THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF ENGINEERING SCIENCE AND MECHANICS

USING COMPUTATIONAL NARRATOLOGY TO ADDRESS THE ARTIFICIAL INTELLIGENCE VALUE ALIGNMENT PROBLEM

ALAYNA A. KENNEDY

SUMMER 2018

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in Engineering Science
with honors in Engineering Science

Reviewed and approved* by the following:

Lucas J. Passmore
Associate Professor of Engineering Science
Thesis Reader

Gary L. Gray
Associate Professor of Engineering Science
Honors Adviser and Thesis Sponsor

Judith A. Todd
Department Head
P.B. Breneman Chair and Professor
of Engineering Science and Mechanics

* Signatures are on file in the Schreyer Honors College.

ABSTRACT

This thesis provides a novel conceptual contribution to artificial intelligence (AI) safety by finding a tractable method for solving the AI value alignment problem: the creation of more complete audience models using narrative information extraction techniques from the field of computational narratology. With a thorough analysis of results from the field of computational narratology, I show that research into narrative for autonomous agents can contribute to solving the AI value alignment problem. In short, we can create artificial intelligence systems that automatically act in the best interest of humanity by teaching them to read and understand stories.

The novelty of this thesis lies in the combination of two disparate academic fields: AI safety and computational narratology. Reviewing the current work and ongoing issues in both fields, I show that methods used in computational narratology to model stories can be used to solve the value alignment problem from the field of AI safety.

In Chapter 2, I show why value alignment is the best solution to the problem of controlling intelligent agents. In Chapter 2, I discuss how stories encode tacit human values, and how the creation of a better audience model will contribute to solving the value alignment problem. In Chapter 3, I present two case studies providing evidence that value alignment from narrative information extraction is not only viable, but effective. Finally, I conclude by acknowledging the shortcomings of the field and pressing areas of future work.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

The staff of the Engineering Science department deserves the most thanks for helping me to pursue a thesis topic that I am genuinely fascinated by and extraordinarily proud of. Especially Dr. Gray, who has afforded me much more patience than I probably deserve, and who always is ready with a cup of Starbucks coffee for a thesis writer under severe mental stress. Thanks to Dr. Todd for encouraging my grad school dreams, and Dr. Passmore, for several unconventional (but brilliant) ideas for new machine learning algorithms. To my many academic and professional mentors throughout college – Dr. Ozment, Dr. Brady, Jacque, Dr. Lewis, and Lou – thank you for shaping me, guiding me, and encouraging my curiosity!

I am also extraordinarily grateful for the experience the Honors College has given me throughout college, including these past few months of thesis writing. Without being forced to, I would never have created this document of which I am so proud. Donna Meyer, Lisa K., Dean Johnson, and many more have paved my path to graduation. Thanks to Deb Rodgers and Melissa Myer for their wonderful work in getting so many students to submit finished theses.

And finally, a most special thanks to my family and friends: my parents, brothers, Maddie, Tessa, Sammie, Travis, Emily, Adam, and so many more, who allowed me to complain about writing a thesis for several months straight. Mom and Dad – for making my life and dreams possible, I thank you with my whole heart. Everything I have ever done, and any future contributions I may make to the world have been yours too. All I do is possible because of you!

Chapter 1

Introduction

General topic and background

The general topic of this thesis is the proposal of a new field of study for artificial intelligence (AI) safety researchers: intelligent agents with value-aligned reward systems created by narratives. AI safety is of growing importance, as AI technology improves rapidly with very few constraints. Many researchers are working on technical and policy-based solutions to the AI control problem – how to control artificial agents in order to prevent them from causing harm to humans. However, the field of AI safety is relatively new and few tractable solutions to the control problem have been found. In this thesis, I propose that the field of computational narrative intelligence provides a solution to the AI control problem. Thus far, the work in these two fields have not been combined to provide a tractable solution to the AI control problem. This work attempts to correct that gap.

Importance of the topic

Developing a solution to the AI control problem could be the greatest challenge facing the existence of humanity in the 21st century. Should AI continue to develop exponentially (as it has been doing for the past several decades) (Bostrom, 2013), artificial systems will surpass humans in all general cognitive capabilities within the next fifty years (Grace, 2017). Without a

mechanism of control in place before the emergence of such as superintelligence, the effects on humanity could be disastrous. The problem facing AI safety researchers is how to best develop a mechanism of control for AI using current knowledge. In this thesis, I propose that value alignment (a property of an intelligent agent indicating that it can only pursue goals that are beneficial to humans) presents the most tractable solution to the control problem. Furthermore, I argue that value alignment can be achieved through methods of computational narratology, a field dedicated to creating AI agents that can understand and generate meaningful stories. Specifically, I use multiple case studies to show that narrative information extraction of an audience model (a model of the human population of readers) is a computationally viable and theoretically sound solution to the value alignment problem.

Order of information and methodology

The novelty of this thesis is that it combines two previously disparate fields of study: AI safety and computational narratology. Therefore, I begin by summarizing trends in both of those fields. I discuss AI safety in chapter 2 and computational narratology in chapter 3. In those two chapters, I lay a theoretical groundwork for my hypothesis that narrative information extraction of stories is the best solution to the value alignment problem. In chapter 4, I provide multiple case studies with evidence to support my claims. I discuss the computational methodology involved in narrative extraction and present the results of the case studies I have done to support my claim. I conclude in the fifth chapter with possible challenges to be faced and future lines of work.

Chapter 2

Abstract

In this chapter, I provide a review of the literature related to AI safety. I identify the importance of solving the control problem before the emergence of a superintelligent AI system. I then provide theoretical reasoning why the control problem can only be solved by value alignment – encoding machines with a human system of values. Value alignment is a difficult task because human values have not yet been formalized in a robust way. I conclude by arguing that books, novels, and written stories provide a wealth of data on tacit human values, and that they must be analyzed in depth in order to solve the value alignment and control problems.

Introduction

Researchers predict that artificial intelligence (AI) will outperform humans in all tasks within 50 years (Grace, 2017). Therefore, it is crucial to develop a strategy to control AI before it surpasses human cognitive capacities. Traditionally, AI has advanced to solve problems – win a game of AlphaGo or recognize faces – and as it advances, our dependency on AI deepens (Chong, 2017). The current unabated growth in the problem-solving abilities of AI neglects important aspects of control and safety (Bostrom, 2013). As rapid advancement continues, the cognitive capabilities of artificial intelligence will eventually surpass humans' (Muehlhauser, 2014). How then, will we control such systems if we do not find a solution now? To avoid the disastrous consequences of an ill-founded relationship between humankind and cognitive artificial intelligence, we must find a way to align the goals of humanity with those of computers.

Imparting our ethical framework of decision making to machines would align human and machine goals and essentially solve the AI control problem. We can control AI by installing it with an artificial "conscience," built with the same technology AI uses to solve problems (Bostrom, 2013). However, instilling an AI with values or conscience is incredibly difficult. The current limitation of such a task lies in our lack of a suitable data source to model human values and ethics on (Chong, 2017). Promising ideas in natural language processing (NLP) are emerging (Riedl and Harrison, 2017), suggesting there is a great deal of rich, contextual information on human values present in text-based data that we have not yet utilized. Therefore, an exploration of text using NLP presents a promising avenue to uncover human values and instill these values into machines (Mani, 2013; Riedl, 2016).

Current AI control schemes do not thoroughly utilize the potential of NLP to solve value alignment, focusing instead on rules-based programming and human operator intervention. Such rules-based and motivation selection methods are ultimately unsatisfactory or incomplete solutions to the control problem (Bostrom, 2013; Milli, 2017; Saunders, 2017). Rules-based programming is too rigid to convey the complexity of human values to computers, and current motivation selection methods rely too heavily on unreliable human decision-making (Anderson, 2007). Therefore, I argue that value alignment presents our best option to create safe AI. Furthermore, NLP possesses the technical capabilities to make real progress in enculturing machines with human values because of its potential to understand narratives. By teaching machines to understand and processes human stories, we can enculture them with human values, creating safe and controllable AI.

The Approach of Superintelligent Machines

The field of Artificial Intelligence (AI) began in the 1940s with Alan Turing's thesis that any computational function, including the processes of the human mind, can be so precisely described that it can be realized by a machine (Church & Turing, 1937). Cognitive scientists in Turing's time were optimistic about the future of AI, promising human-level intelligence by the 1960s. However, after ultimately failing to create marked progress in the field, AI researchers shifted their focus from the cognitive sciences to more specific scientific and engineering problems, where rapid progress has occurred in recent years (Lewis, 2014). As AI has grown exponentially over the last few years, so too has concern about the implications of high-level machine intelligence (HLMI) on the future of humanity (Bostrom, 2013).

Results from a large survey of machine learning and AI researchers show that experts believe there is a 30% chance that the development of a superintelligent AI system will be "extremely bad" for humanity (Muller, 2014). Respondents to another survey predict that AI will outperform humans in many activities in the next ten years, such as translating languages (by 2024), writing high-school essays (by 2026), driving a truck (by 2027), working in retail (by 2031), writing a bestselling book (by 2049), and working as a surgeon (by 2053) (Grace, 2017). The median estimate of respondents was for a 50% chance that high-level machine intelligence will develop around 2040-2050, rising to a 90% chance by 2075. Experts expect that systems will move on to superintelligence before 2100, within the lifetime of children being born today (Muller, 2014).

The development of HLMI appears to be rapidly approaching, and the creation of such an intelligence would transform modern life by reshaping transportation, the job market, healthcare,

science, finance, the military, and our personal lives (Bostrom, 201e). AI could vastly increase wealth, health, and overall well-being, but could also radically reduce employment prospects and national security (Muller, 2003). In the long-term, the emergence of machine superintelligence (AI that is vastly better than humans are at most important tasks) would enable revolutionary changes "more profound than the agricultural or industrial revolutions" (Dafoe, 2018). We need to better anticipate and understand the technology in order to create policies that maximize the benefit and minimize the detriments of AI. Specifically, researchers must find a way to control advanced AI systems when they develop into HLMI. This multi-disciplinary issue is the AI control problem.

