SAM,* and it should know that it knows that it knows its name – and so on. Thus, in much the same way that it should have access to information about the external world, information about its own internal world is available to it.

The problem with Searle's Chinese Room, then, is that the man in the room, portrayed as a part of the room and not representative of the room as a whole, does not have full access to the internal representational model upon which his output is based – if he did have full access to that model, he would effectively know Chinese, because he would have all of the necessary information. In the same way, any of our individual neurological components does not have understanding, but our cognitive system as a whole does.

That said, I don't believe that an entity capable of perception and understanding – in other words, something with pure intelligence - is necessarily “conscious”. I think the capacity for consciousness, like that for qualia, emotions, and motivation, lies outside the scope of entities with only pure cognition, and that these capacities are, rather, characteristic of minds with human cognition.

.....from pure to human cognition

> In the artificial intelligence debate, people seem to be debating two different things: on one hand whether pure cognition is possible via software alone, and on another whether human-like cognition is possible via software alone. I think that it will be possible, one day, for a machine to be self-aware or sentient, but I have serious doubts that fully human-like intelligence will be possible through software alone (unless we invent some kind of revolutionary hardware or technology), and I believe that programs with pure cognition will appear decades, if not centuries, sooner that ones with anything resembling human cognition.

There is a host of qualities that have been cited as necessary for intelligence: emotions, motivations, empathy, creativity, spontaneity, feelings. I believe that none of these are necessary for pure cognition. I will agree, however, that there are certain capacities that are inseparable from *human cognition,* and I think this is because human cognition occupies a special range on the cognition spectrum where a capacity for perception and understanding overlap with *embodiment* and *self-awareness*.

Programming human-like cognition is obviously a tricky goal, since so much of human psychology is dependent on biology. This makes recreating human intelligence in software quite complicated, if not impossible – though that has not stopped some from trying[1]. The subtleties of human communication, along

1 "Henry Markram Builds a Brain in a Supercomputer," TEDGlobal 2009. http://www.ted.com/talks/lang/eng/henry_markram_supercomputing_the_brain_s_secrets.html

with the importance of emotion (and, by proxy, motivation) in interpersonal interaction (Hofstadter 2001), may make it difficult for a program capable of only pure cognition to pass the Turing test, while a program with human-like intelligence would have an excellent chance of passing. However, I think that if a program were written that was able to pass the Turing test, it would be highly likely (if not necessary) that that program was capable of pure cognition. You could say that, while it fails to conclusively prove anything about consciousness, personhood, or self-awareness in a machine, what the Turing test tests for is pure intelligence, and any machine that passes it must be, at the very least, capable of pure cognition.[2]

If a program were written that was capable of pure cognition, I think it would be a genuinely thinking thing. However - and it is here where I think much controversy and queasiness over the matter resides – I think a program capable of pure cognition would not qualify as a person, and would not until it had all the capacities of a mind with human cognition. This leads me another idea: that the difference between pure and human cognition is also deeply related to the difference between personhood and non-personhood, and that if we should ever create a robot with human-like cognition, it should be considered a person.

