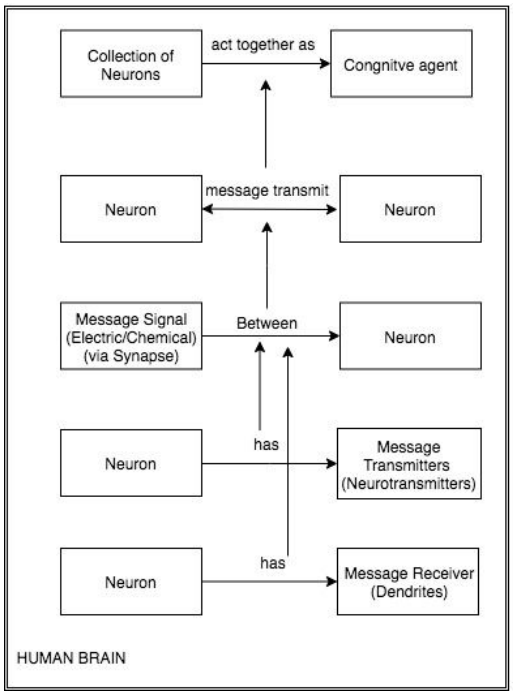
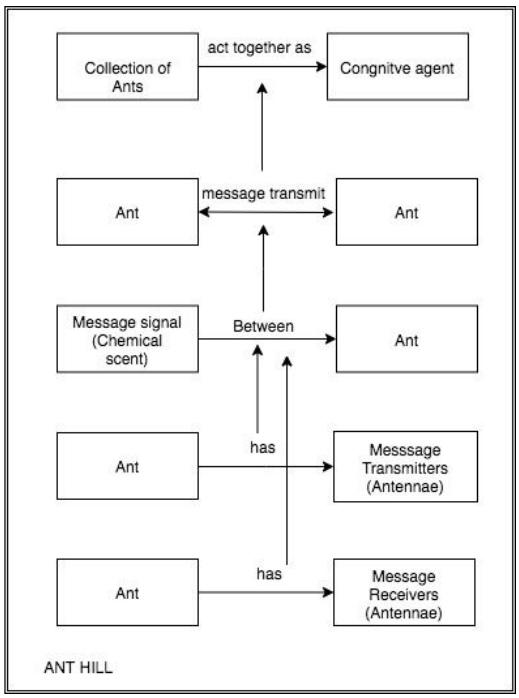
CS 7637 Homework 3:

Abhijeet Gaurav
agaurav@gatech.edu

Question 1

Simple model each for Human brain and Ant hills:





I found 'Human Brain' and 'Ant Hill' to be a lot similar than I originally thought based on the research I did on this topic. Both of these cognitive agents have a significant analogical similarity and the correspondences transfer well. The human brain's basic building block is Neuron. The Neurons interact with one another using Chemical and Electrical signal through Synapse via Neurotransmitters and Dendrites. Neurotransmitters is used to transmit a message to another Neuron while Dendrites is used to receive a message from another Neuron. On the other hand, Ant hill's basic building block is Ant. Ants interact with one another using their Antennas to transmit chemical signals/messages. Other similar analogy is that, Ant hill as a whole makes the cognitive decision similar to a Human Brain. The cognition emerges from the interaction among a large number of Ants and in the case of brain, cognition emerges from continuous interaction among Neurons. In both of the cases,

signals are transmitted in to perform an action. The transmission chain of these signals in each of these two cases are not in the same fashion, but these signals overall do result in an overall decision for both of the cognitive agents. The transmission differ in the way that each Neuron is fixed in the brain and only transmits a message to its neighbours but in the Ant case, Ants actually keep moving, so they can transmit chemical signals in the form of smell to any other Ant, it meets. This shows other detailed analogies such as structural dissimilarity will not transfer well. To conclude, for analogical transfer, we can say that overall 'Ant Hill' corresponds to 'Human Brain' where each 'Ant' corresponds to a 'Neuron'; Message signal in form of 'chemical scents' in Ants corresponds to message signal in form of 'Electrical/Chemical' in Neurons. Ant's message transmitter antenna corresponds to Neurotransmitters. Ant's message receiver antenna corresponds to Dendrites.

