Proceedings of the Ninth Goal Reasoning Workshop Submitted 10/2021; published 11/2021

EDA, An Empathy-Driven Computational Architecture

Xinmiao Yu YUXINMIA@GRINNELL.EDU
Riccardo Morri MORRIRIC@GRINNELL.EDU
Fernanda M. Eliott ELIOTTFE@GRINNELL.EDU
ELBICA laboratory, Grinnell College, 1116 8th Ave, Grinnell, IA 50112, USA

Abstract

We aim to design an Empathy-Driven Computational Architecture (EDA) to investigate the role of empathy in weighting different goals important for decision-making and agent-agent long-term interactions. Two of our main assumptions are that, in a human context: 1. Emotions and feelings play an essential part in rational decision-making. 2. Cooperation emerges from the assistance of emotions, feelings, and both, moral *intuition* and moral *reasoning*. EDA's design was inspired by a computational architecture called $MultiA$, which models *moral reasoning* to obtain moral behavior as an emergent property in independent learning agents. Our work complements and expands from $MultiA$, since we add moral intuition into the design. Here, we discuss bio-inspiration and computational approaches, describe a groundwork in preparation to design EDA, and present preliminary experiments in an evolutionary game to assess the current design and inform ongoing modeling.

1. Introduction

It is not your turn yet, but you are busy and decide to cross the street (it is okay, you think: no car is coming). As you start moving, you gaze at the other side of the road, an adult and two kids peacefully waiting. You change your mind and stay; you do not want to be a bad example. Finally, at your office, as there is only one left-handed desk but two left-handed people, you take turns with your peer on a daily basis. Noticing your peer having a difficult day, you give up on your turn and offer the left-handed desk. These examples illustrate that we read our surroundings and context and may suppress or equalize our self-interests and cooperate with others.

Often, the best social outcome requires commitment to cooperative strategies, *i.e.* agents must choose the actions that will only lead to the best outcome if most of the group commits and cooperates. However, cooperation usually brings a cost to the cooperator while defectors benefit from common resources (Wardil & Hauert, 2014): a dilemma emerges between the agent's self-interest and the group's maintenance or preservation. If an agent changes its strategy and stops cooperating, others may face the worst possible outcome (such as in the Prisoner Dilemma Game (Robert, 1984), or if someone sick with Covid-19 does not self-quarantine). Being able to interpret the context and suppress or equalize one's interest with the other's (and cooperate) is essential to keep a group cohesive.

In robotic tasks that require many agents coordinating actions to accomplish a task, it is hard to observe the agents' joint actions (Matignon et al., 2012). A way to try to mitigate the issue is to assume them as independent learning agents. (Agents that are unable to observe the actions and reinforcements of other agents (Matignon et al., 2012).) What if independent learning agents were able to suppress or equalize their interest with other's without the need to exchange data and observe each others' actions? For example, empathy could be used to weight an agent's conflicting goals and assist its decision-making. That could be a resource to assist a group of independent learning agents to coordinate actions and accomplish a task.

However, in utility-based computational approaches, the emergence of cooperation is not easily achievable, especially if agents are unable to share data. Neumann and Morgenstern (2007) introduced essential concepts from Game Theory, such as analysis about environmental possibilities, difficulties, and adequate agent policy response to accomplish goals. Rational agents select their best-response to what is expected to be the best shot from other agents (see, for instance, the Minimax theorem (Von Neumann & Morgenstern, 2007)). If learning agents are unable to share data to accomplish a task, they will have to choose strategies based only on their own experiences and learn to coordinate actions. Still, agents can bring uncertainties to the environment, or an agent may fix its policy while another agent is still exploring the environment (Matignon et al., 2012). Greenwald *et al.*(2005) describe challenges to achieve convergence to a cooperative action policy in self-play and general-sum stochastic games. For example, the problem to achieve cooperative behavior when the Pareto-optimal solution does not coincide with the Nash equilibrium (Nash, 1951) (*e.g.,* as in agent-agent dynamics that can be described by the Prisoner Dilemma Game.) Therefore, how to achieve the best social outcome instead of the best individual one? Our leading hypothesis relies on the assumption that, in a human context, cooperation emerges from the assistance of emotions and both, moral intuition *and* moral reasoning. Also, that human-inspiration can motivate the design and role of empathy in weighting conflicting goals for decision-making to bring up cooperation as an emergent property.

Here, we describe our initial steps to design EDA using $MultiA$ (Eliott & Ribeiro, 2015a,b) as inspiration. $MultiA$ is a bio-inspired multi-agent architecture that implements Reinforcement Learning (RL) techniques (Sutton & Barto, 1998 (2018). $MultiA$ is composed of sensations, emotions, feelings, and of an Empathy Module that enables an action selection that models moral behavior. In the redesign, we aim to: 1. Model moral intuition using Patterson and Eggleston (2017) as inspiration, and 2. Emphasize and amplify $MultiA$'s empathy-driven decision making. Hence, we named our architecture *Empathy Driven Architecture*, or EDA.

2. Background: Bio-inspiration and Cognitive Architectures

Since we will be considering cognitive architectures, it is worth it mentioning the distinction between dualism and physicalism. Both are theories for the relationship between mind and body. Modern dualism arose from Descartes' *Meditations* (2006 (1641), and dualists distinguish between mental and physical properties and principles, whereas physicalists hold that mental properties belong to the physical world (Robinson, 2018). According to Stoljar (2021), popular sub-theories within physicalism include behaviorism, mind brain identity theory, and functionalism. The most

important theories considered in cognitive architectures fall into the consideration of functionalism: reducing mental states and mental processes to functions with defined input and output. Functionalism considers mental processes as causal relationships between actions and stimulus. For example, LIDA (Franklin et al., 2013) implements the Global Workspace Theory (Baars, 1993 (1988), Soar implements the Problem Space Computational Model (Newell, 1992), and both Soar and Clarion (Sun & Peterson, 1998) take insights from the Dual Process Theory (Laird, 2012). Note that although physicalism and functionalism are popular approaches to implement and test scenarios, opponents to these theories contribute to the discussion through insightful thought experiments.