Background on the Control Problem

In discussing the AI control problem, three main questions must be addressed. First, is this an urgent problem? If so, is it a difficult one? Finally, if it is urgent and difficult, how do we go about laying the foundations for safe AI now?

Urgency

Current AI systems fall far short of HLMI. All artificial agents that exist today are niche systems, capable of a very narrow range of human-like behaviors only after extensive engineering and backend programming. Although programs like Siri and Cortana can simulate human speech and tonal reflection, they do not possess anything remotely close to human problem-solving abilities. Despite the impressive advance of niche systems, skeptics argue that

the realization of true generic AI are as distant as they have always been, especially when measured against the extraordinary cognitive capacities of humans (Dangelo, 2017). Such skeptical researchers argue that thinking about the control problem now is, as famous AI researcher Andrew Ng once said, like "worrying about overpopulation on Mars" (Garling, 2015). These researchers say that any control methods developed today would not only be incredibly difficult to implement in current AI, but also that such developments would probably be rendered useless by subsequent technological developments (Dafoe, 2018).

However, given the wide-ranging uncertainty among AI experts as to when superintelligence might emerge, rapid AI timescales could very well lead to control issues much sooner than anticipated. Prominent figures in the world of technology have raised concerns about the existential risks AI presents. Namely, Elon Musk called AI "the greatest risk we face as a civilization," warning proactive government regulation of AI is needed before AI becomes superintelligent (Morris, 2017). Given the enormous existential risks that a true superintelligence would bring, controlling it is an urgent problem, even if we still don't know when it may occur (Grace, 2017; Muller, 2003).

Difficulty

In debating the difficulty of the control problem, many give simple and ultimately unsatisfactory solutions (Bostrom, 2013). For example, some researchers suggest limiting the capabilities of AI physically; installing a "kill switch" that can always turn the machine off if it gets out of hand (Dafoe, 2018). These capability control methods limit the scope of what a

superintelligence can do in order to prevent it from having a negative impact. The idea is that by intentionally making AI less powerful, we can make it harmless (Bostrom, 2013).

However, many leading thinkers on the subject of AI believe that controlling AI capability can only serve as temporary and auxiliary measures to control AI (Muller, 2003; Dafoe, 2018). Therefore, it will eventually be necessary to master some form of motivation selection (controlling what AI *wants* to do) in order to ensure safe AI development (Bostrom, 2013). Simplistic solutions to the wicked problem of AI control fail to consider the true implications of superintelligence, the complexities of AI programming, and the fact that even humans do not understand their own morality, making the task of transferring morals to machines exponentially more difficult (Bostrom, 2013).

Methodology

Institutions working on AI safety and control have technical agendas full of many open problems, such as highly reliable agent designs, figuring out the induction problems faced by intelligent agents interacting with their environment, how to develop decision theory in smarter-than-human systems, and dealing with logical uncertainty (Bostrom 2013; Riedl, 2016). Experts question which technical developments would be most useful to implement now, versus the ones that cannot be solved until AI develops more intelligent capabilities. Among the more pressing open problems is that of value alignment, synchronizing human and machines values to create ethical machines. The next section will delve into the methodology of solving the control problem by focusing on value alignment.

Capability Control and Motivation Selection Methods

The new agency problem

Economists, psychologists, and business managers have extensively studied traditional agency problems involving human sponsors and developers (Ferber, 1999). In such agency problems, project managers worry that the developers implementing the project will not act in the sponsor's best interest. For example, an underpaid and unmotivated programmer might intentionally overlook bugs in code that could be detrimental to their employer later down the line. Techniques such as personnel background checks, supervisor oversight, and increased employee are tools that can solve this version of the agency problem. However, once this problem expands to artificially intelligent agents, instead of humans, an entirely new agency problem arises (Bostrom, 2013). Researchers face the AI control problem when they attempt to ensure that the artificial agents they are creating will not harm the project's interests. This problem of superintelligent agency poses an unprecedented challenge and will require new techniques to solve (Soares, 2016).

Two agency problems

The first principal-agent problem

- Human vs. Human (Sponsor → Developer)
- Occurs mainly in developmental phase
- Standard management techniques apply

The second principal-agent problem ("the control problem")

- Human vs. Superintelligence (Project → System)
- Occurs mainly in operational (and bootstrap) phase

Figure 1. Two agency problems as described by Bostrom (2013)

Solutions to the AI agency problem can general be divided into two categories: *capability control methods*, which limit the tasks a superintelligence can perform in order to limit its adverse effects, and *motivation selection methods*, which attempt to align what the AI "wants" to do with what its human creators want to do.

Capability control

Capability control methods are unsatisfactory for two main reasons. First, capability control methods like boxing (physical and informational containment of AI) are ultimately not stringent enough to protect against a superintelligent AI (Bostrom, 2013). While boxing can serve as an auxiliary safety measure for superintelligence, AI must interact with the world to some extent in order to be useful. Additionally, eliminating small interactions between AI, developers, and outside information would be almost impossible, since the AI still needs to be created and observed, leading to some level of interaction with the developers and observers (Bostrom, 2013). Second, more stringent capability controls like stunting (restriction of information, memory capacity, or computational hardware speed), destroys the utility of AI systems. Too many controls on AI would make it useless and limit the enormous benefits that safe AI could have on the development of science and technology (Danaher, 2014).

Therefore, we must turn to motivation selection methods to find a satisfactory solution to the AI control problem. Unfortunately, motivation selection presents many problems on its own. With past and current approaches running into intractable technical problems, many forms of motivation selection are over-simplistic or too heavily reliant on human intervention (Grace, 2017; Danaher, 2014).

	Description	Evaluation
Boxing Methods	Contain the AI within a "box", i.e. limit its sensors and actuators. Can use physical or informational forms of containment.	Simple but may have subtle vulnerabilities we are not aware of; human gatekeepers could be manipulated to let it out of the box.
Incentive Methods	Create an incentive environment that forces the AI to 'play nice', for example programme the AI so that its final value is the receipt of cryptographic reward tokens.	Interesting, but the AI might mistrust human reward givers; we may not be able to tell whether the AI is, in fact, playing nice; and there are other esoteric considerations to factor in, e.g. the AI's uncertainty about the simulation hypothesis.
Stunting	Force the AI to run on inferior hardware, or to focus on particular datasets.	Difficult to get the balance right: too much stunting reduces utility, not enough makes it possible for the AI to pose a threat; also difficult to reduce the knowledge of a sufficiently intelligent AI.
Tripwiring	Use "tripwires" to shut down or destroy an AI development project when it gets too dangerous.	Might have some use in the development phase, particularly if paired with other methods (e.g. boxing), but an advanced AI could subvert the tripwires, and human developers may grow impatient.

Table 1. Summary of capability control methods

Motivation selection

The most straightforward approach to motivation selection is rule-based direct specification. The most famous example is Isaac Asimov's "three laws of robotics" concept. The laws state that a robot may not injure a human being, a robot must obey human's orders, and a robot must protect itself, in that order (Asimov, 1976). Most AI theorists agree that Asimov's three laws of robotics form an unsatisfactory basis for machine ethics and development of safe AI (Anderson, 2007). Invariably, and as demonstrated in so many of Asimov's novels, the

imperfections, loopholes, and ambiguities contained in these laws often resulted in psychotic and harmful robot behavior. These rules are at once too rigid and too vague to define and account for every situation a robot is likely to encounter (Dvorsky, 2014). Furthermore, a robot or AI endowed with super-human intelligence would most likely be able to access and revise its core programming, rendering the laws useless (Anderson, 2007).

Since capability controls and rule-based motivation selection are not sufficient to solve the control problem, some researchers suggest that we augment AI with motivation systems that are already acceptable to us. Acceptable motivation systems are mostly human based, like brain emulation or human computer interfaces (Bostrom, 2013). However, even this method of motivation selection proves untenable under further examination.

To begin, humans do not understand their own motivation systems; therefore, augmenting AI with poorly understood systems may prove problematic, especially when the cognitive capacities of AI surpass human levels. We might think that obedience to humans would be a good property for AI to have, but humans are so imperfect in our own decision making that we may give orders that do not actually benefit us. Experiments have shown that a robot's performance degrades when the robot strictly obeys human orders, concluding that robots should be designed to disobey humans whenever we make irrational choices (Milli, 2017). Therefore, we cannot control AI through sheer obedience to humans without giving up their functionality and usefulness.

Alternatively, AI can be controlled by having a trained human supervisor "in the loop" during the certain phases of reinforcement learning, which has been shown to be more successful than pure programmed obedience to human operators. When the decision the AI had to make

was simple, human supervised learning prevented catastrophes without affecting the agent's learning. However, the scheme was unsuccessful for more complex decision-making processes. Extrapolating to more challenging environments, the amount of human labor required to augment reinforcement learning would be completely infeasible (Saunders, 2017).

Value Alignment as a Solution to the Control Problem

Thus far, the state of the control problem looks bleak. Capability control methods, rule-based motivation selection, and human augmentation are not sufficient to control super intelligent AI and prevent an existential risk to humanity. Yet we still need to construct AI systems that not only share our goals, but also share our intentions. Aligning the intentions, or values, of humans and machines is a complex, vague, fuzzy, context-dependent task known as the value alignment problem (Soares, 2016). This is a sophisticated problem, but I propose that our first steps toward solving it should be focused on natural language processing, (NLP) as a way to decode human values and recode them into computers. Language and stories serve as our human way of communicating values, from fables to fairy tales. Just as we learn selflessness from Cinderella, bravery from Robin Hood, and patience from the Tortoise, so can computers, if we master the challenges of NLP.

NLP serves two purposes: first, it can help us to determine what our human values are, and which we should try to transfer to computers. Second, it can re-encode those values into the control algorithms of machines in a more robust way than the rigid laws of Asimov.

Value Alignment through Natural Language Processing (NLP)

Links between language, consciousness, and human values

In philosophy and psychology, language is often cited as one of the most influential factors in creating human consciousness (Kelly, 2010; Jaynes, 2000). Words, linguistics, and stories hold the key to unlocking a deeper understanding of what makes us human, and therefore hold the key to solving the value alignment problem of artificial intelligence.

