2018

In the Mind of the Machine

Marcia Yang

Claremont McKenna College

In the Mind of the Machine

submitted to
Professor Dustin Locke

by
Marcia Yang

for
Senior Thesis
Spring 2018
April 23, 2018

Acknowledgements

I am grateful to Professor Dustin Locke for guiding me through many false starts and dead ends, for reading countless versions of this thesis, for helping me through an area of which he is not entirely familiar with, and for supporting me throughout the entire process. I would also like to thank the entire Philosophy department for being open when I would go down the hall asking for help as I went through my many false starts, for always making philosophy fun an engaging, both in and out of class, and for always encouraging me to dive deeper within the discipline.

I would also like to thank Professor Lisa Koch for always supporting me throughout the process. I will forever be grateful the that Keck Center for International and Strategic Studies assigned me to be her research assistant. I thank her for always supporting me through my pursuits, reminding me to take care of myself when I became overwhelmed, and for being so understanding. I would also like to thank the Keck Center, the Mgrublian Center and the International Relations Department for teaching me valuable skills and giving me many opportunities for research.

Finally, I would like to thank my friends and family for supporting me through all my endeavors and giving me a community to go to whenever thesis was getting to be too much. I thank my friends for spending late nights with me in the computer lab and at our office, enduring my late-night thoughts and rants. I thank my parents for all their sacrifices and hard work raising me and inspiring me to pursue academia.

Table of Contents

Abstract 1

I. Introduction 2

II. Nussbaum’s View 3

III. Delancey’s First Objection 11

IV. Reponse to Delancey’s First Objection 12

V. Delancey’s Second Objection 15

VI. Response to Delancey’s Second Objection 16

VII. Emotions as Representations 20

VIII. Implications 24

IX. Conclusion 28

Works Cited 31

Abstract

As technology becomes more sophisticated, it becomes increasingly important to understand how we should ethically use technology. One question within this area of study is whether we should treat certain types of technology, like artificial intelligence, with more respect. If we do owe these machines some sort of moral status, another question is what level of moral status they have. In order to answer these questions, I argue that machines can be considered as minds under the view of machine functionalism. A significant problem for machine functionalism is whether it can account for emotions within the system it suggests. First, I argue that emotions are able to fit within the system using Martha Nussbaum's framework for emotions. Second, I address Craig Delancey's objections to Nussbaum's view, and I suggest friendly amendments to Nussbaum's framework. Third, I look at how Nussbaum's view fits within the theory of machine functionalism. And finally, I consider the implications of the view that machines can have minds like humans can, and how we should treat machines in light of this argument.

I. Introduction

In philosophy of the mind, there is an ongoing debate about what can be reduced to representations. Whether something can be reduced to a representation is important because it has implications for how our minds and bodies interact, and how the mind itself functions. It has further implications for what should be considered a mind. The larger question that I will tackle in this thesis is whether machines can be considered to have a mind like a human does. If they do have a mind like humans do, then we need to reassess how we treat machines. This does not mean that we should treat all machines like humans, but that we should consider our actions to very advanced machines. In particular, if artificial intelligence is considered to have a mind like a human does, then we should not use artificial intelligence for our own means. We should treat it like we would treat an adult human. This means that we should not use it to advance our own good if we would not use a human in a similar way. We are currently at a point in technological advancement where we tend to think more about what we can do to improve human lives rather than about the possible cost to machine lives. One suggestion that I have from this thesis is that if a machine can meet all the necessary and sufficient conditions for a mind, then we should treat it like we would a human.

Before getting to these larger questions, we must first think about whether a machine can possibly be a mind. The leading theory that allows for machines to have minds like humans have is machine functionalism. Machine functionalism states that minds are processors of representations, and the kind of representation relevant to mind are those that have specific functions and causal relationships. Some disagree with machine functionalism on the basis that it does not properly account for emotions. If emotions are not capable of being reduced to representations, then emotions cannot fit within the system that machine functionalism describes. To answer this question of whether emotions can be reduced to representations, I will first analyze Nussbaum's view that emotions are the acceptance of certain types of beliefs and Craig Delancey's objections to her view. If Nussbaum is correct, then emotions are reducible to beliefs, which are the same as representations. Therefore, emotions can fit within the system machine functionalism describes. If machine functionalism is able to give a robust account of the mind, then machines are able to be considered minds as well.

The structure of the thesis will be as follows. First, I will reconstruct Nussbaum's argument for why emotions are the same as the appreciation of a certain type of belief. Second, I will recount Delancey's position and objection to Nussbaum's argument, as well as other possible objections to Nussbaum's view. I will introduce my own arguments against Delancey's view, and how Nussbaum's view can be expanded in relevant ways. Third, I will analyze how Nussbaum's view fits into the larger argument that emotions can be reduced to representations, and what this means for the theory of machine functionalism. Finally, I will discuss some implications of machine functionalism that stem from the idea that computers can be minds.

II. Nussbaum's view

In this next section, I will further elaborate what is meant by emotions and an identical cognitive belief in Nussbaum's paper "The Stoics on the Extirpation of the Passions". I will begin this section by putting Nussbaum's argument into standard form.