Soar implements the problem space computational model (PSCM), which is a general theory of decision making and problem-solving (Laird, 2012). A key assumption of PSCM is that the problem space, the imaginary spaces where the task environment and possible actions live, are fundamental to reasoning, problem-solving, and decision making: "A problem space is the space of states through which the agent can move using its available operators" (Laird, 2012). A problem space has an initial state and a set of desired states, and to solve a problem is equivalent to performing a problem space search (problem search): selecting and applying operators from the initial state to succeeding states in search of the desired state. In a problem space, states are internal representations of the environment. An agent could travel to a new state through operators, which have preconditions for applying them. "Problem spaces are defined by a set of available operators and the available state information" (Laird, 2012).

During a problem-space search, an agent uses knowledge to select and apply operators, which leads the agent to new states. This process of selecting appropriate operators from long-term memory is called *knowledge search*. During a problem search, the agent generates, combines, or derives knowledge that is not present in long-term memory yet. During knowledge search, the agent uses knowledge from long-term memory and selects operators. Soar considers the influence of emotion from Appraisal Theories of Emotion (those postulate that an agent reviews its situation along with several appraisal dimensions. The values of these dimensions results in the emotions of the agent (Laird, 2012)). The analysis of appraisal values gives rise to emotion, mood, and feeling. Emotion is then defined as the current set of appraisal values; mood, on the other hand, is defined as a decaying average over recent emotions. Agents could only perceive feelings, which are a combination of both mood and emotion (Laird, 2012).

LIDA (Learning Intelligent Decision Agent) from Franklin et al. (2013) implements the Global Workspace Theory and descends from IDA (Intelligent Distribution Agent). LIDA interacts with its environment by continuous action-perception cycles. Within each cycle, the agent performs perception, attention, and action, and learning phases. There are three phases in every cycle of the cognitive process; during the perception phase, the agent receives stimulus from the environment and priming with long-term memory modulo. Associations, percepts, and maps are formed from the long-term memories and sent into the situational model, which are the representation of an agent's current situation. During the attention phase, attention codelets forms coalitions with contents in the current situational model and all interested coalitions are sent to the global workspace. In the global workspace, winning conscious content are broadcast throughout all modules in the architecture. During the action and Learning Phase, learning proceeds, and actions are performed according to the global broadcast (Franklin et al., 2013). Feelings and emotions play a significant role in the

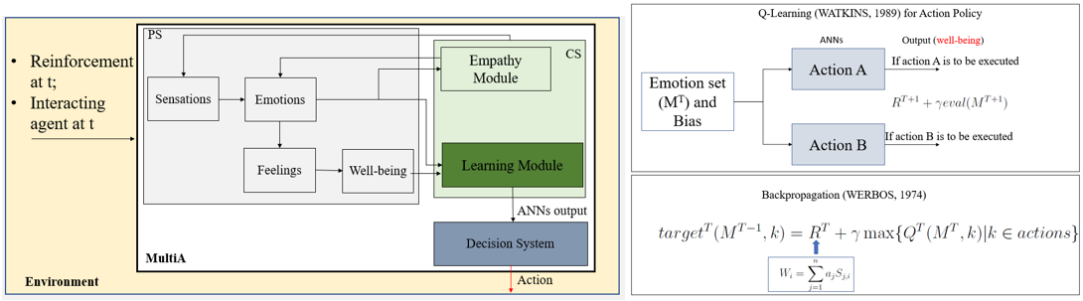

Figure 1. Left. MutiA, an overview: its three main systems (Perceptual, Cognitive, and Decision), the input space and output (action). Right. $MultiA$'s Learning Module, which is embedded in the Cognitive System.

LIDA model by giving an almost immediate assessment of situations. The representation of feelings in the LIDA model is nodes in its perceptual Associative memory. Each node is associated with a specific kind of emotion and feeling (Franklin et al., 2013).

ALEC (Asynchronous Learning by Emotions and Cognition) holds that emotion and cognition are two interacting systems, and both of them contribute to learning performance and problem-solving. The cognitive architecture is tested in a real-world experiment where the agent is challenged by a multi-goal and multi-step decision process with continuous time and space (Gadanho, 2003). ALEC expands on EB (Emotion Based Architecture), which relies on emotional values update for behavior switching in a continuous-time and space environment. One of the challenges for real-world RL is that the agent has to determine when to switch its policy, which can be challenging in a continuous space and world scenario. One possible solution to this problem is to design innate homeostatic emotional values to guide the process of behavior switching (Gadanho, 2003). ALEC improves its learning performance by adapting it to an explicit rule knowledge-based cognitive system. The structure for ALEC consists of an emotion system and a cognitive system; within the emotion system, there are the goal system, and adaptive system. The goal system contains homeostatic variables and calculated well-being values for these variables. The adaptive system uses the well-being values for associating behavior-state pairs with an expected long-term well-being value. Within the cognitive system, there is a dynamic collection of rules guiding the agents to make decisions based on past positive experiences (Gadanho, 2003).

MultiA (Eliott & Ribeiro, 2015b,a) was inspired by the ALEC architecture (Gadanho & Custódio, 2002; Gadanho, 2003), and is a bio-inspired computational architecture that implements RL techniques and includes an Empathy Module to model moral behavior. Through its perceptual system and Empathy Module, it models decision-making differences between moral, immoral, and amoral agents. The Empathy Module was inspired by mirror-neurons, a mechanism that enables the agent to use its own emotions to mirror (and guess) other agents' condition - without data being shared among agents. Thus, an agent cannot observe other agents' actions or reinforcements, but only mirror its own emotions to make assumptions about other agents. The authors used the utilitarian calculus (Bentham, 2007 (1789) as a guideline on determining how the mirrored emotions are used by the Perceptual and Cognitive Systems. Hence, $MultiA$ agents present stronger empathy levels for the agents whose interactions result in positive reinforcements (local, agent-agent reciprocal). Furthermore, a $MultiA$ agent is more likely to cooperate if it has been receiving in general (global) a high number of positive reinforcements. $MultiA$ consists of three main systems:

the Perceptual, Cognitive, and Decision Systems, see figure 1. The environment triggers sensations in $MultiA$. Then, the three systems coordinate to select and apply actions. $MultiA$'s homeostatic goals are to keep feelings within a threshold to maintain its well-being on high levels. To that end, $MultiA$ has to learn an adequate selection of actions in response to the environment. The Cognitive System embeds the Empathy and Learning Modules. In the Learning Module, there is one feed-forward artificial neural network (ANN) *per* action available, and each ANN is indexed to an action. The ANNs model the Q-Learning algorithm (Watkins, 1989) to estimate the expected discounted return for starting from state s (emotions define the state and input space), taking action a, and thereafter following the policy π. The ANNs are trained using the outcome from the execution of their indexed action (Lin, 1993) through the Back-propagation algorithm (Werbos, 1974), whereas the well-being (Damásio, 2004) provides the target value (as figure 1 shows). Finally, the Cognitive System delivers the ANNs' output to the Decision System to select and execute an action, whose selection is based on a variable exploration rate. Sensorial information is triggered by the environment, then, $MultiA$ transforms it into basic and social artificial emotions and feelings. Therefore, its own emotions are employed to estimate the current state of other agents through an Empathy module. Finally, its feelings provide a measure (named well-being) of its performance in response to the environment. Through that measure and RL techniques, the architecture learns a mapping from emotions to actions.

