Joseph Blass
Final Project Report
Multi-Agent Modeling Spring 2015

Modeling Changes in a Population's Exploitative vs. Protective Behavior

Introduction: Humans regularly exploit other people for selfish reasons. We also defend other people, including complete strangers we may never meet, against such exploitations (as well as other harms). Being too protective can be bad, however, since in the absence of a threat, people prefer not to be watched too closely by their neighbors. What circumstances within a society incentivize its members to exploit each other, protect each other, or leave well enough alone? Are there circumstances that select for a particular trait (exploitative, protective, or placid)? Where is the balance between watching out for your neighbors and being nosy? Furthermore, can we define simple circumstances that lead to a balance between these behaviors, as we find in human society, or will certain traits always be selected for over time? I built an agent-based model of social dynamics in NetLogo to explore these questions. I also built a HubNet model with slightly different behavior, so that humans can collaboratively explore this experimental space.

I present a simplified social model of moral behavior. The model features agents, or turtles, who live in a society, represented by the entire model space. Turtles have social currency, which represents wealth, social status, and reputation. As they move through their society, at each tick, turtles decide what to action to take. They decide probabilistically (as detailed below) amongst stealing, protecting, or doing nothing. Turtles who steal have a chance of getting away undetected; if they are not caught by a turtle who is protecting, they siphon social currency from their neighbors. If these thieving turtles are caught stealing, they give social currency to their neighbors who chose to protect, that is, the turtles who caught them in the act. Turtles who protect against a thieving neighbor take social currency from that neighbor. If a turtle who protects fails to catch a turtle trying to steal, it is punished for being a nosy busybody and gives up social currency to all of its neighbors. Turtles who do nothing can be stolen from, or can be rewarded if they have nearby protectors who do not catch any thieves. Otherwise, nothing happens to them (as we will see, this may have been an error in design).

If turtles gain social currency above a particular threshold, set by the user, they can reproduce with another turtle above that threshold (at most once per tick for both turtles). They have one child, whose internal theft and protection parameters have the average of their parents' parameters, plus some randomness. Therefore if the parents were being rewarded for being likely to perform a particular behavior, then the children are more likely to display that behavior. Turtles also die after they reach their life expectancy, or if their social currency dips to zero. Thus we are able to model the change in population over time by observing the change in behavior of turtles in the aggregate.

Detailed Model Description: Users have direct control of five parameters in the model: theft threshold (or the initial overall likelihood of turtles to steal), protection threshold, initial social currency, mate threshold, and probability of getting away (i.e., not getting caught) when stealing. Initial social currency is the same for all turtles at setup and for new turtles born during the simulation; mate threshold and probability of getting away are global parameters that are the same for all turtles and do not change over time. Each turtle has their own theft threshold, protection threshold, and social

currency. Theft and protection thresholds are set for initial turtles by taking the global parameters and choosing a number within 5 of that parameter (that is, by taking the global threshold, subtracting 5, and adding a random number between 0 and 10, inclusive). Turtles born during the simulation have the average of their parents' parameters, again with up to 5 degrees of randomness.

The theft threshold and protection threshold are set such that 0 <= theft threshold < protection threshold <= 100[1]. At each tick, turtles select an action to perform as follows. First they pick a random number X between 0 and 100. If X < theft-threshold (for that particular turtle), they steal. If they steal, they also determine whether they are sneaky enough to get away with their theft. They do so by picking a number between 0 and 100; if the number is below the probability of getting away, they are sufficiently sneaky. If the action selection variable X < protect-threshold, turtles do nothing. Otherwise, protect-threshold <= X <= 100, and they protect. All turtles select an action before performing the action; actions are resolved as previously described.

There is a constant unit of social gain, set within the model to 7. Thieves who successfully steal get that amount, and all the turtles they steal from lose the fraction of that amount corresponding to the number of turtles stolen from. Turtles who successfully protect get that amount from each stealing turtle they catch, divided amongst themselves and the other protectors who caught that same turtle. Turtles who protect and fail to catch a thief lose that same amount, divided amongst all their neighbors. Finally, turtles who have social currency above the global mating threshold reproduce, and turtles who are past their life expectancy or who have no social currency left die (death in this model being is a stand-in for both death and exile from society). Turtles are initialized to be black and turn slightly greener each time they do nothing, bluer each time they protect, and redder each time they steal, providing a means to visualize the population's behavior (for example, a white turtle is one who has performed all three actions more than ten times).

The HubNet model is a simplified version of this model. There is no reproduction or death. Instead, any turtles caught stealing the previous turn move; then all other turtles move; then all turtles select what action to perform, which are resolved as above. In the HubNet model, turtles caught stealing the previous round turn red; all turtles who protected turn blue; and all others turn green.