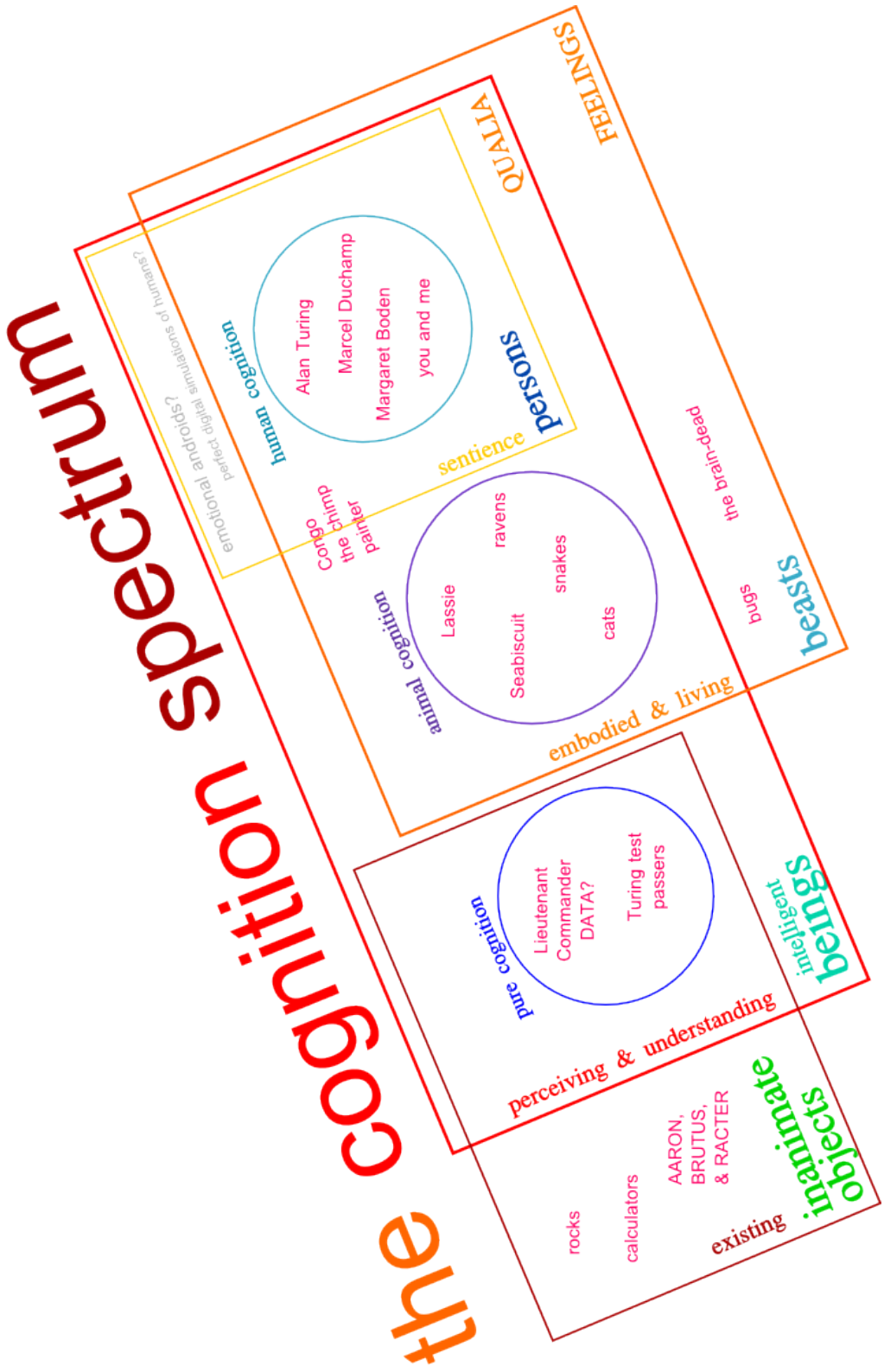
But what do I mean by "person," and what is the difference between a person

2 In other words, pure cognition is necessary, but perhaps not sufficient, for any machine to be capable of passing the Turing test.

and a human (though most humans are persons)? A human being in our times has human rights, citizenship of a country, legal rights and responsibilities, family history, and a physical body of the species Homo sapiens. A person, on the other hand, may be defined as having a mind with high powers of cognition; it is aware of itself, and arguably can be creative; it has intelligence, some form of emotions, intentions, desires, a character, and memories, among other things. A person, even if non-human, has certain essential rights because of the capacities which qualify it for personhood.