Then, I will go through each of the steps of Nussbaum's argument and the reasoning for each of the moves that she makes.

Nussbaum argues that emotions are identical to certain types of beliefs. These beliefs, which will be referred to as v-liefs, are beliefs that assign value to a vulnerable external good whose state is immediately relevant to me (Nussbaum 141). The state becoming relevant to me can mean that the external good has changed or that its presence is at the forefront of my mind. Either way, it means that not only the external good is of value, but that I also care about the state of the external good. For example, a v-lief could be my belief that my dog who is extremely important to me has died. The dog would be the external good, and it has value to me. It is vulnerable because it is subject to outside forces, such as the environment. The state is relevant to me because the dog has changed from living to dead. The emotion from this v-lief would be sadness from coming to accept my dog's death. If I was feeling happy about my dog being alive, this happiness would not involve the state of my dog changing. The state of my dog has not changed because my dog is still alive. When I am happy about my dog, it is because my dog being alive is relevant to me. Nussbaum's argument in standard form is as follows:

1. V-liefs are judgements. (p)
2. Someone holds a judgement when they accept it (Nussbaum 146). (p)
3. Therefore, someone has a v-lief when they accept it. (1, 2)
4. If an emotion is grounded in a v-lief, then the v-lief is at least part of the emotion (141). (p)
5. Emotions are grounded on v-liefs (141). (p)
6. Therefore, the v-lief is at least part of the emotion. (4, 5)

7. If the v-lief was only one part of an emotion, then this would imply that we assess the belief without any emotion (153). (p)
8. We do not assess v-liefs without any emotion (153). (p)
9. Therefore, v-liefs are not only one part of an emotion. (7, 8)
10. If v-liefs are not one part of the emotion and the v-lief is at least part of emotion, then they are the same as the emotion (154). (p)
11. Emotions are the same as the v-lief. (6, 9, 10)
12. Emotions are the same as accepting a v-lief. (3, 11)

Nussbaum begins her paper by talking about the Stoics, the name for the philosophers who flourished during the ancient Greek and Roman empires. Because she is beginning from this tradition, her definitions and arguments are derived from discussions of the Stoics. She specifically focuses on the work of Chrysippus in this paper. Nussbaum's definition of which beliefs emotions are based on comes from Chrysippus' work.

In the definition of v-liefs, Nussbaum argues that they are necessary for emotions. E-iefs are necessary for emotions because emotions do not occur without beliefs, and without these beliefs emotions would be no different from other appetites. Unlike appetites such as thirst or hunger, emotions are not simply our base instincts and desires (Nussbaum 140). Therefore, there must be something that explains why emotions are not like thirst and hunger. The Stoics argue that this difference exists because emotions have an important cognitive element (140). Because emotions have this cognitive element, they can be evaluated like a belief can. If an emotion can be evaluated, it can be rational and irrational or true and false, which seems to align with how we usually judge emotions (140). Therefore, emotions require this cognitive part, which Nussbaum then defines.

V-liefs have three component parts: they are evaluative, they assign value to something, and they are a belief about vulnerable external goods (Nussbaum 141). The belief must be evaluative because it involves assigning values to something. A belief that is not evaluative but also judges something’s value is contradictory. Therefore, the belief that an emotion is based on must be evaluative.

The belief must assign value to something. because we generally do not have emotions about something that does not matter to us. For instance, let us say that I have a belief that assigns value to something and a belief that does not. The first belief assigns importance to my dog. The belief that does not have value to me is that it will rain today. Because I am more invested in one belief than the other, I will naturally be more emotional about one belief over another. This is because it would be difficult for me to become emotional over something that has no value to me. For instance, if my belief about the weather was wrong and it turned out that it was sunny today, I would not be very emotional about it. In comparison, if I have the belief that my dog is important, then I would be emotional if something were to happen to my dog. Therefore, it is reasonable to say that emotions are based on beliefs which assign value to external goods. An implication of this view is that the intensity of our emotions is affected by how valuable something is to us. For example, my dog is extremely important to me, and my pencil may be somewhat important to me. If I lose either, I would be sad, but because I value one more than the other I would be sadder about one instance of losing something than the other.

The belief must be about a vulnerable external good because they must be able to be affected by fortune. Nussbaum does not give a very thorough explanation for why it must be based on vulnerable external goods, so I will try to expand her argument. If something was in complete control by us, then we would not be as emotional about it because we would not be emotional about it. To be emotional about something involves accepting new beliefs that we did not expect. But, if we are in complete control over something, there are no new beliefs that we did not accept. Emotions are the process of accepting a belief. If we have complete control, then we do not have to undergo the process of accepting the belief.

The first premise is implied from the text. Nussbaum gives an account of how the Stoics define judgements, and from there argues that v-liefs work in the same way. Therefore, it is implied that judgements and beliefs are identical. A judgement is a type of belief because having a judgement is the same as having a value-laden belief about something. All v-liefs which emotions are based on are judgements, but not all judgements are this kind of v-lief. This is because v-liefs are not only evaluative but also descriptive about the state of something. If a judgement is both evaluative and descriptive, then it would be considered a v-lief. If a judgement is only evaluative, then it does not meet the requirements of a v-lief.