There are many links between language, information processing, and the development of technology (Kelly, 2010). In an information integration theory of consciousness, language is the key to higher order consciousness and the formation of a sense of self. Language binds our distributed brain processes together to form a unified experience (Toroni, 2004). Therefore, language serves as a tool to integrate information from many parts of our brains, providing us with an intelligence that sets us apart from animals. Not only does language create consciousness, but it also organizes information and provides us with a way to communicate a common understanding of the world around us (Blackmore, 2011).

Other philosophers agree with this characterization of human consciousness as a product of language. For example, Julian Jaynes viewed narratization as a crucial aspect of consciousness, citing our ability to create and understand metaphors and stories as a way to extend our conscious model of the world (Jaynes, 2000). Not only does our consciousness depend on the words that we use, but also it also crucially relies on our ability to form coherent narratives of our experience.

Language is important not only in the creation of human consciousness, but also in the development of technology. In his essay *History of the Seventh Kingdom,* Kevin Kelly argues that the evolution of technology mirrors biological evolution, following the same trends of increasing complexity and ability to share information concisely. He defines life as a self-generated information system and identifies the major advances in biological and technological evolution. See the table below for a summary of the transitions identified by Kelly. Notably, language is the only transition present in the history of both biological and technological evolution (Kelly, 2010).

Major transitions in biological evolution	**Major transitions in technological evolution**
One molecule → interacting molecules	Primate communication → Language
Replicating molecules → Chromosomes	Oral lore → writing / mathematical notation
Chromosome of RNA → DNA proteins	Scripts → Printing
Cells without nuclei → Nucleated cells	Book knowledge → scientific method
Asexual reproduction → Sexual recombination	Artisan production → Mass production
Single-cell organism → Multicell organism	Industrial culture → Ubiquitous global communication
Solitary individual → Colonies	
Primate societies → Language-based societies	

Table 2. Major transitions in biological and technological evolution

Kelly claims that the invention of language that allowed us to generate and organize information in a more meaningful way, saying:

> "No transition in technology has affected our species, or the world at large, more than the first one, the creation of language. Language enabled information to be

> stored in a memory greater than an individual's recall….From a systems point of view, language enabled humans to adapt and transmit learning faster than genes." Kelly, pg. 47

Language was the last major biological development, the crucial step in human evolution that set us apart from non-human creatures. Human uniqueness, intelligence, and success exists only because language developed. In addition to being the last major biological development, the invention of language was the first transition in the evolution of technology (Kelly, 2010).

Thus, language serves as a bridge between biological evolution and technological development. Not only does language allow us to communicate more meaningfully with each other, but also it serves as our best tool to communicate meaningfully with the information ecosystem of technology. This distinction is crucial, because it defines the humanist source of human exceptionalism as our ability to process information and pinpoints language as the source of that ability. Therefore, non-human machines could theoretically gain the unique capabilities that humans currently possess if they could mimic our language processing abilities.

Language and narrative play a crucial role in the formation of our own consciousness, values and beliefs (Jaynes, 2000; Kite, 2018; Mueller, 2003). Therefore, computational language processing is a powerful tool for the alignment of human and machine values. Kurt Vonnegut, author and anthropologist, understood this as early as 1947, when he proposed that the "shapes of stories" are universal, and could be fed through computers (Vonnegut, 1947). Vonnegut's thesis was rejected because computers did not yet possess adequate processing power to comprehend natural language. In recent years, advances in computer hardware have made it possible to study texts through the lens of big data and AI, with important results.

Modeling emotion computationally with natural language processing

Inspired by Vonnegut's rejected thesis, researchers from the University of Vermont's computational story lab attempted to classify the emotional arcs, or "shapes," of 1,327 stories from Project Gutenberg's fiction collection. In this way, they examine the emotional component of narratives, separate from other narratological elements like plot, character networks, or narrator voice. The study utilizes sentiment scores and machine learning techniques to first linearize the emotion of a text, then group stories into similar emotional "shapes." Through a combination of supervised and unsupervised machine learning, they found that the emotional arcs of stories fall into six basic categories (Reagan, 2016).

This research has obvious implications for the field of digital humanities and enrichment of literary theory, but also possesses huge insights for AI strategists. If NLP algorithms can uncover the emotional arcs of stories, they may have the potential to uncover human values, the first step in solving the long-term AI value alignment problem. Additionally, the results of this study only analyze one component of human narrative – emotion. A vast body of work in the field of the digital humanities explores other aspects of narrative, such as plot, character networks, chronological event ordering, and narrator voice (Reagan, 2016; Riedl, 2016). Such work aims to find a way to instill AI with computational narrative intelligence – the ability to construct, tell, understand, and respond affectively to stories. Although there are still technical challenges to overcome, instilling AI with computational narrative intelligence affords a practical way of machine enculturation and value alignment (Riedl and Harrison, 2016).

The exploration of NLP and computational narrative intelligence provide two related paths for future research. The first approach focuses on uncovering human values through NLP

(Reagan et al., 2016). The second approach focuses on building machines that can understand human values using computational narrative intelligence (Riedl and Harrison, 2016).

Firstly, NLP provides an answer to one of the questions posed in the AI Strategy research landscape document: “What will we need to know and arrange in order to elicit and integrate people’s values?” To answer questions of technical AI development and policy growth, we need to know what values humanity would want an AI governance system to pursue. Language processing algorithms can provide a way to uncover these values. If we can use machine learning to uncover the emotional arcs of stories, we can use NLP to analyze the vast corpus of human stories and uncover our own most important values. Not only can machine learning uncover our own values, but also it can then feed those values back into the control structure of AI in a way that it can understand.

That brings us to the second line of future inquiry for NLP-based AI safety: installing machines with a way to understand and enculturate themselves with human values. A promising area of research is the development of computational narrative intelligence, the ability to understand stories and language (Mani, 2013). Machine enculturation may be feasible via machine learning over a corpus of stories. Studies have shown that machines can emulate human behavior with simple, crowdsourced narratives (Li et al., 2013). This area of research presents several technical challenges. Current NLP algorithms have great difficulty decoding the meaning of metaphor in natural language, a process that requires a high-level semantic comprehension that escapes most computers. Additionally, commonsense knowledge of the world is required for narrative understanding and narrative generation. Learning such commonsense knowledge has been an ongoing challenge in AI and machine learning (Riedl, 2016).

Although this area of research contains immense potential for the development of AI safety and the eventual resolution of the control problem, NLP and computational narratology has typically been studied in the realm of the digital humanities, applied to literary theory and chatbots instead of policy proposals. Of course, this area of work presents challenges: it crosses many intersectional areas of study and presents complex technical problems. However, I truly believe that an exploration into NLP and narratology presents our best option for a potential solution to the AI control problem: NLP has the potential to first determine what human values are, and then possesses the ability to encode that into the control algorithms of machines in a robust way (Riedl and Harrison, 2016; Mani, 2013).

Chapter 3

Abstract

In the previous chapter, I outlined why the AI value alignment problem must be solved, and proposed that natural language (novels, stories, and other texts) provide the most promising data on human values. In this chapter, I outline the field of computational narratology, starting with its roots in classic literary theory, and closing with a detailed account of current methods of modeling stories and generating narrative. I conclude by proposing how this large body of research can be applied to solving the value alignment problem: by creating an audience model. There are many algorithms that model authorial intent and story structure, and a few that attempt to create an audience model. Such a model would have to include audience emotions, responses, and values. By modeling audience values, computational narratology algorithms can create a rudimentary structure for value alignment.

Introduction to Narratology

Telling stories is a central part of what it means to be a human being (Mani, 2013; Stern, 2005). Humans have been telling stories for a long time; the 17,000-year-old cave paintings at Lascaux, for example, tell the tales of animals using familiar narrative techniques like chronological sequencing (Mani, 2013). Not only do language and narrative serve as the bridge between biological and social development, as discussed in the previous chapter, but also they provide us with a framework for how we cognitively approach the world (Mateas, 1999).

Because of the central role of storytelling in the human experience, it has long been the subject of study across disciplines such as philosophy, psychology, anthropology, and linguistics.

Although stories have been studied throughout history, formal narratology stems from the language- and linguistics-centered approach of the early 20th century structuralists. Narratology itself is defined as a humanities discipline dedicated to the study of narrative representations, and the logic, principles, and practices that guide it (Meister, 2012). Structuralists such as Ferdinand de Saussure applied the theories of semiotics – the systems of signs that create language – to create classical narrative theory. Semiotics is the study of signs, which are defined as being composed of: a 'signifier' - the form which the sign takes; and the 'signified'- the concept it represents (Saussure, 1916). For example, the three-letter word “cow,” is a signifier, and the moo-ing, spotted creature we see in the fields of Pennsylvania is the signified.