The complexities of the rights and responsibilities of persons without living bodies, as would be the case for an intelligent computer program with feelings and self-awareness, is the stuff of science fiction. In an essay on machine ethics, Drew McDermott argues that in order to experience an ethical conflict, one must struggle with temptation (McDermott 2009). After all, while resisting a temptation to steal is morally good, if one is not tempted to steal in the first place, then not stealing should not be considered an ethically laudable decision. He goes on to argue that temptation and the problem of ethical dilemmas are ultimately a product of the "idiosyncratic architecture of the human brain" - a claim not too different from ideas that I will discuss concerning the role of embodiment in the cognition spectrum.

Illustration 4: The Cognition Spectrum, J. Truong.



BETWEEN THE EXTREMES:
EMBODIMENT, SELF-AWARENESS, & ANIMAL COGNITION

In setting up a paradigm of intelligence as a spectrum between pure cognition and human cognition, the question inevitably arises: what happens in the middle of the spectrum? I want to argue that, along the path between pure cognition in digital machines and human cognition in living, breathing bodies, there are two junctures or sluices, so to speak, past which the nature of the intelligence at hand fundamentally changes. The first of these is embodiment, which for my purposes will have much to do with "life" and the non-interchangeability and interconnectedness of a mind and its body; the second of these is what I will call sentience or self-awareness. I believe embodiment is the prerequisite for such capacities as feelings, emotions, and motivation – all key elements of human cognition, but not pure cognition. Self-awareness, or what I will argue is the prerequisite for personhood, is "farther along," so to speak, on the cognition spectrum, and ultimately I want to argue that it is the combination of embodied *and* self-aware intelligence that is the correct recipe for artistic creativity.

.....embodiment: animal cognition

Embodiment, or more simply having a living body, is clearly not in itself equivalent to being capable of cognition. To use a rather tragic example, those

human beings who have been diagnosed as brain dead may have a beating heart and, aside from their brain, a normally functioning and living body – and yet be completely and permanently incapable of cognition. However, when you consider embodied intelligence along the cognition spectrum, it brings such capacities as feelings, emotions, and motivation into the picture – a starkly different story from the cold, metal-and-wire intelligence implied by pure cognition in a digital computer.

In his argument against computationalism, "The Feelings of Robots," Paul Ziff claims that a machine, by virtue of its very machine-ness, could never have feelings, and could only ever make a performance of feeling them. "Robots may calculate," he writes, "but they will not literally reason. Perhaps they will take things but they will not literally borrow them." (Ziff 1959, 65) He argues that to lie, cheat, and steal are actions that only living persons can perform, and that because robots cannot do these things – they would only print untrue statements, play games incorrectly to the detriment of other players, and remove objects from the possession of their owners - they are not alive. A die-hard computationalist might be tempted to argue that Ziff makes a truism or platitude in saying, "Well, it's different when people do it!" by giving equivalent actions different names when done by humans. But essentially, Ziff is not arguing that computers are incapable of cognition – only that they cannot have feelings – and fundamentally I agree with him. Because robots are not alive,

they cannot feel – the capacity to feel is an undeniably important element of human intelligence.

Humans are clearly not the only creatures with feelings – but while most scientists have found it safe to assume that the philosophical problem of other minds need not be applied to other humans, the unprovability of animal cognition has made it a difficult-to-navigate and controversial field. Though behavioral scientists might cringe at the thought of trying to document the content and quality of animal consciousness, there seems to be a widespread tacit agreement that some animals are conscious some of the time (Griffin 2000). Embodied cognition represents significant shifts in the nature of intelligence from pure, disembodied cognition, and these shifts have to do with the inextricability of a mind from its body, and the emergence of feelings or emotions based on the relationship between its living body and its environment. A being with embodied cognition, yet whose intelligence falls short of human intelligence, can be said to be capable of animal cognition.