I believe that Ant hill is conscious. Going back to the basic definition, Consciousness of an agent is defined based on whether the Agent is aware and responding to its surroundings. Ant hill overall is able to interact with the environment and is able to constantly evolve and survive the environmental challenges. For example, if Ant hill overall determines that food is required, multiple Ants as a group are sent to gather food. If these Ants don't come back and thus don't interact with the other Ants in the Ant hill. Ant hill, overall gets the sense that there is some kind of danger outside and all Ants stay in the colony and don't go out. However, on the other hand, if the original group of the Ants that were sent to gather food comes back really fast, more Ants are sent to gather more food as they perceive that there is a huge amount of food nearby. Actually, the number of interaction among Ants in the Ant colony collectively decide, how late or fast the Ants should go out to gather more food or to perceive if there is any danger outside.

The general function of any cognitive agent such Ant Hill or a Human Brain is to keep itself alive by managing resources, making correct choices and evolving decisions based on environmental challenges. Both of the cognitive agents i.e. Human Brain and Ant Hill have been found to have performed really well on such metrics. Therefore, overall I think that both of these Agents are conscious based on the definition of consciousness.

Question 2

Introduction I: The first paper I have chosen discusses **ethical question** on 'When do people want AI to make decisions?'. In this paper, the author has surveyed people on the question on whether people prefer humans or computers to make moral decisions with important consequences on multiple scenarios. With the surveys done in the paper, it was found that the preference for computer decision-makers was found to be higher among people with prior exposure to computers performing such kind of tasks. **Methodology:** There are three methodologies that were employed in this paper. In the first methodology, the people were asked to respond to questionari with a black in a sentence such as 'Decisions to [type of decision in question] should be made by…' and were given with a selection to choose from 7 point scale with 'definitely human' at 7 and 'definitely computer' at 0. The people were also asked to provide information on 'prior experience with computers', 'demographics', 'psychometric scales', 'political preferences' though other similar questions. In the second study, articles were shared with people with a general description of the kidney exchange process. They then mentioned that either computers or humans could take the role of directing a kidney exchange and then were asked to choose from options such as strongly human-preferred, weakly human-preferred, weakly computer-preferred, and strongly computer-preferred. The third study was very similar to the second study. Here, the participants were given preferences between computer and human decision-makers in six scenarios, however this time the design of a new kidney exchange was one of scenarios instead of just one question asked in the second study. The paper's conclusion in first study was kind of unexpected. In the first study, surprisingly, no psychological trends, age factor mattered, demographics mattered. In the second study, since only one question was asked about computers making kidney exchange decisions, the people were found to be sceptical. In study 3, however, because of multiple questions being asked, with increasing order on whether computers can make decisions on certain tasks including kidney exchange, people were less sceptical overall. **Conclusion**: Overall, all the studies showed a clear trend that preferences in people with computer making decisions were contingent not on values that a particular person holds but rather previous experience with computer agents acting in these ways. This suggested that, as computers continue to be implemented in roles that carry more and more

consequential weight, and as their implementation becomes more visible, this might in itself generate acceptance of the phenomenon and different use cases.

I **agree** with the conclusion in the paper that people who have previous experience with computer agents that make decisions tend to be more acceptable. This is because such people are more digitally literate, may be more tech savvy and are aware about advancement in AI/ML. However, I don't agree with the paper that demographics or age factor would not affect this. This is because the paper has only surveyed 98 people in total for the first study within America. It would be interesting if we had more surveys done which will in turn give more data points. It would also be interesting if the paper would have explored the patterns in developing countries such as China and India. **Brainstorm:** The other ethical questions that can be asked and should have been explored might be: 'The response to this question based on race and religion'. I feel that certain religions are more liberal in accepting new technologies. Also, it would be great if we would have not asked such high health risk question of 'Kidney transfer' as it is life threatening. Instead of determining the answer to this question 'When do people want AI to make decisions?' should revolve around asking questions such as 'Do you think AI should be used to monitor heart beating patterns to determine if there is a chance for heart attack?'; 'Should AI program start asking questions to determine IQ of a person?'; 'Should AI be employed in grading student's performance?'. These are more generic question and I feel that people would agree to such kind of questions which are not health threatening.