The model changed in two major ways over the course of development. First, turtles originally had no chance of getting away if they stole. This was changed because otherwise exploitation was always selected against, since with large enough populations exploiters would always be caught due to the simple mechanics of population density. Also, in the original model, turtles sought the highest social valued turtle in their neighborhood. This led to clusters of turtles, which was a goal of the original model. Each cluster behaved like a society, but did not display different behavior from the overall averages. Furthermore, the whole model was supposed to be of a single society. I experimented with having them move from cluster to cluster, but at that point the model ran very slowly and was functionally the same as simply having them move around without clustering. Therefore I removed the element of clustering.

[1] If the model tries to set a turtle's theft threshold >= protection threshold, for example because of the average thresholds of its parents, then the theft threshold is set to one less than the protection threshold. Although this occurrence is uncommon, it selects for protection as a trait, which is a limitation of the model.

Originally users also had access to many more parameters, including initial population; social currency gain or loss; neighborhood size (the radius within which turtles stole and protected); and life expectancy. These parameters were hidden from the user as a design decision to make the interface cleaner, and are now set by the setup command. Through experimentation with the model I found that changing the initial population or neighborhood size did not have a large effect on the outcome of the model, but instead made the model either end quickly because all turtles died, or become very slow due to population explosion. The effects of changing social currency gain and life expectancy could be more empirically investigated, but in initial tests most settings other than the social currency gain decided upon (7 units) resulted in the model terminating very quickly and not displaying interesting behavior, or not selecting for any traits. Finally, through experimentation I found that varying the initial social currency and mate threshold (which are still exposed to the user) did not have enormous variations on the model except again to terminate it quickly or lead to population explosions. Also, with larger initial social currencies and lower mate thresholds, the model was much less sensitive to large swings and was less likely to terminate early, which is useful for examining the effects of slight changes in the other parameters. That is, with most theft and protection thresholds and probability of getting away settings, along with the default currency and mate settings, the model will run for some time and show some observable phenomena. I left these parameters exposed to show users how sensitive the model could be to initial conditions, that particular design decisions were made which, if changed, would prevent the model from displaying any behavior. Also, if population is dying out or exploding with certain parameters, slightly changing the mate threshold can somewhat stabilize it. The same is true of other hidden parameters, but exposing these two sufficed to make the point.

I was inspired to build this model by Bernie Madoff and the financial crisis caused by banks and other financial institutions in 2008. I wanted to know how it was that people who stole could become so successful, and if it was truly simply a matter of them being able to get away with it. I also looked to the fields of game theory and evolutionary ethics to find what other researchers, philosophers, and modelers had discovered in this area.

Relevant Research: The evolution of ethics and morality in societies and humanity is a field of rich and historied research. Darwin himself described how social instincts, coupled with societal communication, could lead to the evolution of moral norms (Darwin, 1871, Chapter 4). Modern research has confirmed that "it is reasonable to suppose that many psychological facts [including motivations] are part of the factory-installed equipment that evolution built into us" (Haidt & Joseph, 2004). There exist mathematical models which prove that, under the right circumstances, an altruistic gene can be selected for in a population and spread genetically through that population over time (Matessi & Jayakar, 1975). Researchers have also developed biologically-inspired theories that explain cooperation at a variety of levels, from individual cells to bacterial parasites to insect colonies to humans (West et al., 2007). They note that while reciprocity (helping those who helped you) "has attracted a huge amount of theoretical attention, it is thought to be generally unimportant outside of humans," but that enforcement (the punishing of non-cooperators) is "important in a number of vertebrate species [and] has also been suggested to be an important selective force for cooperation in humans" (West et al., 2007), a claim my model explores.

Incorporating social currency as a simple measure of reputation and the possibility of being rewarded for getting away with theft are sensible factors to include when modeling an individual's

reproductive fitness within a society. As Jonathan Haidt, perhaps the foremost moral and ethical psychologist of our time, says:

> "People in all societies gossip, and the ability to track reputations and burnish one's own is crucial in most recent accounts of the evolution of human morality. The first rule of life in a dense web of gossip is: Be careful what you do. The second rule is: What you do matters less than what people think you did" (Haidt, 2007).

That is: reputation matters; stealing helps if you don't get caught; and being nosy hurts if there is no one to catch.

There is also evidence from evolutionary psychology that socially negative traits can be selected for, especially if they are forcibly passed on, such as through deception or coercion. Although it is a hamfisted metaphor, we can view turtles who get wealthy enough to reproduce by stealing from turtles who do not (or are unable to) defend themselves as passing on their traits in the same way that aggression has been passed on as an inherited trait (e.g., Ferguson & Beaver, 2009). Such exploitative agents seize social capital (and therefore, in my model, the ability to reproduce) from other agents; deceptively raising one's status stands in here for coercive tactics. Exploitation is especially likely in situations where it is not likely to be discovered. For example, older people are particularly vulnerable to exploitations such as fraud and abuse because they "possess some characteristics which make them more vulnerable than young people[:] some older persons are less mentally alert and many are more trusting of people," i.e., they are less likely to identify deception, and are therefore more likely to be exploited (Smith 1999).