.....personhood and self-awareness

The next junction in the spectrum between pure and human cognition is self-awareness or sentience, and it is here that persons are distinguished from non-persons. A creature with animal cognition – one with perception, understanding,

and feelings – arguably has only basic cognitive functions. But as you move further up the spectrum and as the intellectual complexity of the organisms at hand grows, they move closer and closer to achieving self-awareness, and at a certain point there is a transition from animal to person – though where exactly this point is remains unclear (and perhaps must remain unclear). There is plenty of evidence of creatures exhibiting behavior to suggest that they straddle the line between animal and person; the fascinating research done with great apes such as Koko the gorilla and Lucy the chimpanzee should give pause to the many who believe personhood is possible in humans alone. To illustrate: Dr. Roger Fouts, a primate researcher, was the sign language teacher for a chimpanzee named Lucy Temerlin, who lived from 1964 to 1987; he recounts the following signed conversation with her during which she lies about a pile of feces she left on the floor (Fouts and Mills 1998) :

FOUTS: WHAT THAT?
LUCY: WHAT THAT?
FOUTS: YOU KNOW. WHAT THAT?
LUCY: DIRTY DIRTY.
FOUTS: WHOSE DIRTY DIRTY?
LUCY: SUE. [i.e., Sue Savage-Rumbaugh, a then-graduate student]
FOUTS: IT NOT SUE. WHOSE THAT?
LUCY: ROGER!
FOUTS: NO! NOT MINE. WHOSE?
LUCY: LUCY DIRTY DIRTY. SORRY LUCY.

Most pet owners will have occasion to project human emotions, like jealousy or resentment, onto their dogs and cats on occasion, but Lucy's lie is in a different

realm entirely; it provides a mountain of implications about the reaches of her mental world, including evidence of guilt, shame, and responsibility – all which suggest a degree of self-awareness. So it seems to be the case that non-humans can, in theory, be persons; but surely, some will argue, something must still be sacred – there must still be something that only humans can do.

The creation of art is considered by many to be among the highest activities a human can partake in. So the question must be posed: if non-humans can be persons, could a non-human person be artistically creative? Is it possible for a non-human to be an artist? As an introduction to Part II of my essay, I will describe two newsworthy anecdotes about non-human "artists": chimpanzee painters Congo and Pierre Brassau.

Illustration 5: Painting by Congo.

A 2005 news piece entitled "Dead Chimp's Art Sells Big" (de Vries 2005) describes an auction at which paintings by Congo, a chimpanzee who lived from 1954-1964 and who allegedly made more than 400 paintings in his lifetime, were sold for $26,352 when early estimates guessed they would be sold for less than $2,000. Howard Rutkowski, the auction house's modern art director, is reported to have said **"We had no idea what these things were worth."** The reports seem to support the notion that Congo was something of a legitimate artist; for instance, it is reported that when painting Congo apparently "knew" when a piece of his was finished or not; if a painting was taken away before he felt he was done with it, he would scream and throw a fit; likewise, if he felt he was finished with a piece, it was impossible to get him to put any more work into it. (Januszczak 2005)

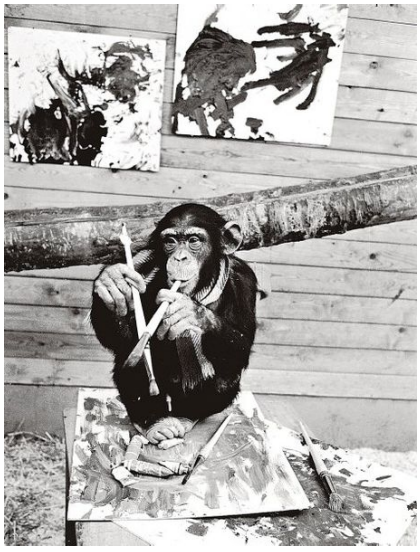


Illustration 6: Pierre Brassau at work.
Photo by Ake Axelsson.