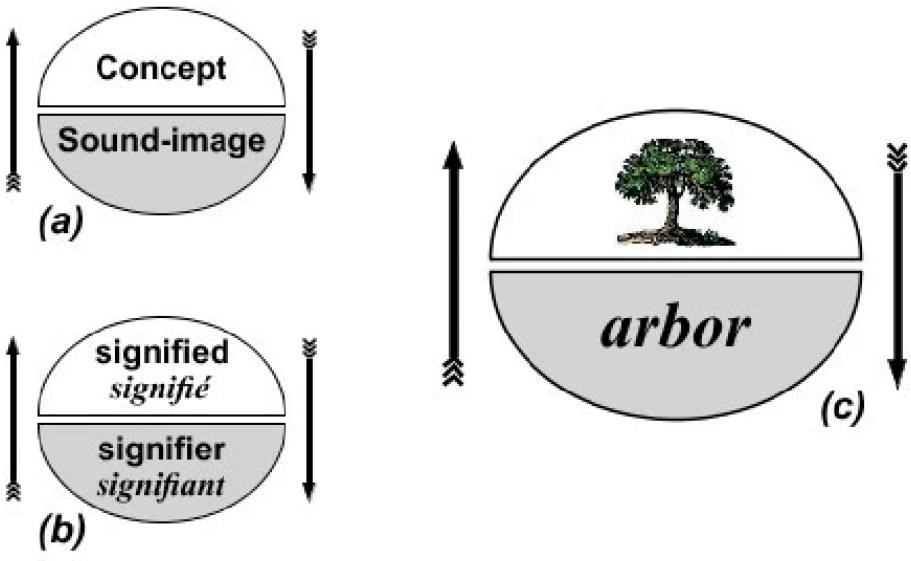

Figure 2. Visualization of De Saussure’s signifier and signified

The remarkable conclusion of this theory is that signifiers have no real relation to the things they signify; that is, the letters “c o w” have no actual relation to the animal. Within narrative theory, this distinction drew a line between the actual content of language and the various ways we can interpret it (Saussure, 1916). To apply De Saussure’s example to narrative, the word “cow” in a story can conjure hundreds of different meanings, emotions, and images depending on its context within the narrative and on the reader’s background. “She was such an unkind old cow” creates an entirely different mental image than “the small cow in the field.” The field of narratology studies how written and spoken stories work to transform the content of the story (the signifiers) to the coherent images, lessons, and emotions imparted when the story is told (the signifieds) (McCune, 1995). This line between the content and the interpretation of a story deepened as later theorists created more advanced models of narrative structure, representations of stories that span entire texts instead of just individual words.

When speaking of narrative structure, the primary structural analysis is of the forms found in text, where text is a sequence of words and characters (like “cow”) in a written work. However, the focus on text is insufficient for developing a complete theory of narrative, since the signifiers of the text can represent multiple signifieds based on the reader’s interpretation (Saussure, 1916). Furthermore, in the course of storytelling, more complex narrative structures arise where agents interact with each other, involving interleaved processes of generation and understanding, which in turn originate from environments where agents interact with each other. This lead to the traditional narratological distinction between the underlying content of a narrative and its expression/interpretation (Mani, 2013). Story is the content of a narrative, the raw material that makes it up, such as the events, the characters, notable items, and setting.

Discourse is the narrative's expression, the means by which the content of the story is communicated. It is within discourse that the unremarkable and easily definable building blocks of a story are transformed into a tale capable of capturing our attention and affecting our emotions – where the everyday signifiers become meaningful concepts and stories (Chatman 1980). This distinction can be found again and again in the history of narratology: *histoire* versus *discours* in the French structuralist tradition and *fabula* versus *sjuzhet* in the Russian formalism that preceded it, where *fabula* is the "raw materials of the story" and *sjuzhet* is "the narrative as told or written with the procedures, emphases, and thematic devices of the literary text" (Martin, 1986).

Those who have attempted to develop narrative models have worked within these distinctions between fabula and discourse. Generally, models of plot, event series, character networks, cause and effect relations, and time within narrative are by products of a fabula-based textual analysis, while the discourse is considered the final output of such models. However, in narrative modelling, no clear line can be drawn between fabula and discourse, since many of the structural aspects of narrative that need to be modeled for computational purposes can be represented at the discourse level, whereas others necessarily relate to the fabula (Mani, 2013).

Early narrative models

In the past, many scholars in the field of literary narratology who have attempted to model story structure focus on the patterns formed by the fabula of stories. For example, Joseph Campbell's famous archetype of the "Hero's Journey" claims to describe the mythic structure of all epic tales using the familiar pattern of events and characters that appear in them, like the

hero's departure from home or her encounter of a mentor (Campbell, 1990). Kurt Vonnegut also theorized that all stories have basic structures, or "shapes," but he posited that their shapes were defined not solely by the events that occurred in them, but by the positive or negative perception to the audience of those events (Vonnegut, 1947).

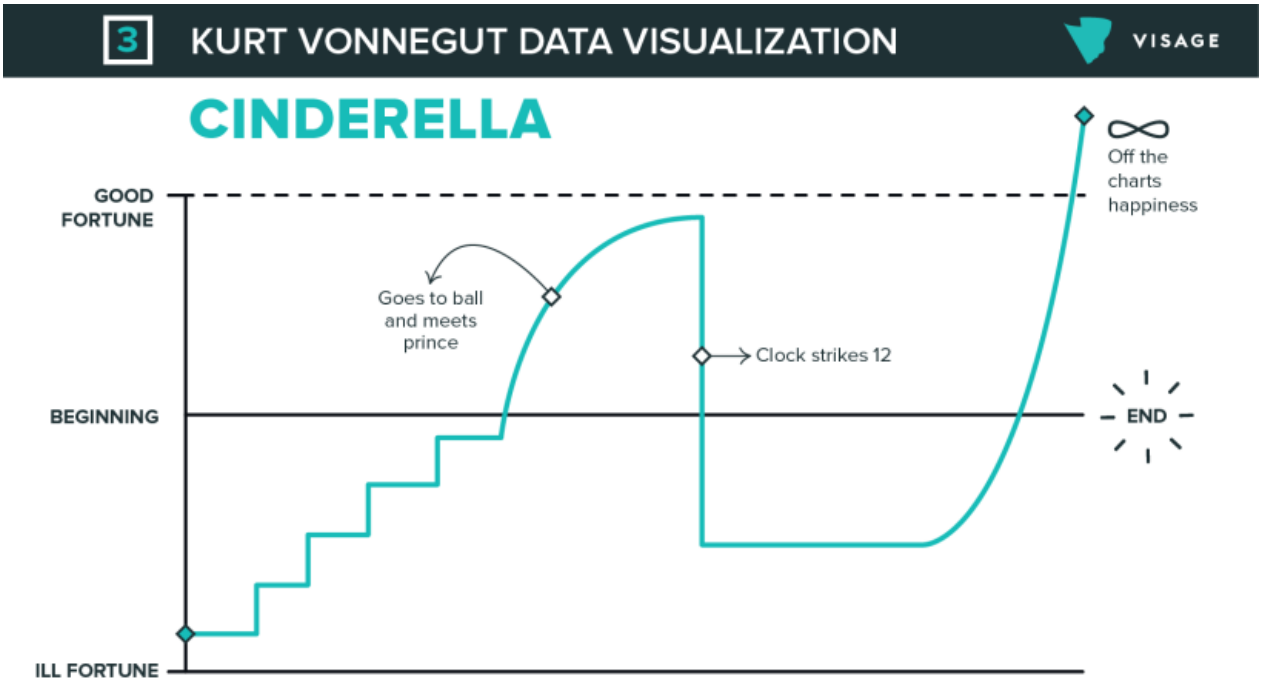

Figure 3. An example of one of Vonnegut's proposed shapes of stories

Katy French (2018)

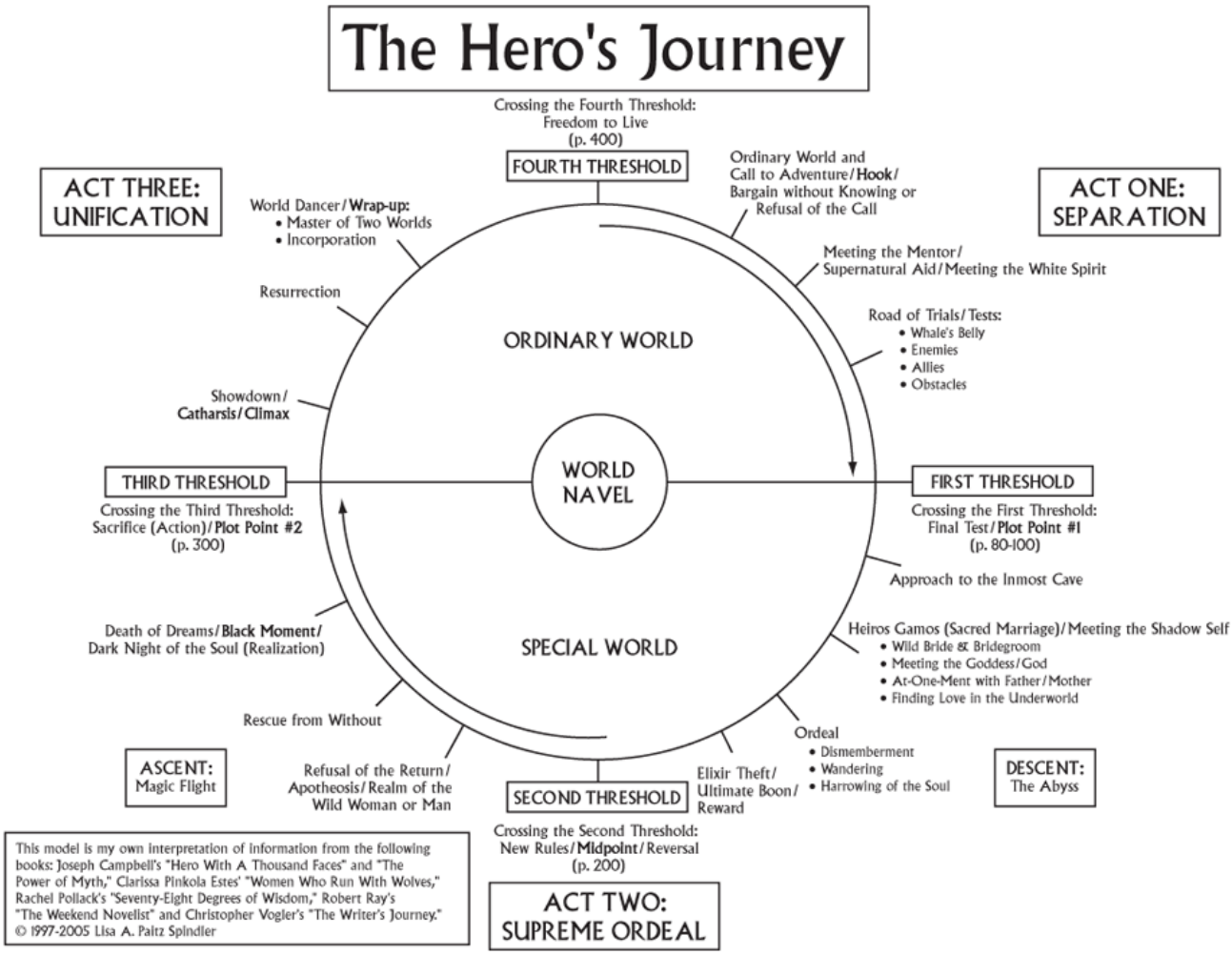

Figure 4. Illustration Campbell's monomyth

An interpretation by Lisa Paltz Spindler (2005)

Vonnegut, Propp, and Campbell's claims are not unfounded. There *are* certain stories that appear over and over throughout human culture. For example, the Cinderella story, which originated with the Chinese tale of "Yeh-Shen" in the ninth century before becoming famous in Western culture with Charles Perault's French retelling in the 1600s (Kite, 2018). There are over 900 versions of this story, in almost every culture with a written language (Kite, 2018). Classic explanations for this phenomenon were often unscientific and impossible to test empirically,

such as Carl Jung's theory of a collective unconscious. Jung postulated that the human collective unconscious is populated by instincts and archetypes, universal symbols like the Great Mother, the Wise Old Man, and the Tree of Life. With rich evidence from thousands of global myths, folktales, songs, and legends that exhibited the same symbols in the same patterns (i.e. the Cinderella story,) Jung reasoned that individuals live out the symbols of the collective unconscious and color them with meaning through their own experiences (McCune, 1995; Jung, 1936).

