Knowledge and Purpose in AI: From Thomistic Passions to Chalmers' Qualia

PCTS, GTU, Berkeley, November 3-4, 2017

Arvin M. Gouw

Artificial intelligence is the next frontier in technology. Given the technological advances in AI ranging from Alexa who can read us our kindle books, to the Deep Blue the best chess player in the world. With the enthusiasm, optimism, and scientific advances, it is presupposed that when it comes to intelligence, the sky is the limit for AIs (1, 2). Let us suppose that that is the case, now I would like to drag scientists back to the middle ages to see what intelligence is all about. Arguably, when we talk about intelligence, we're really talking about knowledge. But with knowledge, comes will and purpose.

Thus first, I will use Aquinas' theory of action to provide a framework for the discussing AI actions. Second, following Noren Herzfeld, I will discuss symbolic vs. embodied AIs (3, 4). Upon discussion of embodied AIs, it was then necessary for me to leave Thomistic framework of action and passions (emotions) to the third part of my paper that will discuss the necessity of qualia for real super intelligent AI to emerge.

PART 1: AQUINAS' THEORY OF KNOWLEDGE AND PURPOSE

Aquinas first argues that every human action is intended for some end, some telos. In other words, there is a purpose to any action. Every action is done for the sake of attaining some goal or realizing some desired state of affairs. To the obvious objection that we do all things all the time without any purpose at all, Aquinas replies by distinguishing between human action (actus humanus) and an action of a human being (action hominis) (5). Human actions, properly called, are defined as actions that proceed from their distinguishing power, through reason and will (6). Anything else that a human being does can be called the action of a human being, but not (in the proper sense) a human action. Human actions, then, are those that are willed on the basis of rational deliberation. And since "the object of the will is an end and a good, it follows that all human actions are for the sake of an end." (7)

What precisely is meant by the "end" of a human action? Aquinas tells us in a.2 that an end is something cognized as good. It must be something apprehended or *cognized* because otherwise it would fall into the simple category of "natural appetite", which is the innate directedness of inanimate things – i.e. that heavy objects fall down (8). An end must be perceived or cognized as good because the will is moved only by what is good. This is an Augustinian notion that the end cannot be perceived as evil because evil is nothing but a deprivation of good, thus having no real existence (9). Even the Devil himself cannot will evil, but only the deprivation of good (10, 11). As Eleonore Stump emphasizes, the will is not a neutral steering wheel; it is "an inclination for what is good, where the phrase 'what is good' is used attributively and not referentially."(6) And because only beings with the capacity for abstract thought can cognize what is good qua good (as opposed to merely cognizing something that is in fact good), only they can have will, the appetite that follows upon intellectual cognition of something as good. "The object of the will", Aquinas

says, “is the end and good universally”(12). By the end of any particular action will not be good-in-general, but some particular thing cognized as good. This end is what gives each particular action its species (a.3), what makes the action the kind of action it is. Since there is an end for every particular action, we can call this **particular end.**

IN a.4, Aquinas argues further that every human action has an **ultimate end**: that is, an end that, with respect to that particular action, is not itself for the sake of some further end. It is impossible for there to be an infinite series of ends, each of which moves the appetite only instrumentally, as a means to some more ultimate end (13). There has to be some mover that moves the appetite on its own account. This non-instrumental mover of the appetite is true ultimate end for the sake of which a particular action is performed(14).

Finally Aquinas argues that this ultimate explanation for all actions is the same for every human being. All human beings have the same ultimate end, for all human beings desire their own perfection, though different people will have different ideas about what perfection consists in. So Aquinas says that all human beings have the same ultimate end as far as the intelligible formula for the ultimate end goes (secundum rationem ultimi finis), but not in terms of the object that people think is aptly descrbed by that formula (secundum id in quo finis ultimi ratio invenitur) (8, 15).

To summarize: Aquinas argues in the first question of ST I-II that

a. All human actions are performed for the sake of an end
b. Any given human action has one and only one ultimate end
c. Ths one ultimate end is the same for all of a given human being’s actions
d. All human beings have the same ultimate end.

Two approaches to intelligence in AI

There is often a generalization regarding the notion of the capability of AI, as if given the unlimited resource of data from the internet, and unlimited memory, suddenly AI will be able to know everything, omniscient, or attaining the point of singularity. However, I believe it is helpful to not assume that the knowledge and purpose of AI are all the same. To construct a framework by which we can have a fruitful discussion, I will first use Herzfeld’s classification of AI, which are then further subdivided based on Thomistic theory of knowledge and purpose that was just discussed in the first section.

There are three approaches to intelligence in AI according to Noreen Herzfeld: symbolic, embodied, and relational AI (7). For our purposes, I will combine the latter two, thus discussing only symbolic and embodied AIs. First I will discuss what symbolic AI is and how the notion purpose becomes very important in symbolic AI. I will use Aquinas’ notion of purpose in the passions to help us discuss the potential implications of symbolic AI’s use of knowledge. Second, I will discuss what embodied AI is and how the notion of passions and qualia becomes crucial in determining whether this project is feasible.