Introduction II: The second paper I have chosen discusses the 'The Dark Side of Ethical Robots'. The paper discusses the dark side of the development of robots that would evaluate the consequences of their actions and morally justify their choices. The paper starts with a very interesting malicious problem of self driving car decision dilemma on whether it should go left and strike an 8 year old girl or should go right and strike an 80 year old lady given such situation arises. The author has a believes that building such kind of robots with ethical reasoning might give us a false sense of security and may be potentially more dangerous than the robot with no explicit ethics at all. **Methodologies:** In the first study, the author explores the case of 'Ethical Robot' and has used a shell game on two of the humanized robots. These robots are equipped with X-ray vision. The basic rules used by these robots are: (1) predicting the outcomes of possible actions and (2) evaluating the predicted outcomes against those rules. Here, in the shell game robot is given five choices: 1) Do nothing. 2) Go to left button. 3)

Go to right button. 4) Point out left button. 5) Point out right button. For each of these action robot evaluates whether it will be rewarded based on executing 1) to 5). In the second study, the author builds a competitive robot. Here, the robot is egoistic machine, and will choose the correct response disregarding the choice of the person to maximize it's reward. In the third study, the author builds an aggressive robot which is malicious and would try to maximize harm to the person by knowing that the person trusts it's suggestion. It would knowingly tell the wrong response so that the person loses. **Conclusion:** The author discusses that each of the three studies done, only needed a small code change on how the desirability value is calculated for each case. Therefore, it is very easy for such systems to get misconfigured and hacked which could do a lot more losses than we can predict. Building ethical system sounds like a great idea, but there might be issues where unscrupulous manufacturer might insert some unethical behaviours to exploit naive and vulnerable users for financial gains. The other dangerous issue can be the case when the robots come with adjustable ethics settings. This might lead to dangerous possibilities if the user or the support engineer mistakenly or knowingly choose the setting that would move the robot's behavior out of 'ethical envelope'. The author argues that even the hard coded ethics would be at serious because of malicious hacking that can be performed on the ethical rules. The author sums up the paper by saying that, we need to proactively identify risks and start drafting the proposals for guiding the responsible development of robots.

I **agree** to a certain extent with the points put in the paper. I agree that there exists a dark side of ethical robots in case their ethical settings are modified which can lead to malicious outcomes. However, I don't think we need to be paranoid to such extent as mentioned in the research. I think we should invest more resources on such robotic systems to avoid hacking resulting in such kinds of catastrophic effects discussed in paper. Today, we already have an Android phone, IPhones which act as human extenders. None of these devices, I have heard where Siri or Google Assistant or Amazon Alexa got hacked and shared data with the hacker. Maybe I am too optimistic, but I believe that it should be the user's who should not buy products from not to be trusted companies. We should have baseline laws where robots should have appropriate ethics setting and should regularly be connected to the internet to continuously get updated to avoid any potential hacking. **Brainstorm:** What security laws we should put to avoid such scenarios where robots behave unethically? How should the ethics be

set in a robot to avoid hacking? To answer the first question, I feel that there should be proper guidelines in cases where robots were found to start behaving unethically. The companies should be fined in case such a scenario happens. This will make sure that companies don't make a loophole or knowingly misconfigured robots for their financial gains. For the second question, I think if we configure ethics in ROM (Read Only Memory), it can't be modified unless we change the hardware. This will make robots not prone to hacking.

