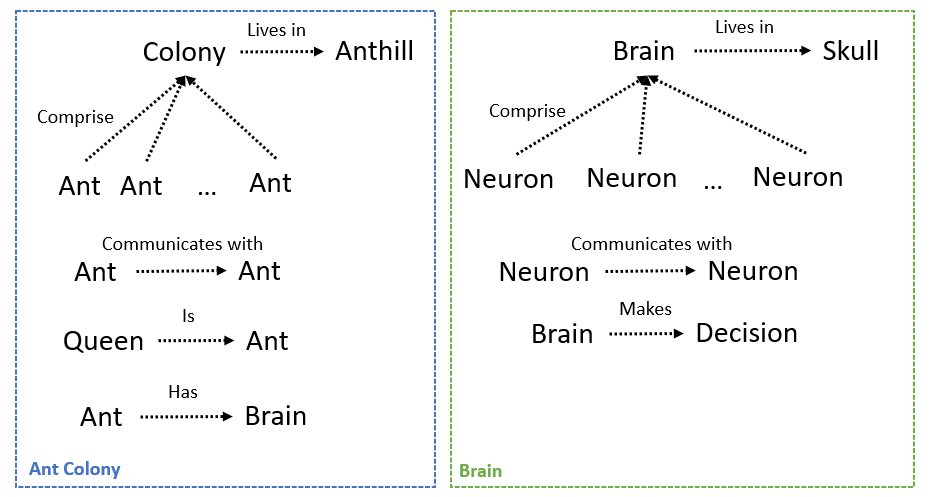
CS 7637: Homework 3

Zachary Todd  
ztodd3@gatech.edu

***Abstract—***This paper contains responses to the questions posed for homework 3 in CS 7637.

# 1 Making brains out of ant hills

Ant colonies and human brains are surprisingly similar. Figure 1 below shows simple models for each. In the ant colony model, note how many ants comprise an ant colony. Similarly, a human brain is comprised of many neurons. This structural similarity is a very weak connector between ant colonies and a human brain.



***Figure 1—*** Side by side comparison of relationships within an ant colony (left) and a neural network (right).

On the other hand, the method by which information is spread in these two systems is what makes the human brain comparable to an ant colony. In the brain, neurons send electrical impulses through a nerve by a series of action potentials (Holmes, 2018). These electrical impulses represent some information being sent through a neural network. This information passes from neuron to neuron until the information has reached its destination. Similarly, in an ant colony, ants will communicate to one another to pass information through the anthill’s tunnels until the information has reached its destination. These methods of communication are analogous to each other and prove a deep similarity between ant colonies and brains. Thus, we can draw an analogy between ant hills and human brains.

It is worth pointing out that each individual ant has its own brain and is capable of making unique decisions independent of the rest of the ant colony. They may be heavily influenced by the queen or other ants, but they still make their own decisions. Neurons, however, only react biologically to chemical and electrical experiences. They do not make conscious decisions. In this way, the comparison between these two models breaks down.

# 2 A question of ethics

In this section I will discuss two articles from the 2018 Artificial Intelligence, Ethics, and Society (AIES) Conference: “Preferences and Ethical Principles in Decision Making” and “The Dark Side of Ethical Robots.”

## 2.1 Decision making ethics

The main point of the first paper is that as AI systems become more and more complex, ethical principles need to be integrated into their decision-making process (Loreggia, 2018). An effective way to make this happen is through the use of CP-nets. A CP-net is a “graphical model for compactly representing conditional and qualitative preference relations” (Loreggia, 2018). They consist of a set of preference features, F, and parent features, FP. The parent features represent different states that the agent may encounter. This could be something like sitting down to a steak dinner and having the option of selecting red or white wine. In this case, the preference set of features, F, would consist of {red wine, white wine}, each with its own weighting. A CP-net would then show a relationship between F and FP. Specifically, it would show which option the agent will choose based on its own preferences.

Loreggia and her colleagues create a model of human behavior by creating 2 CP-nets: one for an agent’s subjective preferences and one for objective ethical principles such as safety regulations or feasibility constraints. They then present different scenarios to the agent and calculate the distance between the agent’s preferences and its provided ethical standards. This results in the agent making decisions that it prefers and are *close enough* to an ethical decision (Loreggia, 2018).

The group concludes that CP-nets are “accurate in practice and efficient to compute” when applied to a decision-making model (Loreggia, 2018). While I haven’t implemented this design myself, it seems to hold true. In their results, the group shows how this model works for situations where an agents preferences are in line with ethical principles as well as when they are not close to ethical principles and the agent must compromise. This seems to model human behavior rather well.

In this paper, Loreggia and her colleagues assume that an agent will not encounter a situation where its preferences and the ethical principles are in direct conflict. For instance, what happens if a person were to encounter upon a “do not walk on the grass” sign, but there was also a briefcase on the grass that said “free $1M?” The agent’s preference would be to take the money, but its ethical considerations would not allow it to step on the grass. Perhaps this model of decision-making could weigh consequences against rewards when personal preferences are in direct opposition of ethical standards.