The second premise comes from the Stoic's definition of a judgement. The Stoics define a judgement as an assessment of an appearance (Nussbaum 146). This means that forming a judgement is a two-part process. First, one must form a belief about something. Second, one rejects or accepts this belief. After a belief has gone through this process, it can now be considered a judgement. For example, I form the belief that the sky is blue. Then, I accept or reject this belief. However, I decide to evaluate the belief is the judgement. If I choose to accept that the sky is blue, then the judgement is my acceptance of the belief. If I choose to reject that the sky is blue, then the judgement is the acceptance that the sky is not blue. Either way, the judgement is framed as the acceptance of some belief.

The third step of this argument is an inference of the first two steps. Because all v-liefs are some form of judgements, v-liefs are actually the acceptance of some belief that assigns value to a vulnerable external good. To hold a v-lief means to accept it.

The fourth step begins Nussbaum's argument for why v-liefs are at least a part of the emotion. This step is necessary because otherwise, one could argue that v-liefs are necessary for emotions, but they are not part of the emotion. For example, kidneys are necessary for the heart to function, but this does not mean that kidneys are included in our definition of the heart. If v-liefs are not at least part of emotion, then they would function in a definition of emotions similar to how kidneys function in a definition of the heart. In other words, v-liefs would be considered completely separate from emotions. Nussbaum believes that this relationship between these two claims exists because when a thing *x* is based on another thing *y*, *y* is included within the definition of *x*. For example, my belief that I love chocolate is based on my belief that chocolate is good. I would not be able to have the belief that I love chocolate unless the belief that chocolate tastes good is included within the belief. This is because the first belief implies the second within it. The belief that I love chocolate cannot imply the belief that chocolate tastes bad or that I have no belief about how chocolate tastes. These other beliefs are contradictory to the belief that the original is based on.

In the next step, Nussbaum argues that emotions are grounded in v-liefs. One way we can tell that this is true is that we believe emotions can be evaluated. In other words, we believe that emotions can be true or false, rational or irrational. This would mean that emotions could not be determined to be rational or irrational, or true or false (Nussbaum 140). Because we believe that emotions are able to be evaluated, they must have some evaluative component to them. Nussbaum asserts that this evaluative component is the v-lief.

The sixth step is an intermediate conclusion from an inference of the fourth and fifth steps. The conclusion that v-liefs are at least part of emotions implies that v-liefs are also necessary for emotions. This conclusion is necessary for the overall argument because if Nussbaum only argued that v-liefs are necessary for emotions but not at least part of them, it would leave open the possibility that v-liefs are not the same as emotions.

The seventh step of the standard form begins a new section of the argument. Nussbaum introduces an argument for why v-liefs cannot be only one part of the emotion. She argues that if v-liefs were only one part of the emotion, then this would imply that the v-lief and the emotion are completely separate. If the v-lief and emotion were completely separate, or one was grounded on the other, then one would first accept the belief without emotion. Then, they would become emotional after understanding the belief. For example, my important dog has died. If the cognitive aspect is separate from emotions entirely, then I would coldly assess the belief that my dog had died. After emotionlessly assessing this belief, I would then start grieving my late dog. However, Nussbaum argues that this is not how the process works.

The eighth step of the premise argues that there is something flawed in this account of emotions and v-liefs. Nussbaum argues that if the v-liefs are only part of the emotions, then the emotions and reasoning are completely separate in the process.

However, this does not seem to be what happens in most cases. Going back to the example of the parent who has passed, I do not emotionlessly consider the belief that my parent has died. In fact, the opposite is true. What occurs in these situations is that I am emotionally coming to accept the belief. If this is not the process through which we become emotional about something, then Nussbaum must give an account for what she believes the process is.

From the seventh and eighth steps, we can conclude that the v-lief is not only part of the emotion. There are two ways to move from this reasoning: (1) Emotions have no v-liefs, or (2) Emotions are the same as the v-liefs. Nussbaum argues for the second option. This is because she earlier argued that v-liefs are at least a part of emotions. If v-liefs were not at least a part of emotions, then she would be able to follow the first option. This leads us to step eleven, which concludes that emotions are the same as v-liefs.

Nussbaum believes that emotions are the same as v-liefs because emotions and v-liefs cannot be separate. In the late dog case, Nussbaum believes that I am emotionally considering the belief that my dog has died. Because emotions and v-liefs are not separate during this process, Nussbaum argues that emotions and having a v-lief is the same. More specifically, she wants to argue that emotions and accepting a v-lief is the same.

Nussbaum makes this final step because to have a v-lief means to come to accept the v-lief. It seems that what leads to emotion is not having the belief in of itself, but the process of fully recognizing the belief is what causes emotional upheaval (Nussbaum 153). The process of acceptance is not preparing oneself for the emotional upheaval to come, but it is itself the emotional upheaval. For example, when I am accepting the death of my parent, it is the process of acceptance that is also the process of grief. These two processes are one and the same. Therefore, the emotion and the acceptance of the v-lief are one and the same.

III. Delancey's First Objection

Now that I have fleshed out Nussbaum's argument for why emotions and v-liefs are the same, in this section I will look at objections to the view. This section will focus on Delancey's first objection to Nussbaum's argument. Delancey argues that v-liefs are not at least a part of emotion, and that one can have v-liefs without having emotions.