PART 2: SYMBOLIC AND EMBODIED AIs

I. Symbolic AI

The first approach to designing AI machines assumed that intelligence is a substantial thing that we inherently have. That substantial thing is the ability to solve problems. This assumption is indeed indicative of intelligent people, i.e. the ability to play chess or solve complicated math problems. This symbolic AI notion is predicated that human thought could be represented by a set of basic facts which could then be combined, synthesized, according to set rules, into more complex ideas. This approach to AI has been called symbolic AI (7). It assumes thinking is basically an internal process of symbol manipulation.

Symbolic AI met with immediate success in areas in which problems can be described using a limited set of objects or concepts that operate in a highly rule-based manner. Game playing is an obvious example of one such area. The game of chess takes place in a world in which the only objects are the 32 pieces moving on a 64-square board and these objects are moved according to a limited number of rules. Other successes for symbolic AI occurred rapidly in similarly restricted domains, such as chemical analysis, medical diagnosis, and mathematics (16).

A. General vs. Specialized AI

I propose to further classify symbolic AI based on their ends. Based on Thomistic distinction between particular and ultimate end, we can argue that there are in fact two general categories of AIs in terms of their ends: sAI (specialized Artificial Intelligence) vs AGI (Artificial General Intelligence). sAI has particular ends, or specific purposes such as playing chess or performing chemical analysis. AGIs have an ultimate end encompassing multiple particular ends that are multi-layered, such as to improve human living condition.

Furthermore, for each category, there should be distinction regarding the capability or knowledge database that an AI has, regardless of whether it is an sAI or an AGI. Some AIs can have limited amount of knowledge that are preprogrammed to it, while others can have an unlimited resource of knowledge as in the case of approaching singularity. Thus we can see that the following is the possible combination of subtypes of AIs based on the notion of purpose and knowledge.

Purpose / Knowledge	Specific	General
Limited	Sai	Human
Unlimited	Sai	AGIs

Related to these two purpose categories of AIs are specific dangers. sAI has the problem of inability to consider more important goals. For example, Bostrom correctly argues it is possible that paper clip producing sAI will deplete all natural resources on earth just to make the most number paper clips (17). Given an unlimited amount of knowledge, an sAI would sacrifice many valuable resources for this particular end, because it knows not of any other end.

Pertaining to AGI's we encounter a different problem, which is how do we know which general purpose or ultimate end is better than another? Aquinas have argued that the ultimate good should be God himself, because he is the ultimate good, but due to limitation of man's reason, we don't always see that. When it comes to robots, it is unlikely that we will program robots just to be religious, because it's such an ambiguous purpose. Thus, how do we prioritize between improving human living condition vs. preserving natural resources or between maintaining safety vs. maintaining privacy. These are all issues which humans can't solve, how can we even decide what to program into these robots.

B. Angels and Robots

Suppose we consider a symbolic AGI that has infinite knowledge, ala singularity, thus they will know better than us which ultimate end is best to pursue. In this case, it might not be that simple, and we might run into problems that medievalists have considered in angelology. For afterall, angels are nothing but simpler creatures when compared to humans yet possessing far greater knowledge than humans.

Before we proceed in comparing symbolic AGIs to angels, let me first argue that there are several advantages to discussing philosophical issues in connection with angels. Indirectly, reflection on angels serves in philosophy as a thought experiment on how a particular idea can be part of the human condition. In other words, in order to know exactly what a lion is, we might start comparing it with another feline, like a cat. Our medievalist ancestors, in order to understand better the specific nature of man, therefore compared and contrasted him with the angel, his cousin in the order of spiritual beings (10). Today, the chimpanzee has replaced the angel in this role, and not to the benefit of the humanities. But when it comes to omniscient symbolic AI, it is time to compare them to angels again.

There are several similarities between symbolic AGIs and angels:

- We want to create them to help us
- They have more knowledge than us
- They are also metaphysically simpler than us, not embodied.

Comparing how the superlative of an attribute in human (knowledge, language, love) is achieved in angels allows us to determine its perfection and features that it assumes in man. Thus, for our case, the study of the angel's sin and error obliges theologians to go to the very heart of the notion of sin and error, toward the sin and error that is chemically pure (1, 12). In our case, by learning from angelic knowledge and sin, we can better see potential problems with omniscient symbolic AGIs.

C. Angelic and Robot's Fall

Now when it comes to considering the strength of infinite singularity knowledge of symbolic AI, we could learn from the folly of angels, or how angels fell into sin.

Angelic knowledge is non-discursive like symbolic AI

Aquinas denies that angelic knowledge is discursive. When discussing human reason (ratio), Aquinas distinguishes three operations or acts. The first operation is forming single concepts and the second one is forming propositions. These two belong to human reason secondarily but not primarily. The third act of reason belongs to reason primarily, and it involves reasoning (ratiocinari), that is, "to proceed" (discurrere) from one thing to another so as to come to knowledge of the unknown through what is known." The first two acts, the formation of single concepts and of propositions, relate to the third one, which reflects the process of coming to know (18).