On the other hand, “Pierre Brassau” was the pseudonym of a chimpanzee named Peter, who was the star of a hoax perpetrated in 1964 by Swedish Journalist Åke "Dacke" Axelsson (Saunders 1980) . In an attempt to mock pretentious contemporary art critics, Axelsson gave Peter art supplies (many of which were apparently eaten, but some of which were used to make paintings) and took the best of the ensuing paintings to a gallery. He exhibited them under the false name with the pretense that the paintings were made by a previously unknown avant-garde artist. Brassau's work received mixed reviews:

"Brassau paints with powerful strokes, but also with clear determination. His brush strokes twist with furious fastidiousness. Pierre is an artist who performs with the delicacy of a ballet dancer."	**"Only an ape could have done this".**

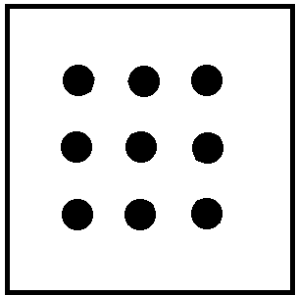
What is art? Who can be an artist? Are the paintings of these chimpanzees works of art, or merely scientifically interesting objects? What relevance do these questions have with the study of artificial intelligence, and the spectrum of cognition?

Part II:

THE "ART" IN ARTIFICIAL INTELLIGENCE

This section elaborates on the discussion of creativity as it relates to artificial intelligence; first is a discussion of the definition of creativity, followed by the question of what is necessary in the creation of art. Ultimately my argument is that the creation of art is only possible in intelligent minds that are both alive and self-aware, which leaves the open possibility of non-human artists.

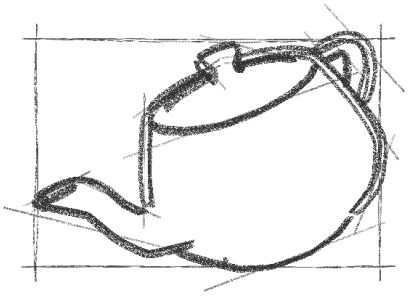
What is creativity? You know it when you see it, in Mozart's piano concertos, in Monet's dancing water lilies, and in Miller's dizzying, provocative prose; creativity also crops up in problem solving – arguably, you need it to answer such "lateral thinking" questions as the nine dots puzzle:



Without lifting your pencil, draw four straight lines such that they pass through all of the dots.

Creative actions, thoughts, or work can be divided into many different categories. The kind you need to solve the nine dots puzzle has a quality of unexpectedness or divergence – you could call this "divergent thinking" or "thinking outside the box" creativity (Partridge and Rowe 1994). Another kind of creativity seems to involve looking

at old things in new ways to form emergent, or completely new, ideas– a classic example is the invention of non-Euclidean geometry. Furthermore, there is the kind of creativity that is behind the work of Mozart, Monet, and Miller – you could call this a kind of generative or expressive creativity. Beyond these distinctions, there seems to be an underlying duality at work in the definition of creativity; on the one hand, creativity implies novelty; ideas or artworks are said to be "creative" when they are totally new, unexpected, or display a certain degree of genius. For instance, if a product designer submits a sketch for a teapot that looks like the one shown here, few people would defend the design as "creative". On the other hand, I think it would be wrong to limit the ascription of creativity only to things that were ingenious or innovative. In fact, it seems clear that bad or mediocre creativity is not only possible, but actually quite abundant[3]. This second kind of creativity seems to be more about the generation of work or ideas for the purpose of (or as a result of) self-expression.



Margaret Boden, one of the leading thinkers in the study of creativity and cognition, divides creativity into (P)-Creativity and (H)-Creativity. H-Creativity is short for historical creativity, and refers to those ideas that are completely novel

3 http://www.museumofbadart.org/

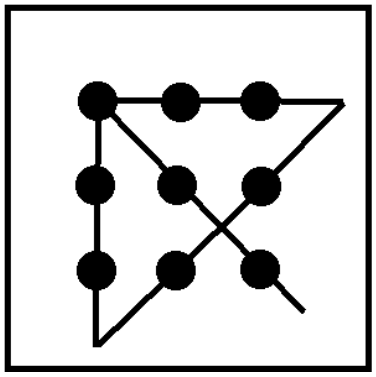


Illustration 7: Solution to the 9 dots problem.