Delancey argues that emotions and v-liefs are not actually the same. Nussbaum is making a flawed inference when she concludes that emotions and v-liefs are the same because the intermediate conclusion that v-liefs are at least part of emotions is false. This intermediate conclusion is false because it is based on the premise that if emotions are grounded on v-liefs, then v-liefs must at least be part of emotions. Delancey argues that there are cases in which one can have a v-lief that is necessary for emotions, but the subject does not have any emotion. Delancey suggests that the reason there is a disconnect is because Nussbaum's argument does not properly take into account the effect of the extended body on the mind.

Delancey defines the extended body as the method through which the mind can interact with the environment and take in information. For example, the extended body for humans is our physical, biological body. The extended body acts as a two-way street between the mind and the environment, but it itself can still be affected. Delancey argues that the extended body can be affected in ways that affect how the mind experiences emotion and v-liefs.

The strongest case Delancey gives of how the extended body affects how we experience emotions is of a patient who has undergone brain surgery. After brain surgery, the patient retains all the cognitive, reasoning aspects of their brain, but they no longer react to emotional stimuli as they used to. The patient underwent a test where the testers would see whether they would respond emotionally. Although the patient understood that they should react in some emotional way to certain stimuli, they felt that they could not. The patient seems to have all the relevant cognitive functions and even seems to form v-liefs, but they do not have any emotional aspect with these v-liefs (Delancey 244).

This case is supposed to be a defeater for Nussbaum because it is supposed to illustrate that emotions and v-liefs are separate. If they are separable, then that directly contradicts Nussbaum's claim that they are the same. Therefore, Nussbaum is making a flawed inference in the eighth step of her argument. Furthermore, this case shows that Nussbaum should take more seriously the role of the extended body in her argument.

IV. Response to Delancey's First Objection

In this next section, I will argue that Delancey's counterexample is not problematic for Nussbaum's view. First, I will bring up possible problems with Delancey's view that ultimately do not work out. Then, I will introduce my own objection to Delancey's argument. I will also acknowledge a possible response to my view, but I will explain why this response is not actually a problem for my view.

A problem for Delancey's view that ultimately may focus on the reason that they are separated as important. The case Delancey introduces is an exceptional case in which someone is not properly able to access their emotions. The patient understands that they should have some sort of emotional reaction to specific stimuli, but because of the brain surgery they do not. This shows that if the patient were able to access their emotions, then their emotions and their v-liefs would be identical. Rather than a case where emotions and v-liefs are separable, this case demonstrates that the extended body can affect the relationships between emotions and v-liefs.

Delancey may argue against this objection saying that the mere fact that someone can have one but not the other shows that emotions and v-liefs are one and the same. If they were truly one and the same, then the absence of emotions would mean the absence of v-liefs as well. Even if one were not accessible to the mind, that should mean that the other should not be accessible as well. Yet, in the case of the patient who cannot feel emotions, they still have the v-liefs. Therefore, emotions and v-liefs would be quantitatively identical, not numerically identical.

To change Nussbaum's view to say that v-liefs and emotions are qualitatively identical and not numerically identical would be to drastically change the view. If emotions and v-liefs are qualitatively and not numerically identical, then Nussbaum's objection that we do not coolly judge our v-liefs and then become emotional about them applies to this case. Therefore, we should find a way to maintain emotions and v-liefs as numerically identical.

I believe that a larger problem for Delancey's argument is that the kinds of v-liefs that are formed by the patient no longer meet the requirements of v-liefs necessary for emotions. The three conditions about v-liefs is that they assign a value to vulnerable external goods (Nussbaum 141). In this patient's case, the beliefs still seem to be about vulnerable external goods. What seems to have changed is whether the patient assigns a value to beliefs that they had assigned value to before. Although the patient seems to have all the cognitive aspects of the v-liefs and nothing seems wrong with their cognition, it seems that the emotional stimuli are no longer as important to the patient as they were before. If an external good is not of value to us, then we would normally not be very emotional about it. Therefore, rather than v-liefs and emotions no longer being numerically identical, it seems that these kinds of v-liefs that the patient has is not like those Nussbaum had thought were relevant.

One could respond and say that it is incorrect to say that the patient no longer assigns value to the belief. To say that the patient no longer seriously values anything they did before seems extreme. Even if the patient has undergone surgery that leads to them not valuing things as they did before, there must be some things that the patient would still value. In the study, the patient themself recognizes that they should assign value to some things (Delancey 244). The fact that they recognize that they should assign value to some things shows that it may just be that they do not assign value to things within the scope of the study.

In this case, the number of vulnerable external goods that are valuable to the patient have greatly decreased. Because we do not know how the patient treated things outside of the study, it is difficult to say whether they experience emotions or assign value to things. Further, acknowledging that they should assign value to things does not mean that they actually assign value to it. For example, although I may acknowledge that

I should assign value to my economics grade, but this does not mean that I actually do. Without more information on how the patient is outside of the study, it is difficult to tell whether this is a true case of someone having v-liefs without emotions. For now, it seems that the patient does not have v-liefs or emotions, but it does not follow that emotions and v-liefs are separate.