In short, rational discursivity is the imperfect state of intellectual insight; it is intellect in the state of becoming. As an imperfect form of intellect and as typical of the human intellect, discursivity does not belong to angels. Angels see instantly (statim), with a simple intuition (intuition simplex), all that can be known are known by their own essence (10). Angels do not come to know gradually, little by little, but, rather they know at once. (ST 1a.58.3.). Thus in this manner, they are similar to symbolic AGIs who come to their knowledge 'pre-loaded' and not through learning.

In Aquinas' mature works, in ST, he explains that angels cannot repent because of internal determinism that comes from intellectual determinism (19). The premise is that "the appetitive power is in everything correspondent or in proportion to the cognitive power by which it is moved, like the movable to the moving cause."(10) Since the angelic superior intellect apprehends in a non-discursive manner, its knowledge is unchanging, because it has thought to have thought of every possible option in decision tree making. Thus, just as angelic knowledge does not change due to its limited omniscience, so the will's adhesion to that knowledge does not change.

In other words, if symbolic AGI in singularity had infinite knowledge, following the angelology of Aquinas, we could speculate that they would be intellectually determined and locked-in, such that we cannot convince or control these AGIs to do otherwise, even it meant the destruction of humanity.

Diagram 2: Dangers of symbolic sAI and AGI

Purpose / Knowledge	Specific	General
Limited	No danger	Being human
Unlimited	Prioritizing ultimate ends	Intellectual Determinism

Part of the solution to this problem is really to consider non-symbolic AI where their knowledge source are not all symbolic, thus robots are able to learn by trial and error and not instantenously omniscient and completely immutable.

II. Embodied AI

The second approach to designing AI is through embodiment that can interact with the environment. To act within an environment means to interact with both the material world and the human community. This means that, first of all, intelligence is embodied. Of course any intelligence agent would be embodied in some way. Deep Blue did not have what we would think of as a body; it could not pick up the chess pieces and physically move them. However, the program was embodied in a bank of supercomputers. So the question is not whether intelligence requires a physical body, but what kind of body (11). Does a human-like intelligence require a human-like body? Thus though embodied AIs are free from the dangers of unlimited knowledge (Diagram 2) because they're not preprogrammed, however they have a unique major problem, analogically, the body-mind problem.

Embodiment has always had its niche in the world of AI. Almost every artificially intelligent computer that has appeared in the realm of science fiction has been an android with a human-like body. In recent years prominent AI researchers, such as Rodney Brooks at MIT, have moved toward embodied AI robotics as well (20). Brooks has noted that the basic problems with symbolic AI are rooted in the fact the problem-solving programs it produces are not situated in the real world and that it is impossible to preprogram every possible scenario a robot could face in the environment (3). Hence they cannot learn from the continuity and the surprises that the real world presents. Brooks and others at a variety of AI labs, have built a series of robots that act within the world on the basis of data acquired through sensors. Brooks began with a series of insects, later moving on to the humanoid robots Cog and Kismet, which acquired some of the rudimentary skills of a baby through interaction with human beings (21). None of these robots come close to human-like intelligence, but some seem to have a niche in their environment. Consider the Roomba, a roboticized vacuum cleaner that navigates around a room looking for dirt, avoids furniture and stairs, and plugs itself in when it needs to be recharged. One might argue that Roomba shows as much intelligence as many animals, in its ability to navigate in a local environment, avoid hazards, and forage for sustenance.

A. The notion of passion as movement

In any embodied robot that is interacting with the environment, it is not sufficient to discuss simply the purpose or end that robot should have (as in symbolic AI), but it is also crucial that we discuss how robots interact with the environment while keeping in mind the aforementioned end. I will utilize Thomistic theory of passions as a model as one possibility in providing explanation for such interactions.

Passions are part of the sensitive appetite. Passions are considered as acts that are "passive" because they require an external agent to elicit them. However, passions are crucial in addition to mere comprehending powers of the intellect that simply grabs an object into the subject where the subject does not move to the object. I can comprehend whether an object is a good or evil by grabbing that object into the phantasms of my mind and analyzing it. However, it is my passions that compels me to move either towards or away from that comprehended object. It is the passions that move the subject in response to the object. But what do concrete objects have in common that makes them activators of passion? Aquinas explains that the general object of

appetite is the good, in so far as the end of any human action is also the ultimate good, as previously mentioned above.

Aquinas then distinguishes two kinds of objects of passion: desirable/good or not desirable/bad. With regards to the attainability of the objects themselves, there are two possibilities: easy or hard to attain. Passions that move in response to easily attainable objects are called concupiscible passions, while passions that move in response to difficult attainable objects are called irascible passions. Within concupiscible and irascible passions, there are also two kinds of movements; moving toward or away from the object. Thus, overall Aquinas says:

"In the concupiscible are three ordered pairs of passion, love and hate, desire and aversion, joy and sorrow. Similarly in the irascible are three groups: hope and despair, fear and courage, and anger to which no passion is opposed. Therefore all the passions differing in species are eleven in number, six in the concupiscible, and five in the irascible, under which all the passions of the soul are contained. (23.4.co)."