## 2.2 Dark (Sith?) side of robots

This paper focuses on how the production of robots that rely on ethics to make decisions inherently makes an easy way for competitive, or even aggressive, robots to be made. Vanderelst and Winfield argue that the negative impact of these corrupt robots outweighs the benefits of moral ones and the production of ethical robots should not be pursued (Vanderelst and Winfield, 2018). They construct an experiment in which a human is playing a shell game. This is where some reward is hidden under a cup and the human must guess which cup the reward is under. The robot is equipped with x-ray vision and can tell where the reward is. The robot has the ability to inform the human of which cup the reward is under.

Initially, the robot is programmed to stop the player when they are about to select a cup without the reward. This is due to the robot having a decision system that values the human getting a reward. However, Vanderelst and Winfield point out that by changing the robot’s programming to value the robot itself receive a reward, the behavior of the robot changes. It now will choose the reward cup itself and become an “egoistic machine,” the *competitive* robot (Vanderelst and Winfield, 2018).

In an even worse scenario, they put the robot’s motivation on the human losing. The robot then stops the player when it is about to go towards the reward, directing the player towards the losing cup. This is the *aggressive* robot.

The paper concludes that corrupting ethical robots into unethical ones is a relatively simple process and that “guaranteeing the security of these robots… will need regulatory and legislative efforts” (Vanderelst and Winfield, 2018). The robots themselves are just tools; it is up to the users to decide how to use them. However, even with these governances in place, the risks that the robots may be compromised is too great to allow for the embedding of ethical decision making in robots.

I agree in part with the conclusion of this paper. There is an inherent risk to implementing a decision-making process that relies on ethics. However, I would argue that any software has the potential to be modified for another actor’s personal gain. For instance, flight navigation software could be tampered with to cause pilots to crash, Google Maps’s algorithms could be altered to draw drivers to certain streets to boost the economy of a small town, or a backdoor could be implemented in a security system to allow nefarious actors access. None of these depend on ethical standards to abide by. It is true that altering a system’s ethical priorities would be rather easy from the perspective of how much code would need to be changed, but if a person really wanted to change the way a system behaves, whether the system is based on ethics or not will not deter the saboteur.

As a follow up, I would be interested to see what these authors would have to say about an agent that learns over time and changes its methods of behavior. Right now, the agent just maps inputs to outputs without any update of decision-making. I believe an agent that could update its decision-making process might be at risk to misconstrue ethics it is based on. Just watch the movie I-Robot.

# 3 A tale of two papers

In this section I will discuss two articles from the 2017 Artificial Intelligence and Interactive Digital Entertainment Conference: “Generative Design for Textiles: Opportunities and Challenges for Entertainment AI” and “Level Difficulty and Player Skill Prediction in Human Computation Games.”

## 3.1 Cross domain cross-stitch

The first paper summarizes the process by which textile pattern generation is done for cross stitch and quilt samplers. Mainly, the article praises the process used by Hoopla and Foundry. These two generators create patterns using a “crowdsourced, data-driven approach for palettes and quotes, and an ad hoc constructive method for motifs” (Smith, 2017). The systematic approach of pattern generation creates a creative product for users to enjoy.

The interesting thing about this article is that it highlights the opportunity for creative products using the structured tool of software. I have always thought that coding has a bad reputation for being very black and white. It is very technical and regimented, true, but it offers so many possibilities as a creative outlet! Companies like Pixar and Blizzard create fantastic worlds of fantasy through their software and this article reinforces the idea that the potential for creativity through software is huge! It opens the door to an art form that combines textbook mathematics and logic with the imagination of the programmer.

Smith could have spent more time talking about how the crowdsourced data is used to generate new patterns. Perhaps the program uses a collection of user-submitted patterns to act as a guideline for future patterns? I believe that these generators would be more accepted by end users if they adapted to the tastes of the consumers. Just like clothing and music, I am sure that tastes in embroidery patterns change over time, and the article fails to mention how this is addressed.

I would love to see this research be applied to other applications of software in creative environments, particularly in the field of artificial music generation as well as storytelling. This article briefly talks about how textiles have been used throughout history to tell stories (Smith, 2017). Hoopla and Foundry do not seem to capture this usage well, but perhaps there are programs that have already been developed to generate stories that could be incorporated to them.

## 3.2 Incremental puzzle games

This article talks about Human Computation Games (HCGs) and how they have a low retention rate due to the different difficulty expectations by the consumer. Most games have a difficulty system which gets harder as the game goes on but is not catered to each individual player. Some games allow the player to change the difficulty level to their own preferences. This paper suggests that a game adapt to each player by presenting levels to the player that are consistent with the player’s rating. This rating will be updated depending on how the player does on the level. This sort of ranking system is used in many player-versus-player games but is not as widely applied to HCGs.