V. Delancey's Second Objection

The second objection that Delancey brings up is that some emotions are not based on the acceptance of a v-lief. This objection is focused on the final step that emotions are the same as the acceptance of a v-lief. Delancey points out that Nussbaum is ignoring the possibility that emotions are not always the same as the acceptance of a v-lief. In particular, an emotion can be the same as the suppression of a v-lief.

Delancey brings up the point that under Nussbaum's view, emotions like anxiety would be the same as the acceptance of a v-lief. Yet, Delancey asserts that anxiety is derived from the suppression of rather than the acceptance of a belief (Delancey 242). Acceptance of a v-lief implies that the belief is clear to the holder of the belief, and they are able to judge whether the belief is true or false. But, when people experience anxiety, the belief is not clear to the holder, and they are not able to judge the belief as true or false. Rather, when someone is experiencing anxiety, then the v-liefs are not easily accessible to them. If we follow Nussbaum's framework, then this would mean that anxiety would not be considered an emotion.

This objection brings up the problem that Nussbaum's view may be too narrow in what it considers to be emotions. Anxiety, depression, jealousy are all emotions that do not seem to be the same as the acceptance of a v-lief. For example, when my brother is jealous that his partner is talking to another person, the v-lief seems to be that his partner who is important to him is treating someone else nicely. However, jealousy is not the same as accepting this belief. Therefore, Nussbaum's theory does not include emotions that we may sometimes consider to be irrational. Yet, we would not consider these emotions to be similar to what Nussbaum considers appetites, such as hunger or thirst. Nussbaum must then find a way to accommodate these emotions within her existing view, find another way to explain how these emotions can be beliefs, or give an account for why these more irrational emotions are the same as hunger and thirst.

VI. Response to Delancey's Second Objection

Similar to the third section, I will begin this section with a possible answer to Delancey's problem that will not work in the end. I will show how Delancey would respond to this first possible objection. Then, I will discuss how Nussbaum's framework can be expanded to include emotions that are sometimes considered to be more irrational. I will argue that an emotion can either be the acceptance or suppression of a v-lief. Then, I will discuss why this modification of Nussbaum's view still works in conjunction with the rest of her argument.

A possible defeater to Delancey's objection is that in the case of anxiety, the v-lief is not being suppressed, but a different v-lief is being accepted. In Nussbaum's argument, beliefs can be accepted or rejected. But, Nussbaum's argument specifically says that v-liefs can only be accepted. If they are rejected, then another v-lief is being accepted. This same framework can be applied to anxiety. For example, if Alex is experiencing anxiety, it could be framed as the suppression of the v-lief could be that Alex is worthy of the job that they want. Nussbaum would rather frame it as the acceptance of the v-lief that Alex is not worthy of the job that they want. This difference may seem to be negligible, but it makes a difference in how we think of the process. The process of suppression is different compared the to process of acceptance. The process of acceptance seems to have an end where we come to fully accept some belief. The process of suppression seems to have no end because there is no point in which the belief will no longer be relevant if the external good is of value.

Delancey could respond by saying that this defeater does not work because the processes are not actually that different. There can also be an end to the process of suppression, namely that there will come a point in which the belief is no longer relevant to the person. For example, if Alex is anxious about an interview, the emotion could be based on the suppression of the v-lief that the interview may go well. This v-lief becomes irrelevant to Alex once the interview is over. In that way, there is an end to some processes of suppression. Therefore, the process of acceptance and the process of suppression are not relevantly different when it involves v-liefs. Thus, emotions can also be the suppression of v-lief. This seems to be how some emotions work as well. It would be strange to say that anxiety about getting a job is the same as accepting that I am not worthy of the job. The anxiety does not end when I have fully accepted the v-lief. That would mean that I have fully accepted the belief that I do not deserve the job. Full acceptance of this v-lief seems like it would lead to more anxiety rather than stopping it. When I am no longer anxious about getting the job, it is because v-lief is no longer as relevant to me as it once was. This seems more similar to the suppression of a v-lief rather than the acceptance of it.

Although Nussbaum's framework specifically says that emotions are the acceptance of a v-lief, I believe that suppression of v-liefs can also fit within this framework. The reason Nussbaum argues that emotions are the acceptance of v-liefs seems to be because when we have emotions, we go through a process of building up the emotion, the climax, and the completion of the emotion. For example, when I am sad about my dog's death, I go through the various steps of grief, beginning with denial. When I go through the grieving process, I am also coming to terms and accepting my dog's death. This also works with other emotions, such as anger. When I am angry because someone cut me off in traffic, I begin surprised that someone has cut me off, and I become angry because of the event. Eventually, I fully accept the belief that I was cut off in traffic, and my anger fades.

If the reason Nussbaum considers emotions to be the same as the acceptance of a v-lief because they have similar processes, then suppression can also fit this process as well. Going back to the case of anxiety about getting a job, Alex starts out a bit anxious because Alex not believe that they are qualified for the job they want. Alex becomes more anxious as they think about this v-lief, because the belief that they am qualified for the job is further suppressed. The emotion and v-lief find closure because eventually the v-lief is no longer as important to Alex as it once was. Once the v-lief is no longer as pertinent to Alex, then the anxiety also fades. This example shows how suppression of a v-lief can also have the three stages that acceptance has: building up, climax, and closure. Therefore, suppression of a v-lief should be able to be the same as emotion.