Thomas knows other passions exist, in questions 26-48, he discusses other kinds of feelings that are described as subspecies of the one of the basic eleven passions.

Diagram 3: Thomistic movement theory of the 11 passions

Passion/Object	GOOD			BAD		
CONCUPISCIBLE	Joy					Sorrow
IRASCIBLE		Hope	Despair	Fear	Courage & Anger	
CONCUPISCIBLE	Desire					Aversion & Hate

B. The passions of embodied robots

If we are truly to have a fully functional AI with free will that are independent and autonomous but also interactive, it also makes sense that these robots shall have 'passions' or ability to decide between moving towards or away a particular stimulus in the environment(22). Though no AI scientists would say that they're trying to program robotic passions, because passions are basically medieval terminology for what we today call emotions or feelings. Indeed, some AI researchers have caught up on the need for robots to feel, such that new robots are created so that they can feel pain, allowing packets of electrical signals to increase in intensity and frequency upon certain stimuli such as temperature and opposing force, mimicking neural impulses from pain nerve endings in humans(23). The motivation of training robots to feel pain is to allow them to avoid undesirable stimuli.

Though it is possible to program robots to feel pain and pleasure by programming positive and negative feedback loops to stimuli as models for reward and punishment, this would only replicate concupiscible passions, but not irascible passions. Irascible passions are by definition 'irrational' in the sense that despite the difficulty of the object, the agent moves *towards* it. This is not something that AIs can be programmed for, because arguably, the split judgment that people do which demonstrate irascible passions such as courage and anger, are not done through

rational reasoning. Most of our decisions when it comes to irascible passions are arguably based on feelings. If our rational calculation can justify the worthiness of the object despite the difficulty to attain it, then it loses the internal conflict that characterizes all irascible passions. Thus an AI that can experience irascible passions and overcome obstacles would need to be able to judge not based on intellect alone, but also based on feelings, or more properly speaking, qualia.

The need for robotic passions is the beginning of the problem with embodied AIs. In the Thomistic sense, any passion is perceived in the sensitive soul, but it is the soul nevertheless. Given Aquinas' hylomorphic notion of a person, thus the soul is the substantial form of a person providing formal cause to the subsistent matter in order to be a person. In contemporary philosophy of mind, such notion could find analogies in the notion of downward causation of the soul on the body if we are to use Clayton's strong emergence paradigm, and the notion of qualia if we are to use Chalmer's notion of supervenience, for example.

PART 3: QUALIA AND ARTIFICIAL INTELLIGENCE

I. Functional consciousness and qualia distinction

Chalmers argues that there are fundamentally two concepts of consciousness or mind. First, there is the functional, or psychological notion of the mind. This is called the functional notion because it understands consciousness as the inner workings of behavior. This is in fact the domain of cognitive neuroscience, and the domain of Thomistic passions. Chalmers then notes that there is a second concept of consciousness, which he calls the phenomenal consciousness. The phenomenal mind can be defined in its most basic sense as the way things feel to us. This is classically referred to as *qualia.* Both concepts explain consciousness, and neither should be understood as the correct concept of consciousness. But this distinction allows us to say that investigating the passions or the psychological consciousness, regardless of how complex it is, is still the *easy problem* when compared to understanding *qualia, the hard problem.* The reason for calling it the hard problem is that we have nothing in neuroscience right now that would be able to address why things feel the way they feel to us (24, 25)

Though many would easily concede to the psychological/phenomenal distinction, they might not easily agree that the phenomenal consciousness cannot be explained away via neuroscience. Many would argue that the hard problem is not much different from the easy problem given the appropriate resources and time to investigate it. In order to defend my position against this tendency to reduce qualia to the brain, I will need to discuss the concept of supervenience.

II. Logical vs. natural supervenience of consciousness and reductive materialism

Generally speaking, objects and properties in nature can be explained in multiple layers. For example, we can explain water biologically as a solvent. Water's biological properties as water can be explained chemically by the hydrogen bond interactions between the oxygen and hydrogen atoms in water. Furthermore, water's atomic structure and these interactions can be explained by force laws in physics. Thus in the reverse direction, it seems that a set of physical facts of an object will determine its chemical properties, and furthermore biological properties. Supervenience is therefore a more formal understanding of how "higher level" properties depend on "lower level" properties. If we want to better understand consciousness, we need to see to what extent consciousness supervenes on neuroscience laws, biochemical laws, and physical

laws. Chalmers and others propose many different kinds of supervenience (26, 27), but I will discuss only two that I think are most relevant to our discussion: logical supervenience and natural supervenience.