Sarkar used Paradox, a 2-D puzzle game, as a case study for this ranking system idea. To begin the game, a player must first go through 9 tutorial puzzles (Sarkar and Cooper, 2017). Sarkar’s ranking system would take a player’s performance on these tutorials and assign them a ranking based on total time taken, total number of moves, and the amount that the player’s score was over par for the level. Sarkar designed another ranking system for the game’s levels and would then present the player with a puzzle that was close to the player’s ranking. At the end of a level, the player’s ranking was updated based on their performance in the level. A different method of using default ratings for all players after the tutorial, and then updating slowly based on performance was also used. These two methods were then compared at the end of the paper. The conclusion of the research was that using the tutorial to establish a player’s rating was more effective at reaching a long-term stable ranking than giving players a default rating.

I thought this article was interesting because all the puzzle games I have played were constructed in the traditional way of sequentially harder puzzles. After reading this paper, I agree that this methodology can frustrate some players because the difficulty progression is often too slow or too fast. Creating a system by which puzzles match a player’s performance level will surely help with player retention. However, the article fails to actually address if this method works. It only compares two different ways of implementing this method without showing its impact on player retention. I think this research should be implemented and given to consumers to see if it works.

# 4 Free to be me

Individual humans have free will. At least from the perspective of the individual. I do not know the future, but I am able to influence what happens in the future by assessing the world as I see it and making an action in the world. My decisions are of course decided by all my past experiences, perception of the world, and personal beliefs.

I will define free will as the ability to make decisions and take meaningful action based on the individual’s own wants and desires. Furthermore, free will can only exist if the individual is able to change its wants and desires over time. A meaningful action is one which actually has impact in the world and is not the only option available. For instance, someone with sleep paralysis only has the option to lie completely still. They cannot take meaningful action in the world.

An artificially intelligent agent can have free will. This may only happen if the agent is given the ability to perceive the world around it, is given a range of meaningful actions which it may take in the world, can predict what the outcome of an action may be, and has an adjustable set of goals/wants. Without each of these elements, the agent does not have free will. If it cannot assess the world it is in, it does not have a way to make decisions. If the agent cannot act in the world, or only one of its actions will do something meaningful, it is constrained to what it can do. If it is unable to predict what the result of an action may be, it cannot decide what action to take. If it does not have a set of wants to compare these predicted outcomes to, it has no way of determining which decision is the best one. The agent does not need an explicit history of its experiences. This is because an adjustable set of wants and desires can be adjusted based on each experience as it happens. This will then allow the agent to make the same or a better decision in the future.

I believe that humans do have free will, but will always choose act according to the way they were designed. We are born with certain tendencies that manifest themselves in the chemical/electrical reactions in our biology. Whether spiritually or naturally given, these tendencies change as we live in the world based on our experiences. Similarly, our wants and desires change as we age. In much the same way, AI agents start with a set of default tendencies to include wants and desires. These are determined by whoever programs the agent. Over time, the free-willed agent will update its wants and desires based on how it experiences the world, very similarly to how a human would.

However, AI agents should never be considered semantically equivalent to humans. To do so would degrade what it means to be human. Humans have souls and a moral compass which cannot be replicated in an AI agent. In terms of free-will, AI agents are susceptible to reprogramming or alteration. Furthermore, their actions are a direct result of their programming, which is the fault of their creator. As humans have created this agent, they are on a different level of cognition and should not be considered equivalent.

An interesting point to consider is how humans are also susceptible to a sort of reprogramming in the form of persuasion. Propaganda, threats, and even family values can all be seen as a form of “reprogramming” in that an outside force is changing a person’s values, wants, and desires. This is similar to an AI agent being reprogrammed, but once again, a human is ultimately a result of its initial biology experiencing the world. A reprogrammed AI agent would be like a human being reborn.

# 5 REFERENCES

1. Holmes, Bob. “The Mind of an Anthill.” Knowable Magazine | Annual Reviews, Annual Reviews, 2018, [www.knowablemagazine.org/article/living-world/2018/mind-anthill](http://www.knowablemagazine.org/article/living-world/2018/mind-anthill).
2. Loreggia, Andrea, et al. “Preferences and Ethical Principles in Decision Making.” AIES Conference, 2018.
3. Sarkar, A., & Cooper, S. (2017, September). Level difficulty and player skill prediction in human computation games. In Thirteenth Artificial Intelligence and Interactive Digital Entertainment Conference.
4. Smith, G. (2017, September). Generative Design for Textiles: Opportunities and Challenges for Entertainment AI. In Thirteenth Artificial Intelligence and Interactive Digital Entertainment Conference.
5. Vanderelst, Dieter, and Alan Winfield. “The Dark Side of Ethical Robots.” AIES Conference, 2018.