When Nussbaum’s framework is expanded in this way, we can give more thorough accounts for some emotions. Specifically, emotions that may seem irrational can be given rational accounts. As elaborated before, anxiety fits better in a framework of suppression rather than acceptance. Fear may fit better in this framework as well. For example, if I am afraid of heights and I am in a skyscraper looking down, then the fear may be the same as the suppression of the v-lief that I am actually safe. If we could only use a framework of accepting v-liefs, then the fear would be the same as acceptance of the v-lief that I am not safe. However, the fear would not end with the acceptance of this v-lief, but with the suppression of the opposite v-lief until I am in a position where the v-lief is no longer significant to me.

Nussbaum may hesitate about expanding her view in to include suppression of v-liefs as being equivalent to emotions because it gives rational explanations for emotions that we may see as irrational. In fact, she may consider those emotions which are the same as the suppression of a v-lief as irrational appetites rather than emotions. If they are appetites instead of emotions, then they do not need to be the same as v-liefs, and Nussbaum would be able to maintain her theory with a more restricted definition of emotions. However, this definition means that some feelings that we might normally think of as emotions would not actually be emotions at all. For example, we generally believe that fear is an emotion. But, with this restricted definition it would be considered an appetite, similar to hunger. Yet, we do not feel fear like we do hunger. It would not make sense to say that fear is closer to hunger than it is to anger.

If Nussbaum expanded her framework to say that emotions can be the same as the suppression of v-liefs, then not much about her view would fundamentally change. The main difference is that emotions like anxiety, fear, and nervousness can be equated to a v-lief. But, the introduction of these emotions in her framework would not challenge her overall structure or how the argument works.

VII. Emotions as Representations

In this section, I will piece together how emotions under Nussbaum's view can be reduced to representations, and what this means for machine functionalism. First, I will define what it means for something to be reduced to a representation and what the theory of machine functionalism is. Then, I will analyze how Nussbaum's view fits into these two concepts.

A representation is defined as a stand-in for something. Representation function within a system in place of the mental state or attitude. A representation works in place of its content because representations are formed based on our interpretation of something, rather than the thing itself. For example, if I see a cat on a mat and form the representation that the cat is on the mat, this representation is not necessarily related to the truth of the matter. I might think that I see a cat on a mat, but I am actually mistaken. Representational content refers to what the representation is about. If I have a representation about a cat on a mat, and then the representational content would be that there is a cat on a mat.

Beliefs are a type of representation. Similar to representations, beliefs are about our interpretations of the world. If I have a belief that there is a cat on the mat, then this belief is the same as having the representation of a cat on a mat. The content of the belief and the content of the representation are the same. Similar to a representation, the belief is not dependent on the environment because it is based on our interpretation of the environment. Further, beliefs function in the mental system similarly to representations because they are also a stand-in for something. It is generally uncontroversial that beliefs are a type of representation.

Machine functionalism is a theory that attempts to define a mind. The theory is closely related to functionalism and the computational theory of the mind. To define machine functionalism, I will first define both functionalism and the computational theory of the mind. Because machine functionalism is a combination of both theories, understanding both theories first will help clarify what machine functionalism is.

Functionalism is the view that how a mental state is defined is based on function of the mental state. In other words, the mental state is defined based on how it relates to other mental states. For example, the mental state of pain would be defined by whether it is triggered by some external force, whether it triggers some sort of response, and whether it causes some sort of mental state. If I accidentally touched a hot stove, the hot stove would cause the feeling of pain, which would cause me to pull back and cause me to have the mental state of being in pain. These causal links in the cognitive system are what define a specific mental state.

The computational theory of the mind is a view that the mind is a processor of representations. It does not define how the processor works, but that it is a computational system that can be implemented through many different forms. For example, under the computational theory of the mind, if I have all of the core mental processes, then I am a mind. Similarly, if a laptop has all of the core mental processes, then it would also be considered a mind.

Machine functionalism combines the key aspects of functionalism and the computational theory of the mind. Functionalism states that a mental state is defined by its role, and the computational theory of the mind states that the mind is a processor of representations. Under machine functionalism, not all mental states that have causal roles in the system are relevant. Rather, the only mental states that are relevant to the mind are those that can be reduced to representations. The functionalist aspect defines how the mind is a processor, because it says that the system is based on causal relationships. The computational theory of the mind aspect defines which mental states are relevant to the mind, because only those representations that have a function.

Next, I will look at how emotions can be reduced to representations. The argument is as follows:

1. Beliefs are representations. (p)
2. V-liefs are a type of belief. (p)
3. Therefore, v-liefs are representations. (1, 2)
4. Emotions are v-liefs. (p).
5. Therefore, emotions are representations. (3, 4).