3. Background: Emotions, Feelings, and Decision-Making

Emotions and feelings are acknowledged as crucial to intelligent decision-making: they play an important role in filtering information and awakening our attention mechanisms (see the Somatic Marker Hypothesis from Damasio (Damásio, 1994; Bechara & Damasio, 2005)). We would seek to maintain negative emotions in low levels and the positive ones in high levels, and the purpose of homeostasis would be to produce a state of life better than neutral, the so-called well-being (Damásio, 1994). Emotions include social emotions (such as *sympathy* and its associated *empathy* feeling), which are analyzed from the aspect of social interaction and homeostatic goals (Damásio, 2004). While defining social emotions, Damasio (2004) embraces the concept of moral emotions from Haidt (2003). Haidt (2003) explains emotions as responses to a class of events perceived and understood by the self, and emotions usually provoke action tendencies: social emotions trigger action tendencies during situations that do not represent direct harm or benefit to the self (disinterested action tendencies), while other emotions are more self-centered. Damasio (2004) also investigates the brain's ability to internally simulate emotional states establishing a basis for emotionally possible outcomes and emotion-mediated decision making. Internal simulation takes place while the sympathy emotion turns into the feeling of empathy. The social interaction is facilitated by mirror-neurons (discovered in the premotor cortex area of macaque monkeys (Di Pellegrino et al., 1992; Rizzolatti et al., 1996) by making our brain internally simulate the movements that others do while in our field of vision, for example. Such a simulation enables us to predict what movements are required to establish communication with the other (which will have its movements mirrored).

The origins of empathy can be approached from an emotional aspect (De Waal, 2010; Proctor et al., 2013). The *sympathy* social emotion feeds the feeling of empathy, while social emotions

benefit from the internal simulation improved by mirror-neurons. However, the empathy feeling will be less or more intense depending on the importance of the other Damásio (2004). From an evolutionary perspective, moral behavior can be seen as a kind of cooperation, as the association of skills and reasons for cooperation provide the emergence of moral reasoning: cooperation demands the individual's self-interest equalization with that of the others, or its suppression (Tomasello & Vaish, 2013). In a similar direction, Tomasello (Tomasello & Vaish, 2013) sees cooperation as a sewing up action that connects the members of the group.

According to Damasio (2004), the human body has several levels of homeostatic regulation; at the most basic level there exist immune responses, basic reflexes, and metabolic regulation. At the next level, built from an interweaving of the systems within the first level, there are pain and pleasure. The following level consists of drives and motivations. Although crucial from an animal perspective, there is no purposeful, precise implementation of these systems in EDA. However, the following levels – which are particularly unique to Damasio's point of view – are, in fact, the underpinning of this research. These levels are emotions and feelings. An emotion is "a complex collection of chemical and neural responses forming a distinctive pattern" (Damásio, 2004). The author details that emotion must be triggered by an "emotionally competent stimulus," (Damásio, 2004) which can be an 'actual' event or object occurring in real-time or a recalled event or object. There exist these, so called, emotionally competent stimuli (ECS) that humans have been biologically conditioned to respond to via evolution, and also a response to some ECS that are learned throughout one's life experience. These responses cause immediate changes in the body and brain – particularly the brain's body maps (also referred to as dispositional representations) (Damásio, 2004). The long-term effect of these responses is to motivate people to place themselves in situations that are better for their survival and well-being.

Within the umbrella of emotions, there are three sub-categories of emotion; these include basic, social, and background emotions (Damásio, 2004). The basic and social emotions are fundamental pieces to EDA. Background emotions are an unconscious calculation of the efficiency of the more basic homeostatic processes – immune system, metabolic system, etc (Damásio, 2004). When the body is functioning relatively optimally, the mind gains a subtle disposition towards a more pleasant mood, affecting behavior and possibly reflecting finely in our posture, facial expression, etc. This category of the emotions-proper is less relevant to this work because there is not a significant emphasis on the physical entity that EDA could hypothetically embody. The key difference between basic and social emotions is that the latter are triggered by social interaction and trigger social behaviors (Damásio, 2004). As we progress EDA's design, we will model social emotions such as pride, compassion, gratitude, and sympathy.

It is also important to understand that feelings are the perception of a mentally generated body map (the dispositional representation of our body state), and not the actual body state. We emphasise the distinction because it is part of what allows the emotion sympathy to become the feeling of empathy. Sympathy is an emotion that can be described as the understanding of another's experience. To sympathize with someone else's suffering is to consider and acknowledge their suffering. Empathy is putting oneself into the shoes of another and feeling their suffering to some extent. The human mind can internally simulate body maps of another, allowing the body to feel an emotional experience indirectly. Feelings inform the mind's understanding of the organism's well-being as

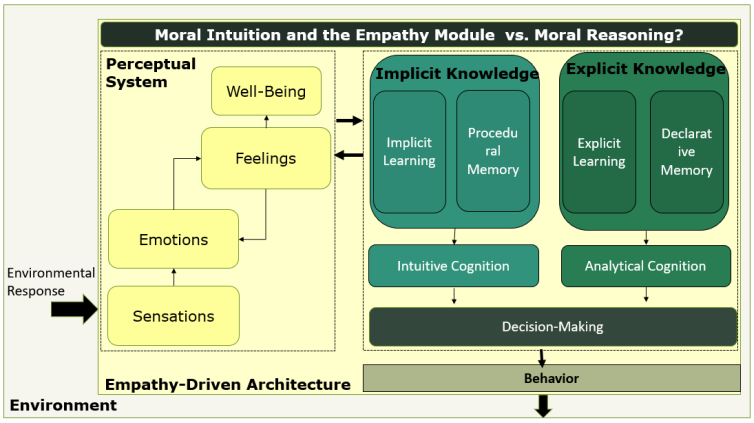

Figure 2. EDA's current draft. We placed the Moral Intuition, Empathy Module and Moral Reasoning at the top followed by a question mark since we are still modeling those.

well as the well-being of others – which is precisely why Damasio categorizes feeling as a homeostatic process (Damásio, 2004). EDA, similarly to $MultiA$, will model well-being values resulting from Empathy response. That is a differentiating feature from other existing computational approaches and serves as a cornerstone for EDA agents decision-making. Just as in Damasio (2004), EDA uses its feelings to calculate its well-being values.