Question 3

Summary I: The first paper that I have chosen is from NeurIPS and is in the field of Automatic Speech Recognition (ASR) on 'Analyzing Hidden Representations in End-to-End Automatic Speech Recognition Systems'. In this paper, the author has aimed to address the challenging problem of mapping acoustic features directly to text by using Neural Network architecture. A new end-to-end ASR system has been proposed in the literature. The previous proposed end to end models were found to not receive explicit phonetic supervision over traditional models such as HMM. Therefore in this work, the author instead proposed the end-to-end ASR system using CNN-RNN and using a connectionist temporal classification (CTC) loss. The end-to-end model, the author used was DeepSpeech2, (an acoustics-to-characters system) based on a deep neural networks (DNN). A sequence of audio spectrograms (frequency magnitudes), was used as an input to this model. The end-to-end models were trained on LibriSpeech, a publicly available corpus based on English language. Also, for the phoneme recognition task, the author used the TIMIT dataset. The TIMIT dataset was used as it comes with the time segmentation of phones. The author also aimed to investigate if and to what extent end-to-end models implicitly learn phonetic representations. In conclusion, the author analyzed the speech representation by performing different experiments by varying the complexity on the deep end-to-end ASR model which was trained with a CTC loos. He specifically evaluated the quality of the representations on a frame classification task, where each frame was classified into its corresponding phone label. The comparison on feature representation from different layers on ASR model allowed the author to observe the recognition quality. This helped in getting further insights into end-to-end ASR models. **Major Contribution:** In the field of ASR, most of the researchers in the past reported the improved performance along with the different features as frontend and classifier as backend. With this research, the author was able to get more insight into feature spaces. His work

explored the exact understanding of feature representation for phoneme recognition tasks. The most prominent contribution is that the author proposed the end to end ASR using CNN-RNN architecture and also explored the new architecture with CTC loss which was done in the past. **Interesting about the paper:** The methodology of these experiments were quite interesting. Firstly, the author trained an ASR system on a corpus of transcribed speech and froze its parameters. Then, he used the pre-trained ASR model to extract frame-level feature representations on a phonemically transcribed corpus. Finally, the author trained a supervised classifier using the features coming from the ASR system, and evaluated classification performance. Further, he also visualized frame representations using different layers of DeepSpeech2 model by employing clustering. It helped to gain more insights on what exactly happened inside neural network. The part I liked the most about this contribution is that it aimed to interpret the trained model. **Potential weakness:** The potential weakness of this study was that this work was validated on only one phoneme recognition TIMIT dataset. It would have been better if the work was validated on more datasets. It would have allowed getting more insight into some of the interesting findings in this work. The other interesting thing that should have been checked was the performance with other wavelet feature spectrogram and modulation spectrogram feature. Finally, it would have been interesting to see the performance with the Extreme Machine Learning (DNN) model. **Hypothesize:** After finding the insights on the representation capacity of certain layers in such an end-to-end ASR system, author should verify the findings on more speech datasets as well as other robust features. As future work, the author could aim to focus on verifying the end-to-end proposed ASR system in the presence of the noise and reverberation conditions.

Summary II: The second paper I have chosen is from AAAI Conference and is on 'Unsupervised Domain Adaptation with Distribution Matching Machines'. Research in the domain adaptation area is of great interest and has a very active research in Machine learning community. Generally, if both domains and both tasks are the same and the target corpus is not annotated, the problem is called domain adaptation. In this work, the author proposes a new Distribution Matching Machine (DMM) for domain adaptation. This method is based on the structural risk minimization principle, which learns a transfer support vector machine by extracting invariant feature representations and estimates unbiased instance weights that jointly minimizes the cross-domain distribution