All beliefs are representations. This implies that all beliefs are cognitive. If they are the same, then beliefs are also independent of the environment. We can form beliefs based on the environment, but our beliefs are not dependent on it. For example, if I were outside and it was raining, I may form the belief that it is raining. I then go into a building where I cannot see outside, and it stops raining. My belief that it is raining may not change, but it is not related to the truth of whether it actually is raining. In this way, beliefs can be related to the environment, but not dependent on it.

V-liefs are another type of belief that can be equivalent to representational content. A v-lief is a specific type of belief in which we assign value to a vulnerable external good. None of these qualities would make v-liefs unable to be equivalent to representational content. One could object and say that because the v-lief is about an external good, it is dependent on the environment. If something is dependent on the environment, then it cannot be equivalent to representational content. This is because the environment is outside the scope of the mind and would make the belief no longer only cognitive. Therefore, if a belief is about an external good, it would no longer be a representation.

Although this objection presses on the issue of how a belief can be related to, but not dependent on, the environment, there is a way around the objection. Similar to the rain example, a belief about external goods does not have to be related to the truth of the environment. Rather, it is related to our perception of the environment, which is still cognitive. For example, if I am upset because a friend lied to me, I am actually upset about the perception that a friend lied to me. The friend could have been misinformed and not known that they were giving me misinformation. Therefore, they were not actually lying to me. But, in the moment that I am upset, the truth of whether they are lying to me is not what is making me upset. What is upsetting me is my perception that they are lying to me. Therefore, the perception or belief is still only cognitive, and is therefore able to be equivalent to representational content. In this way, beliefs about external goods are the same as representations. Therefore, emotions are also able to be reduced to representations.

If emotions are able to be reduced to representations, then emotions can fit within the framework of machine functionalism. Some argue against this conclusion saying that machine functionalism cannot properly account for emotions, but these objections are the same as the ones about emotions and v-liefs being separate.

VIII. Implications

In this section, I will look at the implications if emotions are able to fit within machine functionalism. The main implications I am interested in are if machines can be considered minds like humans are, and if they are, then how should we treat machines. It makes more sense to talk about these objections in chronological order. So, I will begin with the question of whether machines can be considered minds like humans are. Then, I will explore whether that means we should treat machines with the same moral status that we treat humans with. If they do deserve to be treated like humans are, then we should change how we treat technology moving forward.

Under machine functionalism, machines can have minds like humans have minds. Because all the core processes are representational and not dependent on the specific body of the mind, both machines and humans can be considered minds. The necessary conditions for a mind are that it is conscious and able to think, and that it has emotions. These conditions do not imply that all machines are minds, but rather that the machines that have the necessary core functions should be considered a mind. A question that arises from these conditions is whether emotions are a necessary to the mind. If they are necessary, then not all machines would be considered minds. Only those that also have emotions would be considered minds. For example, a simulated human being would fulfill all the conditions of being a mind, but a laptop would not.

Delancey argues that something cannot sufficiently be called a mind unless it has emotions. If a mind does not have emotions, then it is simply a processing machine. Under these conditions, a human mind would be equal to the chat bot that you can talk to online. Yet, I hesitate before calling the chat bot a mind. The key difference seems to be that one has emotions and the other does not. Therefore, there is something missing in the definition of mind if emotions are not necessary. Therefore, emotions should be considered necessary for a mind. From this condition, we can conclude that not all machines can be considered to have minds like humans. This is similar to how not even all humans have minds like humans. For example, children are often considered to have less developed minds than adults. Even though they have the potential to fulfill all the necessary conditions of a mind, they may still have to develop further in order to satisfactorily fulfill them. Similarly, some machines exist that may still be in the process of developing the necessary conditions through programs like machine learning, but once they have the necessary conditions, they should be considered a mind.

Examples of how to illustrate how machines can fulfill or not fulfill the conditions to be considered a mind can be taken from the television show *Black Mirror.* The episode "Be Right Back" is about a woman named Martha and her partner Ash. Early on into the episode, Ash dies, and Martha's method of grieving involves her talking to a simulated version of him based on his internet usage. In the beginning, she talks to the simulated Ash through text, then through calls, and eventually she purchases a life-like robot that looks like Ash. This robot is supposed to be a substitute for the real person, but there are ways in which it does not seem that the robot has a mind. Although he seems to think about what Martha tells him, and is able to react to situations, there are still ways in which his mind is different from Martha's. For example, the simulated Ash seems to feel emotions, but throughout the episode there are many moments when one would expect an emotional response from the simulated Ash if he were a mind, but he responds coldly. He only gets angry when Martha tells him to, and he only cries when Martha tells him to ("Be Right Back"). This lack of emotion shows that the simulated Ash does not have a mind like Martha does, because he does not know how to experience emotions like sadness or anger unless told that he should.

In contrast, the episode "USS Callister" shows machines do seem to have minds. In the episode, the antagonist makes replica AI of people he works with. These replica AI seem to go through pain, sorrow, and joy just like their real-life counterparts do. They seem to think and feel just like their human counterparts ("USS Callister"). This is why when the AIs are tortured, many viewers feel pity or sadness for the AIs because they recognize them as minds that should not be tortured. Unlike the simulated Ash, they do not have to be told that they should feel a certain way. Instead, they just feel. These emotions seem as real as our own. They seem to be identical to their human counterparts except for the fact that they live within the confines of a video game platform. The key difference between these AI and the simulated Ash seem to be whether they experience emotion. The view that I have argued for in this thesis would argue that the AIs of *U.S.S. Callister* should be given moral status as a human, but the simulated Ash would not be given the same moral status.