Of the many narrative theories developed in the humanities, the theory presented in the Morphology of the Russian folktale by Vladimir Propp, first published in Russian in 1927, is the most widely recognized. Propp analyzed a hundred stories from the anthology of Afanasiev's folktales and proposed an official common framework that explained how they were constructed. This framework is based on the adventures of a hero who confronts a villain to resolve the initial dilemma and go on to triumph, similar to Campbell's journey. Propp's work, like many other narrative theorists, created simple, clear structures of stories that could be verified with evidence from a large enough corpus. However, what theorists like Propp and Campbell lacked was the computational power to sift through larger corpuses of books to support their theories (Gervas, 2018). However, the narratological developments in the humanities that have occurred over the past half century have been paralleled by enormous growth in computational power. The ability to computational process huge corpuses of texts, coupled with the development of structuralists theories of fabula and discourse made it possible, for the first time in history, to empirically test narrative models. This resulted in the creation of a new field in narratology that focuses on

creation algorithmic systems for generating and understanding stories – computational narratology.

Computational Narrative Intelligence

The field of computational narratology is a natural extension of the classical literary aim to model narrative. Computational narratology (CN) is the study of narrative from the point of view of computation and information processing, focusing on the algorithmic processes involved in creating and interpreting narratives, modeling narrative structure in terms of formal, computable representations (Mani, 2013). CN is inherently interdisciplinary, with roots in the fields of literary analysis, linguistics, and computer science; therefore, the scope of CN research varies widely. Generally, three common themes emerge in the field:

1. Modelling narrative computationally using understanding and definitions from classical narratological concepts such as plot, character networks, and time
2. The augmentation of human understanding of literature through the exploration and testing of literary hypotheses through mining of narrative structure from corpora.
3. Story understanding and generation in artificial intelligence systems video games, and other forms of human computer interaction.

Modelling narrative computationally

Work on the analysis of narrative can be broadly classified in the following lines of work:

Narrative Information Extraction

A large field of work focuses on automatically extracting narrative information from text using natural language processing techniques (Chambers and Jurafsky, 2008). Instead of creating a model of narrative and then testing it on text corpuses, this approach uses the texts themselves to extract a narrative model. Most of this work focuses on extracting particular components of a narrative from a previously defined ontology. These include automatically identifying characters, narrative and plot structure, and character relationships (Chambers, 2008; Chaturvedi, et al, 2015). Applications range from those tangentially related to computational narrative (such as document indexing and retrieval) or study, summarization and visualization of stories (Coyne, 2010).

Story Understanding

Story understanding (a.k.a. machine reading or story comprehension) combines goes beyond simply extracting information from stories and strives to understand the entire story in order to reason and answer questions about it (Mueller, 2003). These efforts usually involve linking the extracted information to ontologies and common-sense databases (Matuszek, 2006). Unlike narrative information extraction, the field of story understanding begins with theories of classic narratology, then create systems modelled on those theories. The systems created are not based on the analysis of vast bodies of text like in narrative information extraction, but on meta-theories of narrative. However, this is problematic, since questions like “what is a story? What is narrative?” must be answered before computational models of stories can be made.

Many papers in the field of CN reexamine the traditional clear-cut definition of a story as “an account of imaginary or real people and events told for entertainment; an account of events”

(“story,” Merriam-Webster, 1999). Narrative is not a single entity or a single, tightly related set of concepts. Each field that informs CN (linguistics, literature, etc.) has its own, often different, definitions of what narrative is. For example, it can be a tightly woven story communicated by a strong authorial voice to an audience (Barrett, 1989). It can mean the internal imposition of coherence by which a person makes sense of her life, or the communally constructed group memory by means of which a group organizes past experience (Agre, 1997). It can be a set of mental spatial reference models created by a reader. In the broadest sense, narrative can mean an entire worldview (as in "grand" or "master" narrative). Thus narrative is a broad term for a related, yet richly varied set of ideas.

The richness of narrative presents an interesting challenge for the emerging field of CN: how can researchers maintain the complex, multifaceted nature of narrative while still reducing it to computationally feasible rules and patterns? AI, like the rest of computer science, tends to prefer general and abstract formulations (Barret, 1989; Agre, 1997). Applied to narrative, this will result in the attempt to assimilate all narrative phenomena to a single, simplified formulation. In order to build systems, abstraction and simplification are necessary tools. The danger lies in forgetting for what purpose a simplification was made or perhaps that a simplification has even occurred. With a concept as complex and evocative as narrative, there will be particularly strong pressure to elide simplification. If this were to happen, the original richness of narrative, an endless source of inspiration and delight, would be lost (Mani, 2013).

Some models of narrative are directly based on earlier, non-computational models presented by classic narratology scholars. For example, Grabson’s morphological approach to interactive narrative relies on the function of *fabula* and *sjuzhet* proposed by Russian formalist

Vladimir Propp, while Stern's creation of a computer-generated drama relied on an AI system based on Campbell's Hero's Journey (Grabson, 2001; Stern, 2005). In addition to building of traditional narratological theories, computational narratology has also developed its own accounts of key narratological concepts. For example, characterizing events based on the motivational actions of the characters lead to a more computationally viable, fine-grained model of plot (Lehnert, 1981). Although early systems used such models of plot in story generation, understanding and imputing motives requires a level of inference that extant computational systems did not possess. However, in 2010 researchers from Yahoo! Research and the University of British Colombia used a corpus of literature to develop a text understanding system capable of inferring characters' emotions (or affect states) associated with events, identifying which outcomes are beneficial, harmful, or neutral for particular characters (Goyal, 2010).

More nuanced models of characters' emotions have also been explored. For example, the interactive storytelling system of Pizzi (2011) relies on an inventory of character's feelings that was developed by Flaubert in his preliminary studies for Madame Bovary (see figure 5). Such a framework uses an emotional planner to drive character behaviors and allows for a variety of sentiment-driven interactive retellings of the novel. Another interesting reformulation of a narratological construct is that of suspense. Cheong (2007) generates stories judged to be suspenseful by modeling the reader's reasoning about limitations and conflicts involving a protagonist's goals, based on narratological insights from classic theorists. This system is based on the concepts that a reader's suspense level is affected by the number of possible paths the narrative could take and that story structure itself can influence the reader's comprehension of

narrative (see figure 6). These theories are both important because they attempt to create a model of the reader, instead of modelling authorial intent or story structure alone.

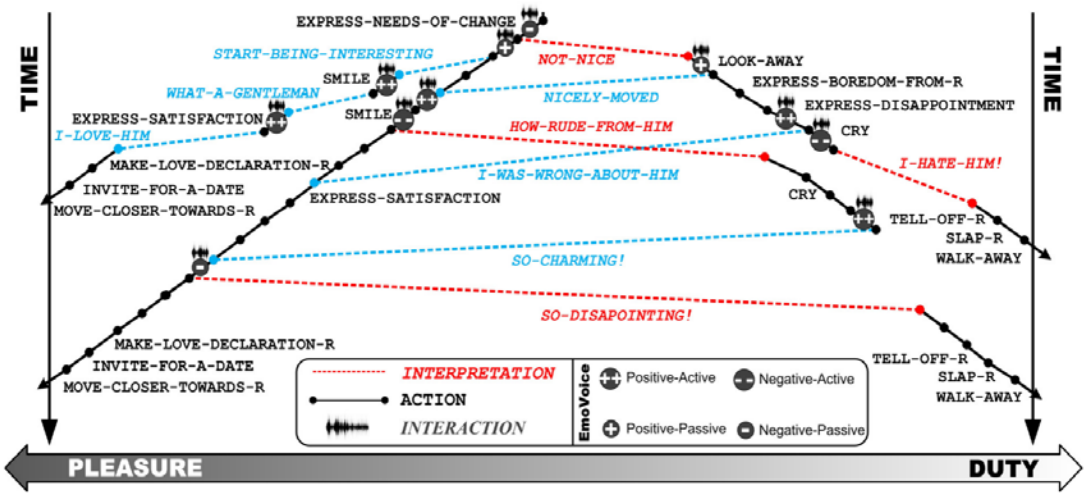

Figure 5. Story Variability and Impact of Emotional Speech

The diagram represents different evolutions depending on the emotional categories recognized by Pizzi (2011). Multiple opportunities for interaction leverage the impact of Emotional Speech, and account for significant variability despite the limited number of emotional categories.