A. Logical Supervenience

Logical supervenience is also known as conceptual supervenience. Higher properties, or B-properties, are said to supervene logically on lower, A-properties if there can be no logical situation where the same A-properties yield different B-properties. This logical situation is not constrained by our natural laws. We can think of alternative situations existing in wholly different worlds in different universes with all different natural laws. This may seem as if there are no constrains to possibilities. Anything from flying phones to flying turtles would count as alternative possibilities. But, there are indeed things that are not logically possible, such as female bachelors, male hens, and round triangles. These are not logically possible because it would be a contradiction in terms for a bachelor to be female; bachelors are by definition unmarried males. It is perhaps easier to think of logical supervenience in terms of whether God could create a set of properties B that supervene on properties A under a different set of laws. Similarly, theologians like to argue that God is omnipotent but only omnipotent in the sense that he is able to do what is *logically* possible: create Adam with six arms, Moses with horns, etc. But God's omnipotence does not imply that God can sin, or create a round circle, or create a world with free will but without suffering.

B. Natural supervenience

Natural supervenience is different from logical supervenience in the sense that natural supervenience is more restrictive than logical supervenience. Natural supervenience, as the name implies, only deals with *natural* possibilities. Therefore there is a preferred frame of reference by which we judge which possibilities would count in the definition of natural supervenience. Overall, B-properties are said to naturally supervene on A-properties if there are no natural possibilities where different A-properties entail the same B-properties. For example, though it is logically possible for water to boil at a lower temperature than alcohol, this is not a natural possibility, because in our world with our natural laws, alcohol evaporates even at room temperature, 25°C, while water boils at 100°C due to the strength of the hydrogen bonds in water. Thus, a naturally possible situation is a situation that could actually occur in nature and which obeys all our natural laws.

Using God as a theoretical device, if B-properties supervene logically on A-properties, then God does not have to do anything else to manifest B-facts once he creates A-facts. In the case of natural supervenience, God still has to establish laws, natural laws once he establishes A-facts so that B-facts can become manifest. In the case of the Laplace demon, he cannot simply "read-off" B-facts once A-facts are established in the case of natural supervenience, unlike in logical supervenience.

C. Arguments for the natural supervenience of consciousness

There are two parts for the argument supporting the natural supervenience of consciousness. First, I will provide two thought experiments which reject the logical supervenience of

consciousness. In refuting the logical supervenience of consciousness, it suffices to demonstrate that there is a logical possibility that in a physically identical world to us, there exists a consciousness that is different than ours. This is what the first thought experiment is all about.

It is logically possible, to imagine, a world identical to ours, in which there is a twin, completely identical to me, whose experience of the color red is different than mine. When I see a color red, I have the feeling of 'redness', while when my twin sees the color red in his world (which is identical to ours), he has the feeling of 'blueness'. Though the perception of my twin's brain still perceives the color as the wavelength of the red color, it is still logically possible for him to have the feeling of "blueness" upon seeing the color red even though he would acknowledge with me that we see the color called red. This thought experiment is often called the "inverted spectra" argument. To attain such inversion in the actual world, some neural circuit rewiring would need to be done. But that is a natural possible situation, not a logical one. It is still logical to have the experiences inverted while the physical structure of my twin's brain is identical to mine. In other words, there is nothing in the wiring of the brain which dictates or could be 'read off' to give a feeling of red as opposed to blue.

A second thought experiment is different but it also refutes logical supervenience of consciousness. Suppose Mary is a neuroscientist who has grown up never seeing any color, only black and white. Being the neuroscientist that she is, she understands completely what it means to perceive various colors. However, since she grew up in a black and white world, she would not be able to know what it feels like to see the color red. Suppose she went to a room which had colors. She will gain some new 'knowledge', which is the feeling of seeing color. This thought experiment demonstrates that all physical information of color and neural processes can give her the *qualia.*

Now that we have presented two thought experiments which refute logical supervenience of consciousness, it is obvious from the aforementioned experiments that there is a mere natural supervenience of consciousness, not logical supervenience. In both cases, the physical facts alone (A properties) do not automatically give us (B facts) or qualia. There has to be something more than just the physical facts in this world for us to have qualia. This means there must be some type of additional laws. These are what Chalmers refers to as **psychophysical laws**.

Thus, in response to the two questions above pertaining to the power of neuroscience in explaining consciousness, the physical explanation of neuroscience is well suited only to the explanation of structure and of function. Structure and function can be 'read off' of physical properties. Thus in this sense, structure and function are reducible and materialism prevails in explaining structure and function. Cognitive neuroscience has explained much of the structure and function of consciousness as shown in the brief review before. But structure and function are both within the realm of psychological or functional consciousness, not in the realm of phenomenal consciousness. The neuroscientific research does not and cannot dictate why certain neural processes should give rise to qualia.

III. Psychophysical law of organizational invariance

From the previous sections we have discounted the power of neuroscience in explaining all aspects of consciousness, therefore we need a better theory of consciousness that is not reducible to neuroscience. As powerful as neuroscience is, psychoneural identity theory which claims that consciousness is identical to the brain must be refuted (25). Following Chalmers, I agree that the

notion of supervenience will be a necessary bridge to create a nonreductive theory of consciousness. Up to this point, we have rejected reductive, materialist point of view. We have accepted natural supervenience of consciousness, which means that there must be psychophysical laws which hold up the supervenience. Chalmers presented several psychophysical principles himself, one of which I will discuss now.