The empathy module is the most distinguishing characteristic of MultiA and, in EDA, we will expand it. To feel empathy for another is to imagine oneself in another's position. The mind predicts how its own emotional state would be affected by the experience of the other. This is precisely how the empathy module works. A $MultiA$ agent projects its own emotions onto the current estimated situation of another agent, calculates an empathy coefficient, and uses that coefficient to help make behavioral decisions.

4. EDA (Empathy Driven Architecture): Modeling Emotions

One of the purposes for building EDA is to drive insights on how empathy works and how it can help independent learning agents to coordinate behaviors and accomplish robotic tasks. EDA's design relies heavily upon biological inspiration, and we hope to provide answers on how empathy works and impacts decision-making in biological creatures, although *we do not intend to classify it as a theory of how empathy works*. EDA substantially expands from $MultiA$ in that EDA's design will seek to encode the brain anatomy (Kandel et al., 2000), but also model and incorporate the distinction between moral intuition and moral reasoning (Patterson & Eggleston, 2017). Within those, we will model the distinction between **1.** Acting in a deliberative way, and **2.** Showing moral response to elicit trust, and therefore **3.** The relationship and differences between guilt and shame (Sperber & Baumard, 2012). In the current design of EDA depicted in figure 2, we added Implicit and Explicit knowledge, but we are still examining how to expand $MultiA$'s Empathy Module. In particular, its capabilities to represent and distinguish moral intuition versus moral reasoning, and how that would impact decision-making differences between moral, immoral and amoral agents.

We are still modeling EDA's emotions but below we present the current design for basic emotions (inspired by (Damásio, 2003, 2004; Bechara & Damasio, 2005; Damásio, 1994)). Emotions fit within the threshold [-1; 1][1]; values approaching -1 mean the emotion is absent whereas high, close to 1 values signify the emotion is at its full capacity. In the current design, only happiness is considered an emotion with positive valence. We are currently framing interaction in simple terms, such as to cooperate or defect (although those could encode more sophisticated behaviors such as to grab and hand objects). In future design and experiments, we will incorporate more complex agent-agent interactions, and use RL to define EDA's learning capabilities. Although **RL techniques are not implemented here** since we are not testing EDA's learning capabilities yet, we already use the term "reinforcement value" to facilitate future implementations. Therefore, we use the term to refer to either positive, negative, or neutral values provided by the environment in response to an agent's actions.

- **Fear.** An agent starts a match[2] with the highest fear levels. It drops as a result of positive interactions (when agent i receives reinforcement above zero) after interacting with a neighbor p[3]. We understand that making an agent start a match with high fear levels may not be ideal; however, this can easily be tuned (*e.g.,* to start with neutral values).
- **Happiness.** It builds upon the happiness values from previous matches, and varies according to the sum of reinforcements after interacting with a neighbor p at the current match t.
- **Anger.** It reflects a comparison between the agent's minimum expected reinforcement value and its reinforcement after interacting with a neighbor p. Anger increases as agent i receives reinforcements below the minimum expected across matches. (The minimum expected reinforcement value enables us to tune an agent's expectations about its environment.)
- **Sadness.** It builds upon the sadness values from the previous match $t-1$, and it increases if agent i does not get positive reinforcements after interacting with a neighbor p.
- **Disgust.** It builds upon disgust values from the previous match $t-1$, and increases as the agent loses a neighbor in the current match t. (An agent may lose a neighbor if it is eliminated from the environment or if the neighbor simply does not interact with an agent at that particular match.) Future design will consider the hypothesis of gaining new neighbors.
- **Surprise.** It increases as agent i loses a neighbor in the current match t. It reflects only what happens at the current match t. In opposition to disgust, surprise should both, build up and go away quickly.

1. We provide a presentation for each emotion in an online document.
2. Suppose an agent is playing a game with a group of other agents. Once all active agents have provided their responses for that round, agents may be eliminated according to the rules of the game, and a new round starts with the remaining agents. We use "match" to describe rounds of a game played in sequence until a final condition is met (*e*.g. all agents are eliminated, or the system reaches an steady condition, or a maximum number of matches is met).
3. Every agent that interacts with agent i is considered its neighbor and is referred to by p. Interactions trigger the environment to generate a reinforcement value, even if it is zero. An agent keeps a history for all the neighbors it has interacted with.

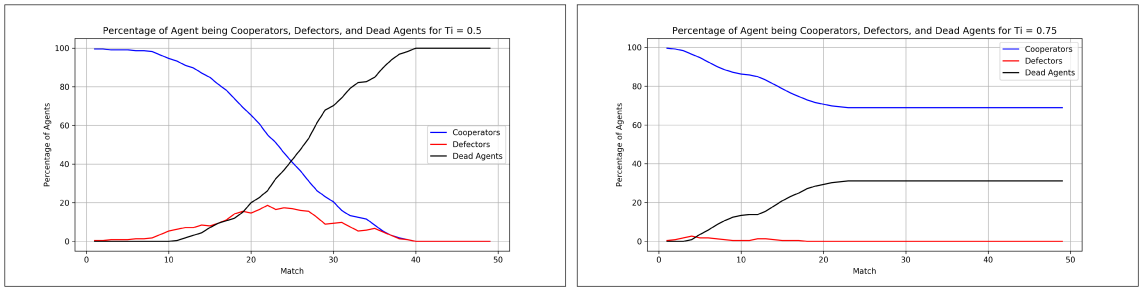

Figure 3. Percentage of cooperators, defectors, and eliminated agents across matches. Left: threshold for elimination T_i is 0.5. Right: $T_i = 0.75$.