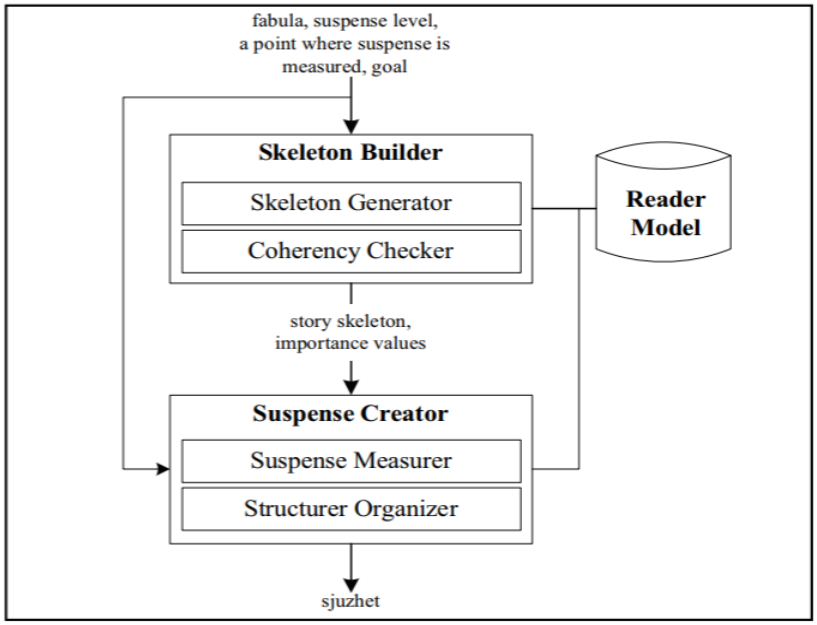

Figure 6. The Suspenser Architecture

This AI system uses fabula as its input to output the discourse, or sjuzhet.

Much work has been done on developing computational equivalents to classic narratological theories (Grabson and Braun, 2001; Mani, 2013). The theories discussed above go a long way towards the creation of a story understanding system. However, ultimately, the question "what is narrative?" cannot be answered in clear cut computational terms. The analysis of narrative with computational methods face many shortcomings. Therefore, to utilize CN technology, researchers must make do with imperfect models.

Supporting human narrative intelligence

Modeling narrative completely requires an understanding of too many variables to be computationally viable. Therefore, the next main subfield of CN deals with the creation of imperfect models, designed for use in supporting human narrative intelligence. An excellent example is the Nora natural language processing tool that performs text classification tasks with either Naïve Bayes or Scalable Vector Machines (SVM), two unsupervised clustering algorithms that process the text based on sentiment analysis (Plaisant et al., 2006). The user chooses from one of three text collections (non-fiction materials from Documenting the American South, several hundred poems and letters by Emily Dickinson, or a small set of sentimental novels), and can configure the classification task to test some hypothesis; for example, the user might be interested in the characteristics of erotic language in Dickinson's poetry (a well-turned question in the scholarship). The goal of Nora is not to be able to read and analyze literary text as well as a human, but to provide specific computational tools that can supplement human scholars' interpretations of text (Plaisant et al., 2006). By classifying large volumes of text using machine learning techniques, Nora can provide insights that might otherwise escape notice. Scholars

know that while reading and rereading text is an essential step in literary analysis, human cognition and language processing tend to miss subtle meanings or fail to connect larger patterns, tasks that machines, with greater processing power, can accomplish. How effective is a system like Nora amongst practicing literary critics? This long quotation captures one scholar's experience, someone who has spent a career as an Emily Dickinson authority:

> "the data mining "search and find the erotic" exercise made me put together the blending of the erotic with the domestic. And thus, I was surprised again because I've written extensively on the blending of the erotic with the domestic, of the familial with the erotic, and so forth…. So, the data mining has made me plumb much more deeply into little four- and five-letter words, the function of which I thought I was already sure, and has also enabled me to expand and deepen some critical connections I've been making for the last 20 years" (Smith, 2005).

There are many limitations to computational narrative models, since there are many aspects of narrative, including the modeling of style, subtle lexical connotations, metaphor, humor, and irony that narratology does not concern itself with (Mani, 6). However, the method of "reading" texts provided by data mining techniques has its roots in long-standing practices within the humanities. CN offers a non-traditional way of approaching literary criticism and machine learning and provides a way to augment human understanding of narrative (Kirschenbaum, 2009).

Interactive story design

The two main areas in the field of CN, narrative analysis and narrative generation, both address key conceptual problems discussed in narratology. However, story generation and story understanding algorithms use different techniques in practice. Story understanding (a.k.a. machine reading) has been discussed previously in the section on narrative analysis and narrative information extraction. Computational models for narrative generation use story understanding systems as their input, with the further goal of outputting some coherent and engaging computer-generated narrative.

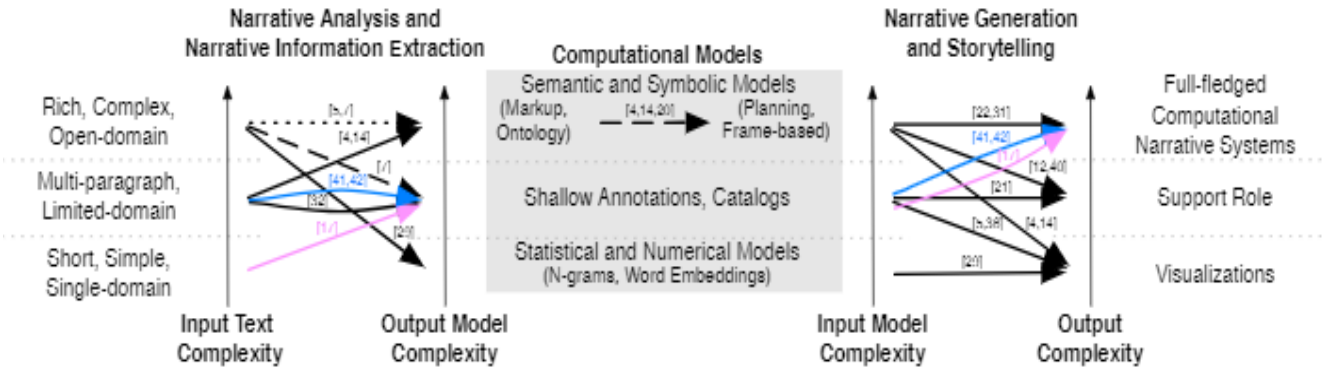

Figure 7. Overview of input and output in computational narrative applications in the main two areas of research within computational narrative

Black arrows indicate existing approaches (solid arrows indicate automated approaches; dashed arrows indicate manual and semi-automatic approaches). The pink and blue arrows illustrate the related work by Vargas et al (2017) for text-based end-to-end computational narrative systems.

Artificial Intelligence uses two techniques in Story Generation: planning/problem solving, and production grammars. Specific rules used in their algorithms might be influenced by insights from literary studies or other fields (e.g. psychology of reading and writing).

The most popular AI technique used to generate stories is AI planning – a technique that constructs plans to achieve a desired conclusion given a set starting point. The path the plan takes is built as a sequence of linked actions that lead from the initial set up to the desired outcome. When the actions in this sequence are interpreted as events and the sequence as a narrative thread, these plans constitute a good approximation of what is expected of a story (Sanghrajka, 2018). By being solutions to a planning problem, all the events in the resulting narrative are, by constitution, linked by cause and effect, which provides coherence and thematic unity. These models include logic and plan-like models representing a story space or the rules defining a simulation, narrative theory, agent behavior or author's goals (Vargas, 2017). Plans obtained in this way are very lineal and have little variation, creating only simple and monotonous stories (Gervas, 2018). Because of the specialization of the different systems, planning systems tend to be ad-hoc, manually authored, simplistic, and not robust – they cannot be reused across genres or texts.

Another common method of constructing story generators is to try to formalize knowledge gathered in the fields of literary studies and classic narratology into a working production grammar. Story generation algorithms tend to be overly simplistic, usually utilizing only one guiding heuristic from narratology, while in reality, dozens of narratological theories are needed to understand the intentions of a human author. Many story generators utilize only the notion of *fabula* in creating stories. While they are capable of linking plot events together in a coherent time sequence, they lack *sjuzhet* - the tone, emphasis, and other methods of delivering the story.

Bailey (1999) further distinguishes between three different approaches to automated story generation:

1. **Author models:** These models attempt to imitate the process of a human author writing a story. Such models rely heavily on logic-based systems.
2. **Story models:** These are based on an abstract representation of the story as a structural or linguistic object. Systems based on story grammars, such as Propp's morphology or Campbell's journey, fall under this category
3. **World models:** In these models, generating a story is seen as constructing a world governed by realistic rules and peopled with characters with goals. The rules of the world must be programmed into the story generation system, and the story is created as the character agents attempt to achieve their goals while constrained by the rules of the world system. An example is Tale-Spin (Meehan, 1977), the classic story generator inspired by Aesop's fables.

Authorial intent

Story generation systems, like human authors, must evaluate which narrative trajectories belong in their story. Computer systems could have entirely human-generated narrative paths, of which it simply selects different options, or it could have the freedom to generate and select its own narrative paths. Branching story graphs are examples of logic-based planning systems that give all authorial intent to the human creators of the system. In these graphs, nodes represent chunks of authored narrative content and directed arcs represent explicit choices users can make (see figure 8). Every possible narrative trajectory is manually authored, preserving complete and

immutable authorial intent (Riedl and Bulitko, 2013). An excellent example of this type of story is a choose your own adventure book. In these stories, human authors must write out every possible ending to the story, with the user (or computer) simply choosing which narrative path to take between nodes (Mani, 2013). AI path finding heuristics simply plan a final narrative from a branching story graph without authoring content. However, the amount of narrative content that must be authored can grow exponentially with the number of choice points and the authoring of large graphs with many opportunities for user agency quickly becomes intractable (Mateas and Stern, 2005).