compared to all past ideas by all people. P-Creativity, or psychological creativity, is described as creative thought relative only to oneself and one's ideas; someone experiencing an "Aha!" moment realizing that you can use discarded seat belts to make handbags is experiencing P-Creativity but not H-Creativity, since someone else has already had that idea. Conversely, if an idea is H-Creative it is necessarily P-Creative (Dartnall 1994). Boden writes that all creative ideas are ones that "not only did not happen before, but which ... could not have happened before". In other words, Boden's definition of creativity requires not only novelty, but a kind of emergent novelty that amounts to more than a rearranging or recombination of old ideas. Accordingly, she would argue that not all generative acts are creative; she uses the following made-up sentence as an example:

"The mangoes are in the oak-chest,
next to the socks that belonged to Dante."

In all likelihood, neither she nor anyone else had ever thought of that sentence before she wrote it to make her point. It is new, she would argue, but not *creative.* I disagree with Boden – or, rather, I prefer to discuss a different kind of creativity than she does. For the purposes of my discussion, novelty and unexpectedness will be largely irrelevant because they are relative measures that rely on events and truths that reside in the past, or independent of the

creating agent. Instead I will focus on creativity that needs no such reference - the kind at work in the production of all artwork – good and bad, genius and mundane.

ART MUST ARTICULATE:

THE ROLE OF EMOTION IN ARTISTIC CREATIVITY

Boden's sentence about Dante's socks is creative because it is expressive of the state of mind she had when she wrote it, complete with emotions and feelings; even if the sentence has nothing to do with what she was feeling at the time, her intelligence, inevitably affected by the feelings of the body in which it resides, intentionally gave rise to a set of words which are, as a result, expressive by default. Because I can infer that Boden is human and an intelligent person, I can even make guesses about her intellectual landscape based on the arbitrary and seemingly nonsensical sentence that she constructed; I might guess that she likes mangoes, that she has an object in her house that she refers to as an oak-chest, or that she has read Dante's Inferno. Perhaps none of these guesses are correct, but that is irrelevant. The point is that we can take it for granted that because she is a living person, her creations are backed by emotions, feelings, memories, and intentions.

.....aaron & brutus: natural beauty vs. art

So what about the creations of computer programs like AARON, a painting

program developed by Harold Cohen (Cohen 1999)? Or BRUTUS, a story-writing program specializing in tales of betrayal (Bringsjord and Ferrucci 1999)? I think I can safely say that no computer program written today is capable of having feelings, or using human-like cognition. Because there can be no emotion underlying the work of these programs, their creations can't be art. Art is essentially an expressive, communicative act, and if there is nothing there to express, then there can be no art. For instance: if the wind knocks a house painter's paint cans from his ladder onto a tarp in his absence, defending the end result as a legitimate art piece is tricky. Who is the artist? The absent house painter? You could make a case for that, but it seems wrong. The wind? Surely not. Even if the resulting paint splatter looks, inch by inch, identical to a piece by Jackson Pollock, one of the paintings is art, and the other is just an accident. Beauty or aesthetic merit is not enough to qualify something as an art piece; even though a field of sunflowers may be beautiful, it is not art. Likewise, for programs like AARON and BRUTUS, there is hesitation to call their productions “art” because there is something “windy” about the way their pieces come about. Something crucial, which might be described as intentionality or “on-purposeness,” is missing from the work of these programs. If you think about intentionality as essentially an emotional state consisting in a desire to

Illustration 8: A painting by AARON: Two Friends with Potted Plant, 1991. Oil on Canvas.

do something, then since AARON and BRUTUS have no emotional capacity, their actions can never have intention, and their creations can never be expressive. Harking back to the argument made by Ziff, only living things can have feelings – therefore I must conclude that for an entity to make art, it must not only be intelligent – it must be alive.

THE "I" IN CREATIVE

PERSONHOOD & AGENCY IN AUTHORSHIP

> In this section I attempt to argue that an artist must be self-aware, and that while the role of agency (the capacity for free and deliberate action) in art has been challenged by artists like Marcel Duchamp, there is a minimum amount of agency that is required for an object to be considered an art piece.