A. The Principle of Organizational Invariance

Chalmers stipulates that consciousness arises out of the functional organization of the brain. By functional organization, he means the abstract pattern of causal interaction of various parts within the system. Before presenting his arguments, let us first get some sense of what he is trying to demonstrate. Chalmers seeks to argue that when two systems have the same functional organization, or functional isomorphs, then they share qualitatively identical experiences. According to this principle, consciousness is organizationally invariant. Consciousness is a property that remains the same in all functional isomorphs, regardless of the underlying components of the system. The parts of the system could be silicones, plastics, or billiard balls. The lower levels do not matter as long as they are functionally isomorphic to us, then we will both share the same kind of qualia.

Chalmers presents his arguments for this principle of organizational invariance by using his fading, inverted, absent, and dancing qualia *counter*arguments. These qualia arguments can be divided into two groups for our purposes. The first group of arguments uses what is called absent qualia. An absent qualia is a situation where consciousness must be absent because of the underlying functional organization. For example, the underlying functional organization could be composed of citizens of India. Even if all citizens of India were functionally organized identical to our neural functional organization, it would be bizarre to believe that consciousness can arise out of such a functional organization. The second group of arguments uses what is called inverted qualia. An inverted qualia argument basically says that if our brain had been composed of other physical materials, then we would have a different kind of experience/qualia. These two classes of arguments are arguments *against* Chalmers' organizational invariance principle. Chalmers then proposes several counterarguments. Against the absent qualia arguments, he proposes a fading qualia argument, and against the inverted qualia arguments, he presents his dancing qualia argument.

B. Fading and Dancing Qualia Thought Experiments

First, the fading qualia argument goes as follows. If we were to replace one neuron in the brain with a functionally identical silicon neuron, there would be no reason for our qualia to change. If we were to replace two neurons with silicon neurons, it would be difficult to see why that would change our qualia as well. If absent qualia was true, when the brain is completely replaced by the silicon brain, then the qualia would be absent. This means if we keep sequentially replacing neurons in our brain with functionally isomorphic silicon neurons, then our qualia must be fading to the point that it disappears completely when all the neurons had been replaced with silicon neurons. It seems that this is an absurd concept, because how can *fading* qualia take place? It is implausible to have 20% qualia, 30% qualia, 50% qualia, etc. Therefore, Chalmers claims that if the fading qualia is implausible, the absent qualia arguments are false as well.

Against the inverted qualia argument, Chalmers designs a dancing qualia argument. Let us recall that an inverted qualia argument claims that it is possible for a functional isomorph to have an inverted qualia compared to us. Suppose this was true, and suppose there are two functionally isomorphic brains: one composed of neurons, and the second composed of silicon. A surgery is then performed connecting the visual cortex of the neural brain to the visual cortex of the silicon brain. The connection is made such that we can switch on and off whether we want to use the backup silicon brain's visual cortex. Being presented a red book, my neural brain has a red color qualia. Now if the inverted qualia argument were true, then upon switching to the silicon visual cortex, I would then experience a different qualia, perhaps the color blue. If we play around with the switch, and keep switching between the neural and the silicon visual cortices while staring at the red book, then there would be a dancing qualia of red and blue in my mind. This is certainly absurd, because the two visual cortices are identical in their processing mechanisms. There is no reason why upon switching 'wires' with the same connections, you would get a different qualia. Thus Chalmers claims if dancing qualia across functionally isomorphic brains is impossible, then inverted qualia arguments are to be rejected.

So this is the flow of logic in Chalmers' qualia arguments: if absent qualia are possible, then fading qualia are possible; if inverted qualia are possible, then dancing qualia are possible; and if absent qualia are possible, then dancing qualia are possible. Since it is implausible that fading and dancing qualia are possible, therefore absent and inverted qualia are implausible as well. It stands to reason then that the principle of functional organizational invariance is true.

Thus far, I have agreed with Chalmers on the natural supervenience of consciousness and the implausibility of logical supervenience of consciousness. This means I also believe that reductive materialism is false, and that some psychophysical laws are needed to construct a nonreductive theory of consciousness. Chalmers proposes several principles, and I chose one of his principles that are relevant to this thesis. Up to this point I only presented Chalmers' arguments for proposing the two principles. It is difficult to counterargue Chalmers' counterargument to the materialists' counterargument to Chalmers, because they are very detailed and complicated at this point. However, it does not mean that they are not testable.

C. Current Artificial Qualia Constructions

Cognitive scientists working on AIs and machine conscsiousness (MC) have in fact attempted to create artificial phenomenal consciousness, or qualia in their robotic systems. For example, Chella and Gaglio created an active process integrating internal and external flows of information that is designed to build a two-dimensional, viewer-dependent, field of view. The matching between this reconstruction and the internal perceptual data of the system represents the artificial qualia of the AI (28, 29).