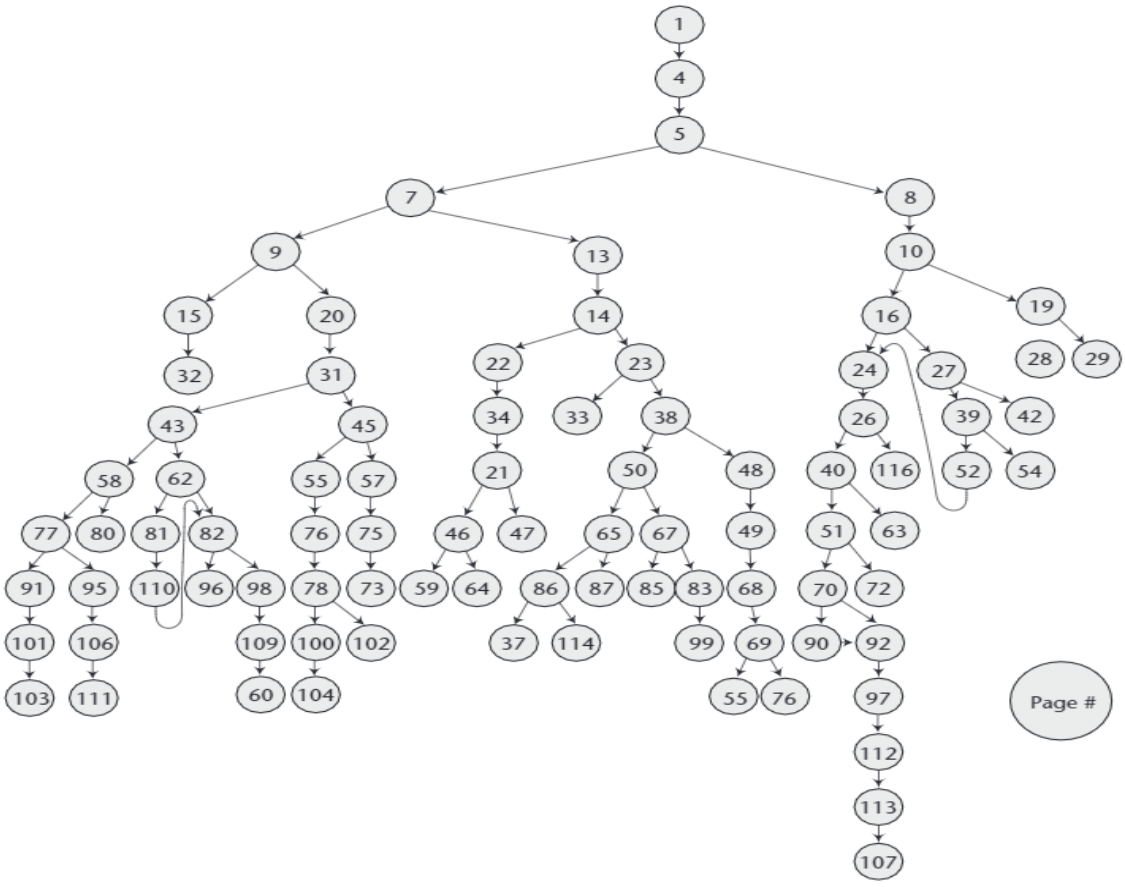

Figure 8. A branching story graph from "The Abominable Snowman"

In reality, human authors of narrative must express his or her authorial intent through some means other than enumerating all possible plots, then choosing one. A common approach to encoding more nuanced authorial intent is as a plot graph in which nodes are narrative events and arcs denote precedence constraints such that no event can occur unless all events constrained to occur prior to it have already occurred (see figure 9). A plot graph thus defines a space of possible narrative trajectories; constraints prune out sequences that do not logically make sense or should not be allowed to occur in the user's narrative experience. A search process — adversarial search, reinforcement learning, or case-based planning — generates possible trajectories and evaluates the trajectories according to an author-defined heuristic (Riedl and Bulitko, 2012).

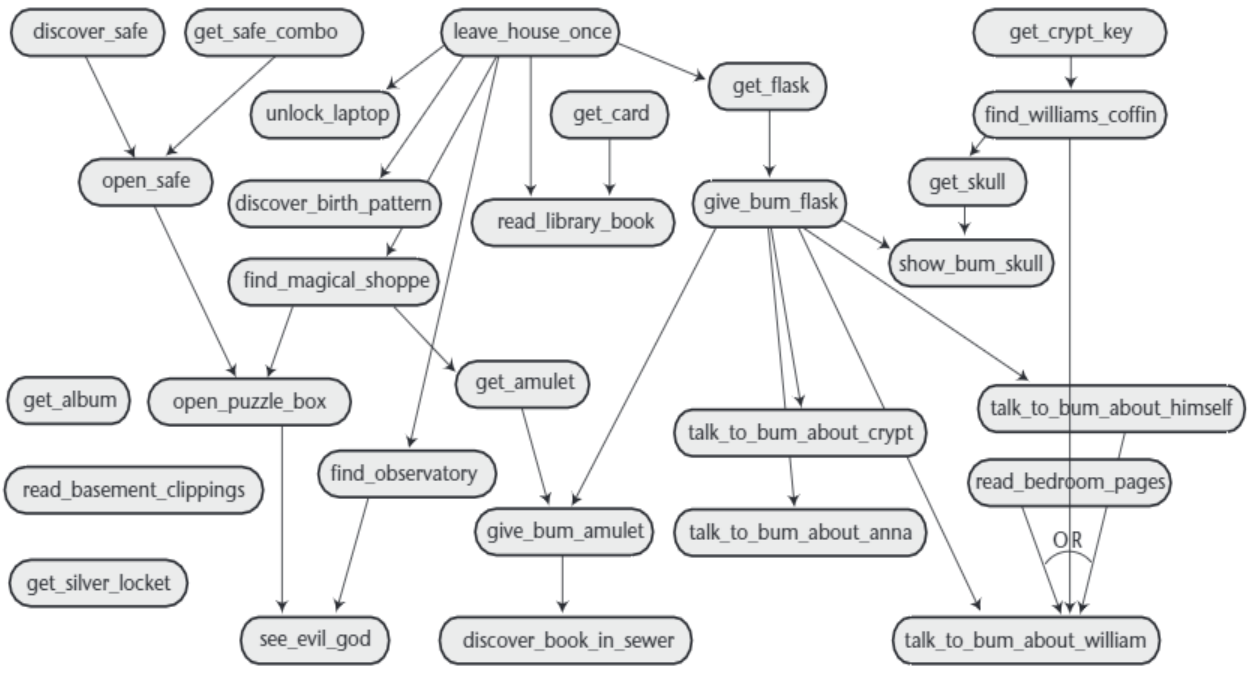

Figure 9. A portion of the plot graph for Nelson and Mateas' interactive fiction system (2005).

Story models

Many different structures of narrative have been discussed already, including the theories of Propp and Campbell. Story grammars are driven by narrative functions expressed via rewrite rules (Mani, 2013). However, such story grammars are difficult to model computationally because they make abstract claims that often cannot be formalized (Mani, 2013; Lehnert, 1981). In his evaluation of general story grammars, Black (1979) examined the form of the grammar rules and evaluated them with results from mathematical linguistics. He found that only one type of story grammar was formally and empirically adequate; however, even this fully adequate transformational story grammar would not aid human understanding, since it was unintuitive to most human readers. Black (1979) concluded that grammars would add nothing to semantic computational models that focus on story content. Mani (2013) lists several more disadvantages of story grammars, including their inherent brittleness and their failure to decouple the order of events specified in the grammar from their discourse order. For example, in a story such as the *Odyssey* in which the order of events in the discourse does not match the order of events in the text, story grammars would not suffice. Both Mani and Black conclude that structural story grammars do not have any potential to further the field of CN. While story grammar theories contribute insights and provide a basis for good hypotheses in the field of CN, focusing on them is unlikely to yield further progress. A more productive research direction, according to Black (1979) is "exploring the kinds of knowledge needed to understand story content."

The authorial intent and story grammar models used in the field of interactive story design highlight the overall shortcomings of computational narrative intelligence. In general,

representations of stories in AI systems present a simplistic approach to narrative in several senses.

Linear vs branching stories. Representation of stories in AI systems, particularly by story grammars, consider the events in the text as linear. This is problematic, primarily, because events as they occur in a text sequence often do not match the temporal sequence of such events, such as in the *Odyssey* (Mani, 2013). Secondly, stories more complex than the simplest children's book have several branching stories whenever more than one character is doing something relevant to the story in different places at the same time. Stories rarely follow the actions of a single character in a linear sequence without involving the actions of another character, a flashback, or other such complicating event.

Role of Causality. AI systems model plots in terms of a graph of causal links between the motivations of character agents and the mental states of these characters after an external event. These causal links can involve a single character or include cross-character interactions (Mani, 2017). However, this planning paradigm is biased towards producing plans in the shape of an inverted tree: a number of branches (causes) all converge toward a final goal (the result). Real stories rarely have a single end point where the goal of the story can be said to be achieved (Vargas, 2017).

Modelling the Reader. Bailey's (1999) approach is based on the idea that something is a story if and only if some reader identifies it as such when being exposed to it. This defines a story only in terms of a particular reader, but Bailey tries to abstract a general description of what makes all readers recognize something as a story. This requires having some way of modelling

and/or measuring the reader's reaction to a story. As Bailey himself confesses in his paper, there is still a gap between existing work on this topic from the point of view of AI and the Humanities, in the sense that there is a large body of literature on the influence of narrative on the reader that has not been applied to AI research (Vargas, 2017).

Challenges in the field of CN

As shown by the shortcoming of story generation systems, representations of stories in AI systems present a simplistic approach to narrative. While linguistic models from classic narratology are under defined, vague, and predominantly descriptive, AI models are based on a highly reductionist concept of a story which ignores the humanities insights in the field of narrative (Gervas, 2006).