Though not all things made by persons are art, only to the extent that something is made by a person can it be considered art. The windblown paint splatter is not art, but not only because there is no emotional or expressive element behind it. An artist must not only have feelings to communicate; she must also have some degree of self-awareness, and she must manifest her expression with some amount of directed agency. A dog might feel exuberant and proceed to tear apart a stuffed bear; the feelings behind the making of such an object are evident, but the resulting mess is not an art piece. A frustrated fast-food chef might put together a messy burger – but that is not art either.

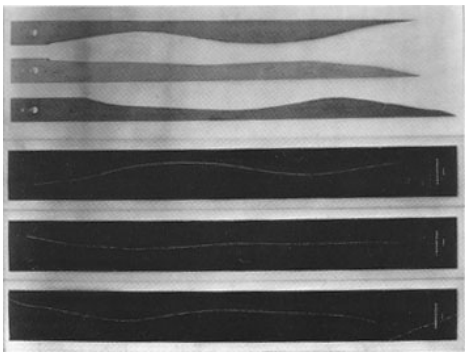


Illustration 9: Duchamp's 3 Standard Stoppages, 1913.

The role of agency in authorship has been challenged by the likes of the Dadaists, among others, with the incorporation of chance into the creation of art. Take, for instance, Marcel Duchamp's 3 *Standard Stoppages*. To make the piece, he cut three lengths of string, one meter each, and let them fall haphazardly to a flat surface. The curves created by the fallen strings were made, without modification, into wooden measuring sticks. And yet, despite the large role that Duchamp has allowed chance to play in the creation of this piece, there is still a sense that the piece belongs mostly to Duchamp - not nature or chance, as is the case with the field of sunflowers. Of course, Duchamp even further questioned the role of agency in art with his infamous *Fountain*, a so-called "readymade" that was an upturned urinal signed with the alias R. Mutt. Though many are tempted to deny that *Fountain* is art, the essentials are there: it is certainly expressive, albeit conceptually rather than directly[4], of some kind of emotion on Duchamp's part; and even though he did not himself make the urinal, without his agency there certainly would have been no *Fountain* – just another urinal. Whether *Fountain* is *good* art is beside the point – for better or for worse, it is art, and the fact is, I'll argue, that a being without self-awareness could never have come up with the idea for *Fountain,* or any other legitimate

4 A painting, for instance, would be directly expressive in this sense. Even if Duchamp considered *Fountain* to be a joke, this in itself is an expressive act.

work of art, because expressive action or production without personhood or sentience is merely the manifestation of instinct. And insofar as the obeisance of instinct is not the realization of free agency, it is effectively equivalent to the following of an algorithm; and insofar as something follows an algorithm, it is not art, as we proved with the cases of AARON and BRUTUS.

CONCLUSIONS

What this all boils down to is that because only the coincidence of self-awareness and embodiment in an intelligent mind can create a habitable zone, so to speak, for creativity, digital computers as they are today can never really be creative, even though they can, in theory, have intelligence in the form of pure cognition. Artists and art lovers may warmly note that on the spectrum of cognition, personhood and creativity perfectly coincide — and it's no coincidence. It may be a bit of a surprising conclusion to say that the realm of artistic creativity is not limited to humans, but if it turns out that either Congo or Peter was as self-aware as Lucy seems to have been, I stand by the claim that their paintings would have to be considered works of art. That's not to say that I would buy one, of course, but that is an entirely different matter.

sources cited

Bringsjord, Selmer, and David Ferrucci. 1999. *Artificial Intelligence and Literary Creativity: Inside the Mind of Brutus, A Storytelling Machine*. Psychology Press, September 1.

Cohen, Harold. 1999. Colouring Without Seeing: a Problem in Machine Creativity. crca.ucsd.edu/~hcohen/cohenpdf/colouringwithoutseeing.pdf.