Another significant work on constructing artificial qualia is H-CogAff, where a meta-management layer is added as a reflective process for the AI to express the virtual machine functions and operations (30, 31). Thus not only the machine is programmed to be interactive, but the AI contains a second layer of meta-management to fully express what it is like to do certain operations (i.e. 'what is it like to see a moving dot on a screen?').

D. Implementing fading and dancing qualia experiments as qualia detection methods

Thus it seems that for embodied AIs, it is important to discern the feasibility of qualia in these systems. Philosophers have had fun stipulating the various thought experiments ranging from angelic nondiscursive knowledge, angelic error/sin, to inverted and fading qualia thought experiments. But unlike the medieval angelic thought experiments, inverted qualia thought experiment is actually experimentally doable.

It is experimentally possible to determine if dancing qualia could take place. One does not need to create a perfectly silicon brain to compare. One can simply progressively replace the visual cortex of our brain with silicon chip, perhaps one neuron at a time, to determine if the qualia changes. Back to the previously mentioned thought experiment, being presented a red book, my neural brain has a red color qualia. Now if the inverted qualia argument were true, then at some point upon switching to the silicon visual neurons, I would then experience a different qualia, perhaps the color blue, or perhaps a novel qualia altogether. The qualia of a cyborg perhaps. If we play around with the switch on these silicon neurons, we can keep switching between the neural and the silicon visual cortical neurons while staring at the red book, then there would be a dancing qualia of red and blue in my mind. This experiment would determine if upon switching 'wires' with the same connections, you would get a different qualia. Though it is difficult to find wiling subjects for this experiment, it is possible that a patient with an inevitable degenerating visual cortex would be willing to volunteer. In fact, Princeton faculty members have filed a patent on this very experiment as a qualia diagnostic test for AIs (32, 33).

Au contraire, given an existing android with a silicon brain, assuming it's a humanoid brain, isomorphic to our neural brain, it would be possible to perform the experiment replacing their silicon neurons one by one with a human neuron and ask the android if the android 'feels' any different upon the human neural substitution. If this experiment was successful, then it is possible to identify if an android is conscious.

E. Concluding Remarks

In the first part of the paper, I argue that for any AI to be intelligent, it is crucial to consider the notion of knowledge and purpose in Herzfeld's classification of AIs into symbolic AI and embodied AI. The second part of the paper elaborates on applying the Thomistic framework of knowledge and purpose onto these two categories. First, pertaining to symbolic AI, using the Thomistic framework of knowledge and purpose, I classified symbolic AIs into sAI and AGI with limited or unlimited knowledge. Assessment of the dangers of sAI include failure to weigh in unspecified preprogrammed ends, while the dangers of AGI include the failure of prioritizing competing ultimate ends. Assessment of the dangers of unlimited knowledge by looking into medieval angelology as thought experiment, reveals that AIs may fall into intellectual determinism and unable to adapt. Second, pertaining to embodied AI, it is necessary to consider the notion of emotions as driving force behind the will of the robot interacting in the environment. Building on Thomistic theory of passions, the problem arises that embodied AIs will have a hard time exhibiting irascible passions due to the lack of qualia. The issue of qualia brings us to the discussion of qualia in the third part of the paper. Utilizing Chalmers' distinction between functional and phenomenal qualia, I argue that it is crucial that AIs exhibit phenomenal qualia. Following Chalmers, I demonstrate that the notion of natural supervenience demands the

principle of organizational invariance for phenomenal qualia to emerge. Chalmers proposed fading and dancing qualia thought experiments to defend his principle of organizional invariance and natural supervenience. A cursory look into the literature shows that AI scientists have indeed attempted to design an artificial qualia for AIs. However, they do not seem sufficient and I bring the discussion further by proposing modified thought experiments as real experiments that can be done as diagnostic tool for the existence of qualia in an AI. Yet, one can only wonder that the question one day will arise from an AI - How can I know that you are conscious?