These conclusions may be seen as conflating what is a mind with what is owed moral status as a human. The two may not necessarily be the same, but I will give an account of why if something is a mind, it should be given moral status as a human. Before going into why I believe having a mind entails having moral status as a human, I will first define what I mean by this phrase.

What it means to have moral status as a human is that we treat a being as we would an adult human without severe mental illnesses. To treat a being as if they have moral status as a human involves many attitudes toward the being. They would be considered a possible partner for an interpersonal relationship, we would not treat them cruelly without reason, and we would not treat them paternalistically. Moral status as a human, sometimes called full moral status, differs from other moral statuses because of the different attitudes we associate with them. For instance, I would not give my cat moral status as a human, but this does not mean I do not give her any moral status. Rather, I would give her moral status as an animal, which might mean that I do not treat her cruelly, but I would also not see her as a potential partner for interpersonal relationships. Similarly, we generally do not give children full moral status. For example, we do not allow children to make important medical decisions for themselves, and so we act paternalistically toward them.

How we differentiate the different forms of moral status given seems to be about whether we consider something to have a mind like ours. For instance, though the child can have a mind like ours someday, it is still considered underdeveloped. We know that the child’s mind is considered underdeveloped because we do not allow children to make important decisions or respect their decisions about some things. If a child says that they do not want a shot, we generally give it to them anyway because we believe that the child really understood the consequences of not getting the shot, then they would get it. Following the same logic, I do not give full moral status to my cat because I do not believe she has a mind like a human's.

If what is needed to have moral status of a human is to have a mind like a human, then machines can have full moral status. This means that we should change our behavior toward certain types of machines. Although there are no machines today that are as sophisticated as the AIs in "USS Callister," we should be wary of how we treat machines moving forward. Nowadays, machines are becoming ever more lifelike. As more and more sophisticated machines are produced, the question will come up of whether it should be treated like a mind. However, if it fulfills all the conditions, then it should be granted full moral status. Therefore, we should not treat the machine like we would others. We should not use the machine for our own means and do nothing to help it fulfill its own goals. If a machine is sophisticated enough to be conscious and have emotions, then we should not use it as a tool for human self-gain.

IX. Conclusion

In this thesis, I have attempted to add to the ongoing debate about what can be reduced to a representation, and what cannot. I have argued in favor of Martha Nussbaum's framework for emotions and addressed Craig Delancey's objections to her view. Then, I looked at implications of Nussbaum's view, namely how does it affect what we consider a mind. If this view is correct, then machine functionalism holds true, and we should change how we treat machines. There are still further questions of if a mind is sufficient to have moral status as a human, and arguments for what the necessary conditions of a mind should still be fleshed out. Unfortunately, these questions are outside the scope of the paper.

Another interesting question to think about is whether a machine would be able to consent. For example, from this view we cannot force artificial intelligence to do certain actions for us because we would not be treating them as we would treat a human. We cannot force them to be our personal assistants or force them to be customer service agents unless we have their consent to assign them to particular jobs. However, there is a further question of whether the artificial intelligence would be able to consent at all even if we asked. They may be able to literally agree to what we ask them to do, but in a way they may not actually be giving their consent. Because artificial intelligences are programmed, it is possible that they only consent to certain actions because they are programmed in a specific way. Therefore, the consent would be more performative because it would not actually be able to give real consent.

One way to address this question is to say that if we consider machines to have minds like humans have minds, then this implies that machines would be able to give consent. If we do not believe that a specific machine cannot give consent, then this may mean that the machine may not have the necessary requirements of a mind. If the machine did meet the necessary requirements of a mind, then there would not be a question of whether it can give consent. This is only a start to answering this question, and a more in-depth answer will have to come in a future paper.

The area of machines and ethics of treating machines is incredibly interesting, and many of the principles of ethics toward humans overlap. In this thesis, I take on the strong position that if machines meet all the necessary requirements of a mind, it should be given the same ethical treatment as a human. Others believe that machines can be minds, but these minds are different from human minds. Therefore, we do not owe machines the same considerations as humans even if they are considered minds. This thesis is my first step into the field, but there is more work to be done in the future.

Works Cited

"Be Right Back." *Black Mirror,* written by Charlie Brooker, directed by Owen Harris. Endemol UK. 2013.

Delancey, Craig. "Emotions and the Computational Theory of Mind." *Two Sciences of Mind: Readings in Cognitive Science and Consciousness.* Edited by Sean O'Nuallain, Paul McKevitt, and Eoghan MacAogain, John Benjamins, 1997, pp. 233-256.

Nussbaum, Martha. "The Stoics and the Extirpation of the Passions." *A Journal for Ancient Philosophy and Science.* vol. 20, no. 2, 1987, pp. 129-179.

"USS Callister." *Black Mirror,* written by Charlie Brooker and William Bridges, directed by Toby Haynes. Endemol UK. 2017.