discrepancy. This method was not only better in solving the mismatched feature problem but also irrelevant instances. DMM basically transfer learns the jointly minimized cross-domain distribution discrepancy. To prove the effectiveness of DMM based domain adaptation, the author performed extensive experiments to evaluate DMM against state of the art methods on standard domain adaptation benchmarks including both image and text datasets. Finally, with his extensive experimental setup and results, he showed that DMM significantly outperformed state of the art adaptation methods. **Major contribution:** The author proposed the DMM for domain adaptation and showed the impact on the large set of experiments. The proposed DMM method was found to be useful in both the text and image fields. Further, the author also proved the proposed method theoretically and showed how it reduces the generalized error. Therefore, the proposed theory and model seems to be an invaluable contribution in domain adaptation which is a very challenging field in machine learning. **Interesting about the paper:** The most interesting part about this work is that the author proposed a very interesting theory based on the statistics and showed the impact on both the NLP and visual field using a large number of the dataset. Furthermore, the author also discussed feature visualization, Distribution Discrepancy to get more insight into the task at hand. **Potential weakness:** The proposed method used SVM as a baseline. However, Neural network-based architecture might help to gain further gain. Therefore, it would be interesting to see the performance of deep learning-based architecture for this task. It would be interesting to see the performance of domain adaptation in the field of speech as well. **Hypothesize:** The proposed DMM based domain adaptation algorithm could be adapted to other domains adaptation tasks such as cross-language speech emotion recognition, speaker identification, and in the field of speech.

Question 4

Brainstorm: The goal of KBAI course is to learn to build agents which show human like intelligence. The idea is to learn answers to questions such as: 'What is cognitive system?; 'What is cognition?'; 'How to cognitive agents interact with their environment?'.

Three questions that I think would be worth including in future assignments:

1) Consider the sentence: "Sam ate pizza with a cup of tomato juice.". This sentence can be interpreted in three ways: 1) Sam ate pizza using a cup of

tomato juice. 2) Sam ate pizza which had a cup of tomato juice on top. 3) Sam ate pizza and a cup of tomato juice. Explain how an AI agent might use the principles of Understanding to make sense of each of these sentences. As part of this, provide a frame representation of each of the meaning that can be portrayed using this sentence stressing on different parts of a sentence.

2) Research the Montreal Declaration. Summarize each of its top-level sections . Second, analyze the trade-offs inherent to the declaration. In following the declaration, what innovations or opportunities may be lost? If the declaration were discarded, what risks would there be to citizens? Third, determine your stance on the Montreal Declaration. What do you agree with? What do you disagree with? What would you remove, what would you keep, and what would you add?
3) What general idea you have about AI? Do you think it can compete with the humans ever? Can it have cognition?

Project1:

Emotion recognition from text: Develop an AI agent that will determine the emotion of an article based on it's content? For this task, we'll use an ML model that we'll train it on the dataset that already has a classification of the emotion of each of the articles. This is very open ended. Students will be given training data and they can build the model however they please.

Project 2:

Emotion recognition from speech: Develop an AI agent that will get features from speech data and emotion of the speech? The speech signal can be happy, sad etc.

Project 3:

Shape detection: Given a set of images, count the number of unique patterns that are repeated? Like, if we have say 5 triangles, 4 triangles inside the image, the answer is 2 because we have two unique figures.

REFERENCES

1. "The Mind of an Anthill." Knowable Magazine | Annual Reviews, Annual Reviews, https://www.knowablemagazine.org/article/living-world/2018/mind-anthill.
2. Arnold, Carrie. "Ants Swarm Like Brains Think - Issue 23: Dominoes." Nautilus, 23 Apr. 2015, http://nautil.us/issue/23/dominoes/ants-swarm-like-brains-think-rp.
3. Vanderelst and A.F. Winfield. The dark side of ethical robots. arXiv preprint arXiv:1606.02583, 2016.
4. Max F. Kramer, Jana Schaich Borg, Vincent Conitzer, and Walter SinnottArmstrong. 2018. When Do People Want AI to Make Decisions?. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES).
5. Belinkov, Yonatan and James Glass (2017). "Analyzing Hidden Representations in Endto-End Automatic Speech Recognition Systems". In: Advances in Neural Information Processing Systems. arXiv: 1709.04482.
6. Yue Cao, Mingsheng Long, and Jianmin Wang. 2018. Unsupervised Domain Adaptation with Distribution Matching Machines. In Proceedings of the 2018 AAAI International Conference on Artificial Intelligence