BIBLIOGRAPHY

1. Kurzweil, R. 2012. *How to create a mind : the secret of human thought revealed*. New York: Viking. 336 p. pp.
2. Kurzweil, R. 1999. *The age of spiritual machines : when computers exceed human intelligence*. New York: Viking. xii, 388 p. pp.
3. Herzfeld, N.L. 2002. *In our image : artificial intelligence and the human spirit*. Minneapolis, MN: Fortress Press. xi, 135 p. pp.
4. Herzfeld, N.L. 2009. *Technology and religion : remaining human in a co-created world*. West Conshohocken, Pa.: Templeton Press. viii, 167 p. pp.
5. Davies, B., and Stump, E. 2011. *The Oxford handbook of Aquinas*. Oxford ; New York: Oxford University Press. xv, 589 p. pp.
6. Stump, E. 2003. *Aquinas*. London ; New York: Routledge. xx, 611 p. pp.
7. Kretzmann, N., and Stump, E. 1993. *The Cambridge companion to Aquinas*. Cambridge ; New York, NY, USA: Cambridge University Press. viii, 302 p. pp.
8. Miner, R.C. 2009. *Thomas Aquinas on the passions : a study of Summa theologiae : 1a2ae 22-48*. Cambridge, UK ; New York: Cambridge University Press. xi, 315 p. pp.
9. Thomas, Regan, R.J., and Davies, B. 2003. *On evil*. Oxford ; New York: Oxford University Press. xviii, 540 p. pp.
10. Bonino, S.-T. *Angels and demons : a Catholic introduction*. viii, 332 pages pp.
11. Kreeft, P. 1995. *Angels and demons : what do we really know about them?* San Francisco: Ignatius Press. 157 p. pp.
12. Gondreau, P. 2009. *The passions of Christ's soul in the theology of St. Thomas Aquinas*. Scranton: University of Scranton Press. 516 p. pp.
13. Gilson, E. 1994. *The Christian philosophy of St. Thomas Aquinas*. Notre Dame, Ind.: University of Notre Dame Press. x, 502 p. pp.
14. Gilson, E., Maurer, A.A., Shook, L.K., and Pontifical Institute of Mediaeval Studies. 2002. *Thomism : the philosophy of Thomas Aquinas*. Toronto, Ont.: Pontifical Institute of Mediaeval Studies. xiv, 454 p. pp.
15. Lombardo, N.E. 2011. *The logic of desire : Aquinas on emotion*. Washington, D.C.: Catholic University of America Press. xiii, 319 p. pp.
16. Hansell, G.R., Grassie, W., Blackford, R., Bostrom, N., and Dupuy, J.-P. 2011. *H± : transhumanism and its critics*. Philadelphia, PA: Metanexus Institute. 278 p. pp.
17. Bostrom, N., and Ćirković, M.M. 2008. *Global catastrophic risks*. Oxford ; New York: Oxford University Press. xxii, 554 p. pp.

18. Davies, B. 1992. *The thought of Thomas Aquinas*. Oxford
New York: Clarendon Press ;
Oxford University Press. xv, 391 p. pp.
19. McCabe, H., Davies, B., and Eagleton, T. 2010. *God and evil in the theology of St Thomas Aquinas*. London ; New York: Continuum. xviii, 205 p. pp.
20. Brooks, R.A. 2002. *Robot : the future of flesh and machines*. London ; New York, N.Y.: Allen Lane. x, 260 p. pp.
21. Boden, M.A. 2006. *Mind as machine : a history of cognitive science*. Oxford
New York: Clarendon Press ;
Oxford University Press.
22. Trappl, R., Petta, P., and Payr, S. 2002. *Emotions in humans and artifacts*. Cambridge, Mass.: MIT Press. viii, 390 p. pp.
23. Fellous, J.-M., and Arbib, M.A. 2005. *Who needs emotions? : the brain meets the robot*. Oxford ; New York: Oxford University Press. xv, 399 p. pp.
24. Chalmers, D.J. 1996. *The conscious mind : in search of a fundamental theory*. New York: Oxford University Press. xvii, 414 p. pp.
25. Kim, J. 2011. *Philosophy of mind*. Boulder, CO: Westview Press. x, 374 p. pp.
26. Murphy, N.C., and Knight, C.C. 2010. *Human identity at the intersection of science, technology, and religion*. Burlington, VT: Ashgate Pub. viii, 243 p. pp.
27. Murphy, N.C., and Stoeger, W.R. 2007. *Evolution and emergence : systems, organisms, persons*. Oxford; New York: Oxford University Press. xii, 378 p. pp.
28. Gaglio, S., and Lo Re, G. *Advances onto the Internet of Things : how ontologies make the Internet of Things meaningful*. ix, 352 pages pp.
29. Chella, A., and Manzotti, R. 2007. *Artificial consciousness*. Exeter ; Charlottesville, VA: Imprint Academic. 284 p. pp.
30. Society for the Study of Artificial Intelligence and Simulation of Behaviour. Conference (9th : 1993 : University of Birmingham), and Sloman, A. 1993. *Prospects for artificial intelligence : proceedings of AISB93, the ninth biennial conference of the Society for the Study of Artificial Intelligence and Simulation of Behaviour, 29 March-2 April 1993, The University of Birmingham*. Amsterdam ; Washington, DC: IOS Press. viii, 291 p. pp.
31. Sloman, A. 1978. *The computer revolution in philosophy : philosophy, science, and models of mind*. Hassocks Eng.: Harvester Press. xvi, 304 p. pp.
32. Schneider, S. 2009. *Science fiction and philosophy : from time travel to superintelligence*. Chichester, U.K. ; Malden, MA: Wiley-Blackwell. x, 350 p. pp.
33. Velmans, M., and Schneider, S. 2007. *The Blackwell companion to consciousness*. Malden, MA ; Oxford: Blackwell Pub. xviii, 744 p. pp.