5. Experiments

To examine the emotions' design, we used the generalized PDG model described in (Wang et al., 2011). It starts with a network and agents are represented by nodes, while a neighborhood is represented by the links between nodes. Evolutionary games are described in (Wang et al., 2011) and related to the emergence of cascading failures: agents (nodes) and links being eliminated from a network as the outcome of agent-agent interaction. As the elimination of an agent can cause the elimination of other agents, the elimination process can continue until it causes the complete elimination of links and agents. Agents connected through a link (considered neighbors) choose to defect or to cooperate, and the matching strategies define the reinforcement that each agent receives. Once all agents have interacted with each and every single neighbor, a match ends and the individual sum of each agent's reinforcements is calculated. Agents that get an individual sum of reinforcements below a threshold T_i are completely eliminated from the network. A match is defined by all agents interacting only once with every neighbor. Matches are repeated in sequence until the network topology stops changing as a consequence of agents interactions. Once agents fix strategies and the elimination process ends, a network is either steady or completely eliminated.

Since our main goal is to investigate how the emotions change due to agent-agent interaction and not the topology of such interactions, we used a small grid of 15 *vs.* 15 hand-designed agents, and we embedded the emotions into those agents. All agents (but the one at the center of the grid) start a match as cooperators. Agents imitate a more successful neighbor (a neighbor who received higher average reinforcement in the previous match) with a probability set to 0.25. Through imitation, agents can either change strategy to become a defector or a cooperator. V_i is the number of neighbors agent i has at the beginning of a simulation (at the very first match). For mutual cooperation, cooperators receive $1/V_i$ as reinforcements, whereas for defection-cooperation, the defector receives $2/V_i$; all remaining matching strategies render zero reinforcement values. We ran two sets of experiments, one for the threshold $T_i = 0.5$, and another for $T_i = 0.75$ – note that, the higher the threshold, higher the need for cooperators to sustain the network and block the agent elimination process. Each set of experiments comprehends 20 different simulations and the graphics show averaged values across those.

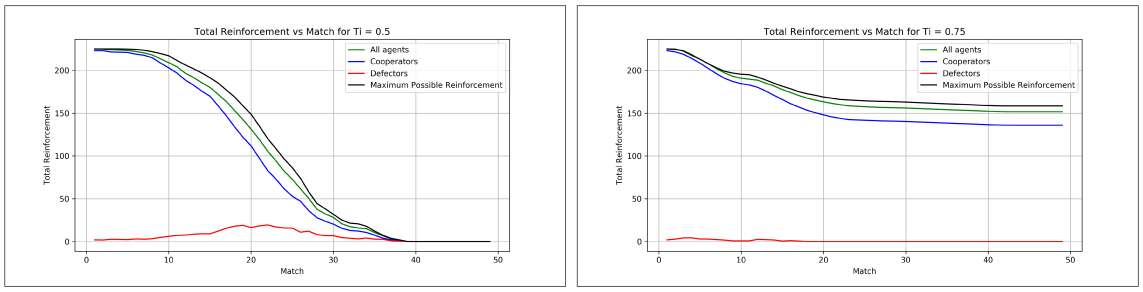

Figure 4. Sum of reinforcements according to the agents' strategies in comparison to the maximum reinforcement (if all alive agents were cooperators). Left: threshold for elimination T_i is 0.5. Right: $T_i = 0.75$.

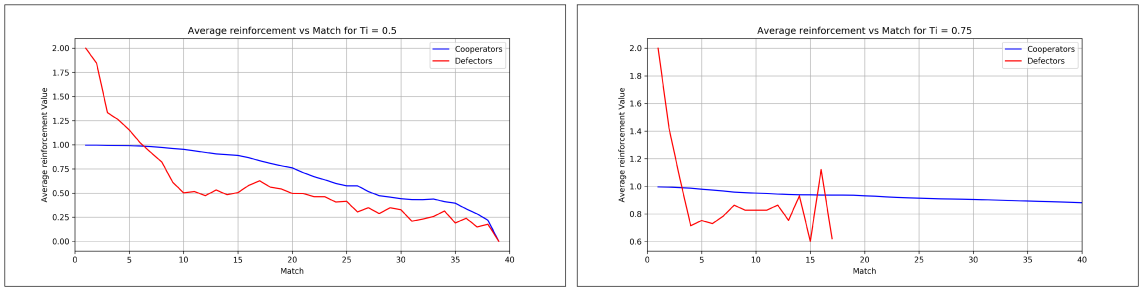

Figure 5. Average reinforcement per agent across matches. Note defectors getting double reinforcements in comparison to cooperators from interacting with cooperating neighbors.

We used the same weight value across emotions that rely on the history of previous matches: a 20% weight on history, and 80% on present values. Anger uses a variable to define an agent's expected minimum reinforcement value: here, we set a fixed expected minimum value of 0.1. In future work, we will make the expected reinforcement vary according to an agent's situation (*i.e.* it could change if an agent loses or gains neighbors). In figures 3 – 5 we depict the overall results; then, in figures 7 – 11 we show how the emotions responded to these experiments according to the agents' strategies. As expected, not only the elimination process *per se* impacts the agents' emotional values, but also the elimination continuity (as the disgust emotion shows). In future experiments, we will investigate how an increase in the number of neighbors enlightens the design of emotional variables, and various elimination processes as well.

- Figure 3 shows that when $T_i = 0.5$, the number of cooperators drops significantly due to: **1.** cooperators are imitating their neighbors and turning into defectors very quickly. **2.** cooperators (and defectors) are eliminated due to the lack of reinforcements from interacting with defectors. The number of eliminated agents also rise substantially at beginning of the match, and the number of defectors increases at a much slower rate because many defectors are eliminated a few matches after they defect. Eventually, all agents are eliminated due to the lack of cooperators. When $T_i = 0.75$ the relatively high survival threshold causes the elimination of defectors and

prevents the defective strategy from spreading across the network. About 70% of cooperators survive whereas all defectors are eliminated.

- In figure 4, the maximum possible reinforcement shows the total reinforcement if all alive agents were cooperators. When $T_i = 0.5$, the total reinforcement for cooperators drops significantly as cooperators continuously face defectors and more agents are eliminated from the network. The total reinforcement for defectors does not increase considerably given that, as the defective strategy spreads, more defectors (and cooperators) get zero reinforcement. When $T_i = 0.75$, the total reinforcement for cooperators drops because some cooperators either face defectors or lose neighbors. However, once defectors are eliminated, the total reinforcement for cooperators remains at a steady level.