Dartnall, Terry. 1994. *Artificial Intelligence and Creativity: An Interdisciplinary Approach*. 1st ed. Springer, August 31.

Dennett, Daniel C. 1991. *Consciousness Explained*. 1st ed. Little Brown & Co (T), October.

———. 2001. *The Mind's I: Fantasies and Reflections on Self & Soul*. 0th ed. Basic Books, January.

Fouts, Roger, and Stephen Tukel Mills. 1998. *Next of Kin: My Conversations with Chimpanzees*. Harper Paperbacks, September 1.

Goldenberg, Jacob, David Mazursky, and Sorin Solomon. 1999. Creative Sparks. *Science* 285, no. 5433. New Series (September 3): 1495-1496.

Griffin, Donald R. 2000. Scientific Approaches to Animal Consciousness. *American Zoologist* 40, no. 6 (December): 889-892.

Hofstadter, Douglas R. 1999. *Godel, Escher, Bach: An Eternal Golden Braid*. 20th ed. Basic Books, February 5.

———. 2001. The Turing Test: A Coffeehouse Conversation. In *The Mind's I*, 69-95. Basic Books, January.

Jackson, Frank. 1982. Epiphenomenal Qualia. *The Philosophical Quarterly* 32, no. 127 (April): 127-136.

Januszczak, Waldemar. 2005. Congo the chimpanzee - Times Online. *Times Online*. September 25. http://entertainment.timesonline.co.uk/tol/arts_and_entertainment/article569970.ece.

Krulwich, Robert. 2010. Lucy. *Radiolab*. WNYC, February 19. http://www.wnyc.org/shows/radiolab/.

McDermott, Drew. 2009. What Matters to a Machine? August 7.

Partridge, Derek, and Jon Rowe. 1994. *Computers and Creativity*. Ablex Publishing Corporation, November.

Saunders, Richard. 1980. *The World's Greatest Hoaxes*. Playboy.

Searle, John. 1982. Minds, Brains, and Programs. *Behavioral and Brain Sciences* 5, no. 02 (June): 339-341. doi:10.1017/S0140525X0001236X.

Singer, Peter. 2007. *The Way We Eat*. Point of Inquiry. February 9. http://www.pointofinquiry.org/peter_singer_the_way_we_eat/.

Tijus, Charles Albert. 1988. Cognitive Processes in Artistic Creation: Toward the Realization of a Creative Machine. *Leonardo* 21, no. 2: 167-172.

Turing, A. M. 1950. Computing Machinery and Intelligence. *Mind* 59, no. 236. New Series (October): 433-460.

de Vries, Lloyd. 2005. Dead Chimp's Art Sells Big - CBS News. News source. *CBS News*. June 20. http://www.cbsnews.com/stories/2005/06/20/entertainment/main703057.shtml.

Wolfe, George. 1983. Creative Computers: Do They "Think"? *Music Educators Journal* 69, no. 5 (January): 59-62.

Ziff, Paul. 1959. The Feelings of Robots. *Analysis* 19, no. 3 (January): 64-68.

*Anon. Pierre Brassau, Monkey Artist. http://www.museumofhoaxes.com/hoax/archive/permalink/pierre_brassau_monkey_artist/.

*Anon. 1964. Art: Zoo Story - TIME. Time Magazine. February 21. http://www.time.com/time/magazine/article/0,9171,870835,00.html?iid=chix-sphere.

BETWEEN MAC, MAN, & MANET:

an interdisciplinary exploration of artificial intelligence, art, and personhood

8584 WORDS/34 PAGES

JACQUELYN TRUONG

a senior thesis for completion of the bachelors program in the humanities

at yale university in the spring of 2010

many thanks to my advisor, Professor David Gelernter, who guided &
inspired me in the exploration of such a satisfying, yet impossibly rich, field of study

also, many thanks to the Davenport Mellon Forum Committee

and the Davenport Class of 2010 for allowing me
to present the topic of my thesis to my colleagues on March the second, 2010