In our model, a turtle who has a high social value because it protects can breed with a turtle who has high social value because it gets away with stealing a lot. My model can account for how a population might change stochastically through intergroup mingling, rather than be driven forward inexorably through the refinement of selected traits. As Richerson & Boyd (2005) point out in their work on how culture has in part driven human evolution, "children acquire beliefs and values […] from their parents. Then, as they grow older, their beliefs and values may also be affected by other adults. Next, as adults, they marry," and their children's beliefs and values differ to varying degrees from their children's grandparents. While my model does not account for individual turtles changing over their lifetime, it does model how children change relative to their parents, and can change perhaps drastically from their grandparents, if a turtle likely to steal reproduces with a turtle likely to protect.

There are two crucial ways that my model does not align with standard views of evolutionary ethics: we can interpret this model as having protectors protect because it benefits them, rather than out of pure altruism; and the model does not involve any sort of emotional processes. In most human evolutionary models, protecting behavior and altruism are defined as actions that help others even when they do not help the self, or potentially when they harm the self. In my model, turtles directly benefit from protecting others (since catching thieves nets them social currency), and it benefits them to protect themselves, since otherwise they themselves would be stolen from by an undetected thief. Furthermore, my model has no implementation of emotion whatsoever. The consensus among moral and evolutionary psychologists is that "the building blocks of human morality are emotional (e.g., sympathy in response to suffering, anger at nonreciprocators, affection for kin and allies) and that some early forms of these building blocks were already in place before the hominid line split off" (Haidt,

2007). My model encodes instant punishment, but no modeling of emotions, reciprocity, or true altruism: no turtle ever sacrifices itself for other turtles, and turtles will happily protect a turtle who stole from them the previous turn.

The domain of finding equilibria between agents seeking to maximize their own utility, as my model does, is primarily the realm of game theory. Game theory models of evolutionary ethics have already discovered that self-interested agents can learn to cooperate and develop heuristics such as the golden rule (Skyrms, 1996). The fact that, in my model, some of the simulations eventually select for doing nothing is a good demonstration of this: in some cases, the best possible thing for agents to do is to leave each other alone[2]. If my model rewarded turtles for doing nothing, albeit at a lower rate, perhaps doing nothing would be selected for more as a genetic trait; as it is, doing nothing is only rewarded when other turtles protect in the absence of an attacker.

In game theory, Attacker-Defender games involve a Defender setting up defenses based on a theory of where the Attacker will attack, then the Attacker (knowing what the Defender has done) chooses how to attack (e.g., Cox, 2009). One particularly relevant instantiation was built by Nochenson & Heimann (2012), who used agent-based modeling to simulate the attacker-defender game. In their model, a single defender in a network had to prevent attackers from breaching that network. Through their model, the Defender learned from experience to find an optimal defense strategy, and the attackers had access to a variety of strategies. In my model all movement happens simultaneously, and turtles do not know what their neighbors will do when they select an action; therefore my model represents a simplification of the Attacker-Defender game theory decision model.

One final relevant finding from the game theory literature is the fact that whether we believe someone is watching us affects how we behave. In a clever study, Haley & Fessler (2005) found that simply putting stylized pictures of eyes on the background of a desktop computer (instead of some other image or a blank screen) made people more likely to split money evenly with a partner in the Dictator game.

Observed Phenomena: I began by running a behavior-space exploration, varying all four user-set parameters broadly and running multiple runs of the simulation to a maximum of 200 ticks. I measured the proportion of turtles that performed each type of action at each turn to see how those proportions changed over time. I also measured the proportion of each 'type' of turtle; a turtle was defined as being of a type if it performed a particular action ten or more times. Therefore a turtle who stole ten times was labeled as a "thief". Note that turtles can be of multiple types.

I did indeed find that behaviors were selected for and populations changed over time, as measured by the proportion of turtles selecting one behavior or another. One interesting behavior that appeared in some runs was that turtles would die out if all the thieves died. That is, if stealing was steadily selected against as a trait the population would quickly dwindle to zero once all turtles stopped stealing. This is because the protectors would lose their social currency for being nosy in the absence of thieves, and the placid turtles doing nothing would not gain enough social currency to reproduce and pass on their own genes. This is actually a reasonable demonstration of the fact that societies need balance: thieves need people to exploit, or eventually they will kill all their victims and have no one to

[2] Very few settings selected for doing nothing, and even those only do so once theft is nearly or entirely gone in a population.