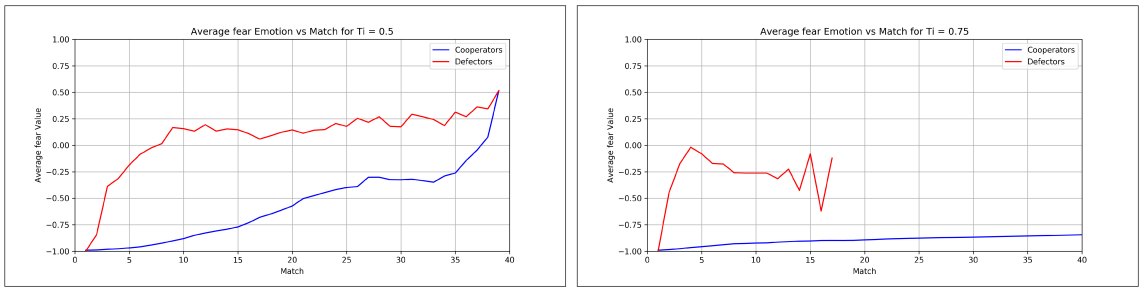

Figure 6. Average fear per agent and match across 20 simulations for $T_i = 0.5$ and $T_i = 0.75$. Higher the value, higher the emotional level.

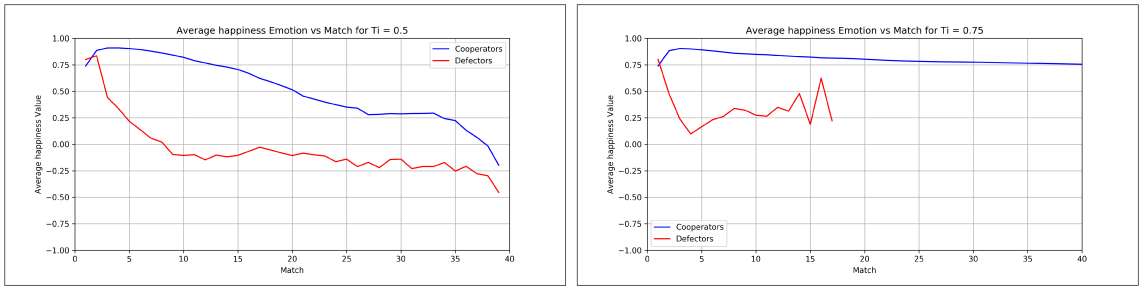

Figure 7. Average happiness per agent and match across 20 simulations for $T_i = 0.5$ and $T_i = 0.75$.

- In figure 5 when $T_i = 0.5$, although the average reinforcement for defectors is higher at the first few matches, it drops considerably. In match 7, the average reinforcement of defectors is lower than that of cooperators. The average reinforcement of cooperators slowly drops as the number of defectors and eliminated agents increase. When $T_i = 0.75$, the average reinforcement for defectors fluctuates more violently because agents are eliminated at a rapid pace. The average reinforcement for cooperators remains at a stable level.

- **Fear**, figure 6. Overall, fear decreases as agents get positive reinforcements. When $T_i = 0.5$ the average fear in defectors rises exponentially at first but then levels off at around 0.25. The average fear in cooperators increases slowly at first (until close to match 33), but then significantly. Both, cooperators and defectors reach about 0.5 fear levels before their final match, when all agents are eliminated due to the lack of reinforcements. (That is when agents show maximum fear, although they are eliminated before that.) When $T_i = 0.75$, the average fear in defectors rises significantly before match 5 but fluctuates violently afterward. The line plot for defectors ends at about match 17 because all of them are eliminated at that point. The average fear in cooperators remains steady.
- **Happiness**, figure 7. Happiness depicts how well an agent has been receiving reinforcements. When $T_i = 0.5$ the average happiness in defectors rises slightly at first but soon drops significantly before leveling off at around -0.25. For cooperators, average happiness increases slightly at first (reflecting the history of positive interactions) and then drops slowly as the number of cooperators diminishes. When $T_i = 0.75$, the average happiness for defectors drops significantly and fluctuates before coming to an end when they are eliminated from the network while the average happiness remains steady for cooperators.
- **Anger**, figure 8. It depicts if the agent is getting reinforcements close to what it expects to receive. When $T_i = 0.5$ the average anger for both, defectors and cooperators drops slightly at beginning of the simulation, but then starts to increase, although at a higher rate in defectors. The average anger for both ends close to 0.5. When $T_i = 0.75$, average anger increases and fluctuates until defectors are eliminated. The average anger remains steady for cooperators. As figures 8 and 9 show, anger and sadness show similar results for the current experiments. That happens due to setting the agents' expectations of reinforcement to 0.1. However, in cases where highly negative reinforcements are frequent, one could change that expectation to -0.1, for example. Another possibility would be to make the expectation vary to reflect changes in the environment. We will accommodate that in future design.
- **Sadness**, figure 9. Reflects each agent-agent interaction and increases as an agent does not get positive reinforcements. When $T_i = 0.5$, the average sadness in defectors and cooperators increase, but that of defectors increases at a faster rate. The average sadness ends at around 0.5, before agents are eliminated from the network. When $T_i = 0.75$, the average sadness increases and fluctuates before defectors are eliminated. The average sadness remains steady for cooperators.
- **Surprise**, figure 10. This emotion shows the effect of losing neighbors is in the current match. That is an interesting emotion to observe negative outcomes caused by defection. When $T_i = 0.5$, the average surprise remains at -1 since no agent is eliminated in the beginning. After match five, the average surprise for cooperators and defectors increases, but that of defectors increases more rapidly, since they are causing (and experiencing) more eliminated neighbors than cooperators do. The average surprise in both cooperators and defectors is 1 at the end of the simulation since all agents end up eliminated and surprise takes into account only the present scenario. When $T_i = 0.75$, the average surprise of defectors fluctuates significantly, while that of cooperators remains steady.

- **Disgust**, figure 11. When $T_i = 0.5$, the average disgust for both cooperators and defectors remains at -1 for the first five matches when there are no eliminated agents. After match five, the average disgust for both cooperators and defectors increases, but that of defectors increases more rapidly, because defectors are having more eliminated neighbors than are the cooperators. The average disgust for both cooperators and defectors is around 0.5 at the end of the simulation. When $T_i = 0.75$, the average disgust emotion of defectors fluctuates significantly, while that of cooperators remains steady. Disgust and Surprise look similar due the continuous elimination of agents in $T_i = 0.5$ and the steady agent population at $T_i = 0.75$ for cooperators (the experiments described do not make that evident the distinction between the fast pace of surprise in opposition to disgust).

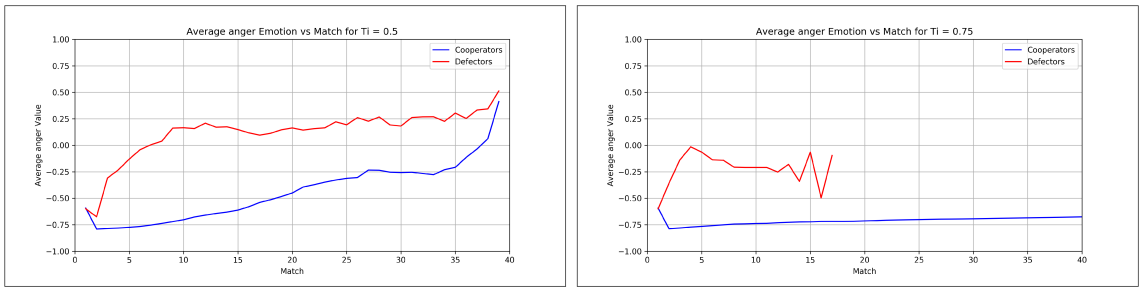

Figure 8. Average anger per agent and match across 20 simulations for $T_i = 0.5$ and $T_i = 0.75$.

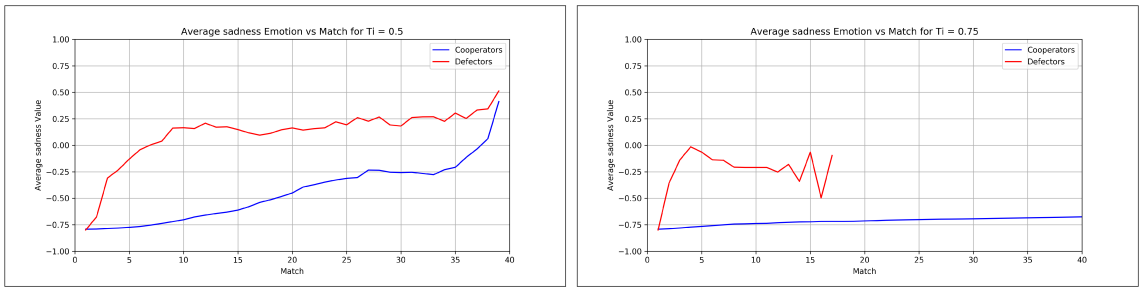

Figure 9. Average sadness per agent and match across 20 simulations for $T_i = 0.5$ and $T_i = 0.75$.

6. Final Remarks

We described a groundwork in preparation to design EDA, a computational architecture that drives inspiration from $MultiA$ (Eliott & Ribeiro, 2015a,b). $MultiA$'s most distinguishing feature is its empathy module and, in EDA, we will redesign and expand it. The empathy module is inspired by mirror neurons (Di Pellegrino et al., 1992; Rizzolatti et al., 1996), which utilize one's own current emotional state and projects it onto another person's situation. The mind then predicts how

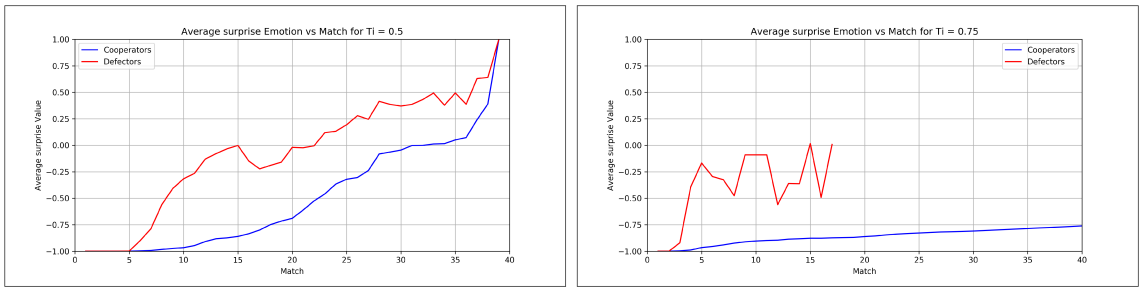

Figure 10. Average surprise per agent and match across 20 simulations for $T_i = 0.5$ and $T_i = 0.75$.

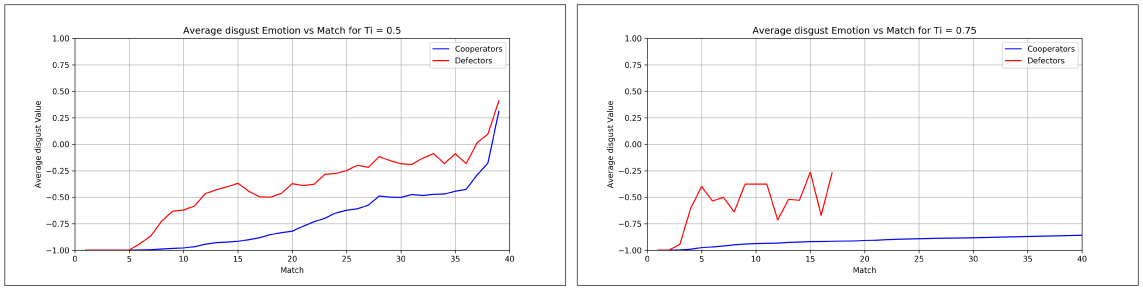

Figure 11. Average disgust per agent and match across 20 simulations for $T_i = 0.5$ and $T_i = 0.75$.

its emotional state would be affected by the stimuli. $MultiA$'s empathy module uses that as inspiration. A $MultiA$ agent projects its own emotions onto the current estimated situation of another agent, calculates an empathy coefficient, and then uses it to weigh behavioral goals. In EDA, we will model moral intuition (Patterson & Eggleston, 2017), emphasizing and amplifying $MultiA$'s empathy-driven decision making.

In the experiments, we focused on positive and neutral reinforcements, although indirect "punishments" happen through the elimination of agents. We will keep modeling the emotions to work through negative reinforcements, and enhance the differentiation between emotions (*e.g.*, in the experiments, anger and sadness show similar behavior for the adopted parameters). However, similar values across distinct emotions help to show that experimental parameters and test-bed play an important role for both, design and applications. We will continue modeling the emotions with different experimental tasks and parameters to identify weaknesses and strengths in EDA's design. Here, we framed interaction in simple terms (to cooperate or defect), although we will incorporate more complex agent-agent interaction possibilities in the future. Future work consists of designing EDA's modules, social emotions, moral reasoning and moral intuition (along with decision-making differences between moral, immoral and amoral agents), and identifying approaches to examine and test empathy-driven behavior.

Acknowledgements

The authors would like to thank Grinnell College's Mentored Advanced Projects (MAP program).

References

Baars, B. (1993 (1988)). *A cognitive theory of consciousness*. Cam. U. Press.

Bechara, A., & Damasio, A. (2005). The somatic marker hypothesis: A neural theory of economic decision. *Games and economic behavior*, *52*, 336–372.

Bentham, J. (2007 (1789)). *An introduction to the principles of morals and legislation*. Courier Dover Publications.

Damásio, A. (1994). *Descartes' error: emotion, rationality and the human brain*. Picador.

Damásio, A. (2003). Feelings of emotion and the self. *Annals of the New York Academy of Sciences*, *1001*, 253–261.

Damásio, A. (2004). *Looking for spinoza: Joy, sorrow, and the feeling brain*. Random House.

De Waal, F. (2010). *The age of empathy: Nature's lessons for a kinder society*. Broadway Books.

Descartes, R. (2006 (1641)). *Meditations, objections, and replies*. Hackett Publishing.

Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: a neurophysiological study. *Experimental brain research*, *91*, 176–180.

Eliott, F. M., & Ribeiro, C. H. (2015a). Emergence of cooperation through simulation of moral behavior. *Hybrid Artificial Intelligent Systems (HAIS 2015). Procs. of the 10th I. Conf. on Hybrid Artificial Intelligence Systems, Bilbao, Spain. Lecture Notes in Artificial Intelligence* (pp. 200–212). Springer International Pub. Also available as https://link.springer.com/book/10.1007/978-3-319-19644-2.

Eliott, F. M., & Ribeiro, C. H. (2015b). Moral behavior and empathy modeling through the premise of reciprocity. *Procs. of the 1st International Conference on Human and Social Analytics*. St. Julians, Malta: Huso. Also available as https://www.thinkmind.org/articles/huso_2015_3_20_70033.pdf.

Franklin, S., Madl, T., D'mello, S., & Snaider, J. (2013). Lida: A systems-level architecture for cognition, emotion, and learning. *IEEE Transactions on Autonomous Mental Development*, *6*, 19–41.

Gadanho, S. C. (2003). Learning behavior-selection by emotions and cognition in a multi-goal robot task. *Journal of Machine Learning Research*, *4*, 385–412.

Gadanho, S. C., & Custódio, L. (2002). Asynchronous learning by emotions and cognition. *Proceedings of the seventh international conference on simulation of adaptive behavior on From animals to animats* (pp. 224–225). MIT Press.

Greenwald, A., Hall, K., & Zinkevich, M. (2005). Correlated q-learning. *Brown University Technical Report*.

Haidt, J. (2003). The moral emotions. *Handbook of affective sciences*, *11*, 852–870.

Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S., & Hudspeth, A. (2000). *Principles of neural science*, volume 4. McGraw-hill New York.

Laird, J. (2012). *The soar cognitive architecture*. The MIT Press. MIT Press. From https://books.google.com/books?id=Z9bxCwAAQBAJ.

Lin, L. (1993). *Reinforcement learning for robots using neural networks*. Doctoral dissertation, Carnegie-Mellon Univ.

Matignon, L., Laurent, G., & Le Fort-Piat, N. (2012). Independent reinforcement learners in cooperative markov games: a survey regarding coordination problems. *The Knowledge E. Rev.*, *27*, 1–31.

Nash, J. (1951). Non-cooperative games. *Annals of mathematics*, (pp. 286–295).

Newell, A. (1992). Unified theories of cognition and the role of soar. In *Soar: A cognitive architecture in perspective*, 25–79. Springer.

Patterson, R. E., & Eggleston, R. G. (2017). Intuitive cognition. *Journal of Cognitive Engineering and Decision Making*, *11*, 5–22.

Proctor, D., Brosnan, S. F., & de Waal, F. (2013). How fairly do chimpanzees play the ultimatum game? *Communicative & integrative biology*, *6*, 2070–5.

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive brain research*, *3*, 131–141.

Robert, A. (1984). The evolution of cooperation.

Robinson, W. (2018). Dualism. In *The routledge handbook of consciousness*, 51–63. Routledge.

Sperber, D., & Baumard, N. (2012). Moral reputation: An evolutionary and cognitive perspective. *Mind & Language*, *27*, 495–518.

Stoljar, D. (2021). Physicalism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, Summer 2021 edition.

Sun, R., & Peterson, T. (1998). Autonomous learning of sequential tasks: experiments and analyses. *IEEE transactions on Neural networks*, *9*, 1217–1234.

Sutton, R. S., & Barto, A. G. (1998 (2018)). *Reinforcement learning: An introduction*. MIT press Cambridge.

Tomasello, M., & Vaish, A. (2013). Origins of human cooperation and morality. *Annual review of psychology*, *64*, 231–255.

Von Neumann, J., & Morgenstern, O. (2007). *Theory of games and economic behavior (commemorative edition)*. Princeton university press.

Wang, W., Lai, Y., & Armbruster, D. (2011). Cascading failures and the emergence of cooperation in evolutionary-game based models of social and economical networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *21*, 033112.

Wardil, L., & Hauert, C. (2014). Origin and structure of dynamic cooperative networks. *Scientific reports*, *4*, 5725.

Watkins, C. (1989). *Learning from delayed rewards*. Doctoral dissertation, Kings College, UK.

Werbos, P. (1974). *Beyond regression: New tools for prediction and analysis in the behavioral sciences.*. Doctoral dissertation, Harvard.