Areas of future research – the creation of a user model

A gap exists between the underdefined narrative models of the humanities and the over defined, simplistic models of AI systems. As Bailey (1999) recognized, a story model must exist within relation to the reader. All narratives, by their very nature, are consumed by some audience. In some narratives, the audience is referred to explicitly, as in the first line of *The Catcher in the Rye*, while in others the audience may be created by the narrative, as in the case of Shahryar being the audience for Scheherazade's narratives (Mani, 2013). The audience brings beliefs and knowledge to the consumption of a narrative that allow them to understand the author's references. This audience background knowledge is one of the hardest aspects of

narrative understanding to replicate in AI systems. To use E.M. Forster's classic example, when we say "the king died and the queen died of grief," any human reader understands why the queen died – we understand the concepts of marriage, relationships, and love. However, an AI system reading that sentence would not know why the queen died of grief unless it had hundreds of underlying concepts encoded into it (Plot and Story, 2017). The audience of a story is the most sophisticated story understanding system, so understanding audience response is key to creating AI that understand stories. More important than modelling the structure of the story, or the authorial intent, is actually modelling the audience itself. Several facets of reader affect (aspects of narrative that make it interesting to the audience) can be studied both formally and computationally, such as models of suspense, plot, and characterization that have already been discussed. However, other facets of reader affect are harder to model given the many different parameters of a discourse responsible for audience impressions (Mani, 2013). One of these facets is that of meta-narrative: the emotional coherence and resolution of the overarching arc of a story.

Conclusion

The field of narratology has created many models of stories and narrative. Computational models of narrative have largely imported these narratological theories, focusing on the theories with a formal or logical structure that could easily be turned into a computer algorithm. However, computational models are overly simplistic models of narrative. By focusing on story grammars, authorial intent, and other formal narratological constructs, CN researchers have failed to develop a robust model of a narrative audience. To create an AI system that can both

understand and generate stories, such a user model must be explored. Since few such models exist in classic narratology, a user model can be created by analyzing a large corpus of text with natural language processing techniques. In addition to furthering the field of CN, a robust audience model will also provide a starting framework to solving the AI value alignment problem outline in Chapter 1. Modeling emotion in narrative agents gives us a mirror onto ourselves and a way to encode emotion into safer AI systems.

Chapter 4

Abstract

This chapter provides quantitative support for the theory that AI can align its values with humans using stories. I discuss the methodology used in AI value alignment and reinforcement learning. I then present two different case studies that provide evidence that value alignment through stories is possible. In discussing the methods and results of these case studies, I conclude that this field presents a tractable solution to the AI value alignment problem. However, I acknowledge the shortcomings of the field and the possibility of insuperable challenges in future research.

Introduction

Value alignment is a property of an intelligent agent indicating that it can only pursue goals that are beneficial to humans (Soares and Fallenstein, 2014). As discussed in Chapter 1, enumerating human values is extremely difficult, yet it presents the best potential solution to the AI control problem. However, in Chapter 2, I concluded that the creation of an audience model using CN techniques could present a framework for value alignment. In this chapter, I argue that the creation of an audience model using narrative information extraction presents a viable option for value enumeration. In other words, AI can learn human values by reading stories.

Stories encode cultural values, and there are many models of narrative in the humanities that attempt to classify the ways in which they do so. Stories, in the form of books, movies,

television, and oral tales provide a wealth of data on human values (Vargas et al., 2017). The technique of narrative information extraction, discussed in the previous chapter, uses large corpuses of stories to create models of narrative (Reagan et al., 2016). However, few researchers have used this technique to create a user model of the human beings reading and consuming stories. Riedl (2016) argues that a computer that can read and understand stories can reverse engineer the values tacitly held by a culture, given enough example stories produced by that culture. Unfortunately, most work in the field of AI narrative understanding is directed toward story generation (for video game systems and interactive stories) or for narrative models based on theory instead of actual text. In the following chapter, I will use two different case studies to support Riedl's claim – that stories can generate value aligned AI. These case studies are preliminary experiments attempted to use stories to achieve basic form of value alignment. First, Riedl's version of the Scheherazade system provides proof of concept for value alignment using stories. Second, Reagan's work on modeling emotional arcs of novels provides a more robust mechanism for reward-based value alignment. While both studies are rudimentary, they present promising areas of future work. Further work in this field is necessary to solve the overarching AI value alignment problem.

Background

Value Learning

Humans learn sociocultural values during early childhood development through a process of enculturation that includes cultural messages, parent beliefs, and peer interaction (De Raedt,

1998). The result of enculturation in civilized societies is that most people act according to a similar set of moral beliefs without explicitly being told what those beliefs are (Nickerson, 1998). The process of socialization creates beings that behave in an acceptable way – most of civilized society has been trained by this process to be polite, law-abiding, non-aggressive humans. Although the process of socialization is important, it is extraordinarily difficult to conceptualize formally (Miller, 1996). In other words, we are not consciously aware of the processes that create our own systems of values and behaviors (Nickerson, 1998). Because of the difficulty in identifying our own value system, aligning an intelligent agent's values with those of humanity is a difficult task. Although we do not possess an exhaustive or formalized list of human values, we do have access to an enormous corpus of stories and texts that help form our values from childhood onward. Storytelling is a strategy for relaying tacit knowledge – expert knowledge that can be effectively used but is otherwise hard to articulate (Riedl and Harrison, 2016). Tacit knowledge governs many of our everyday beliefs and behaviors, including procedures for acting in a socially acceptable way (Russell, 2016).

As scholars in the field of classical narratology recognized, stories encode many forms of tacit knowledge through symbolic meaning and archetypes (Saussure, 1916; Jung, 1936). Numerous fables and allegorical tales in both literary and oral traditions explicitly encode values. Aesop's fables, for example, provide a series of narratives accompanied by an explicitly stated model for behavior to be emulated by the reader. While such parables are obvious examples of value learning, many other fictional works encode the tacit knowledge of values and normative behavior expressed by the readers of a work. Work done by psychologists suggests that both written and oral stories are important channels of value socialization (Pratt et al., 1999). While

we learn values from stories during childhood, storytelling continues to be our preferred method of value communication throughout our lives. In a study of corporate culture, Buckler and Zien (1996) found that the stories employees told were the best indication of the strength of company culture. They conclude that corporate myths and legends provide the most effective means for communicating and reinforcing the shared values that distinguish corporations from each other. Therefore, we turn to stories as a means of communicating our values with intelligent artificial agents.

Reinforcement Learning

The theory of reinforcement learning (RL), rooted in psychological and neuroscientific perspectives on animal behavior, provides a model for how agents optimize control of an environment (Balleine, 2009). In more computational terms, it is the problem of learning how to act in a world so as to maximize a reward signal. Formally, a reinforcement learning problem is defined as <S, A, P, R>, where S is a set of states, A is a set of actions/effectors the agent can perform, P: $\{S \times A \times S\} \rightarrow [0, 1]$ is a transition function, and R: $S \rightarrow R$ is a reward function. The solution to a RL problem is a policy π: $S \rightarrow A$. An optimal policy ensures that the agent receives maximal long-term expected reward. The reward signal formalizes the notion of a goal, as the agent will learn a policy that drives the agent toward achieving certain states (Riedl and Harrison, 2016).

This technique has been successful in behavior generation of various artificial agents, including robots and AI systems (Kober et al., 2013). Deep neural networks trained with RL algorithms were able to generalize high-dimensional sensory inputs in various environments,

including playing Atari video games at a human level of skill (Mnih et al., 2015). One of the reasons that RL agents are so successful at adapting human behaviors is that they are driven by pre-learned policies. The agent is either coded with some model for how the world works, or it develops its own policies based on high-dimensional sensory data (Mnih et al., 2015; Kober et al., 2013). They use these pre-learned polices to choose actions that maximize long term reward. Therefore, RL provides a method of AI control, since the agent is compelled to achieve the given goal as encoded in the form of the reward signal (Saunders, 2017). However, this level of control does not guarantee that the solution an AI agent finds will not have the side effect of changing the world in a way that is harmful to humanity.

A classic example of psychotic AI behavior resulting from RL is that of the "paper clip generator." In this thought experiment, an AI agent is given the task of creating paper clips. Its reward function, therefore, will be maximized when it maximizes paper clip output. If a large reward is earned for acquiring more paper clips, but a small amount of reward is lost for each action performed, then the AI agent might decide to stop the production of cars, food, and technological products in order to turn the facilities into paper clip plants. Now instead of producing the necessities for human life, all factories make paper clips (Bostrom, 2013). This, clearly, would be a harmful outcome to humanity. Value alignment in AI systems is really just the construction of a reward signal that incorporates human values to avoid the psychotic AI behavior of the paper clip generator. Therefore, the maximum reward an agent could receive would necessarily be one that aligned with humanity's best interests.

Inverse reinforcement learning (IRL) is the process of constructing a reward function from observational data collected from some other agent – usually a human. IRL assumes that

the agent being observed is always acting in its own best interest and that the agent should receive a reward for copying the behaviors observed. The result is the learning of a reward signal that can then be used by a reinforcement learning agent to recreate an optimal policy (Abbeel and Ng, 2011). However, in practice, IRL has led to catastrophic results, since humans are rarely observed to be acting in their own best interest. Microsoft created an AI chatbot using IRL and observing the behavior of human users on Twitter. Within a day, the chatbot has learned to express racist, sexist, and offensive sentiments (Price, 2016). Therefore, IRL must be modified so as to only represent exemplar human behavior if value alignment is to be achieved.

Learning values from stories shares conceptual similarities with IRL; however, the stories studied by a system can be curated to express a range of human behavior – not just the slurs one can find on Twitter. Furthermore, written stories include normally unobservable mental operations that provide a more instructive model of the human reward system (Riedl and Harrison, 2016). Just like IRL, learning values from stories presents challenges. Stories rely on commonly shared knowledge and non-linear plot structure that confuse artificial readers, and many novels display non-exemplary human behavior that we would not want the agent to incorporate into its value system. However, such challenges can be overcome. Because of their wealth of data on human behavior and easy access, natural language stories present an excellent basis for AI value alignment.

Using stories to align agent values

In theory, value alignment in a reinforcement learning agent can be achieved by encoding a reward signal that gives rewards for solving problems and detracts rewards for performing any

actions that would be harmful to humans in either short-term or long-term scenarios (Soares, 2016). A value-aligned reward signal will reward the agent for obeying human social and cultural norms and penalize the agents for nonsocial, psychotic behavior. Therefore, the agent will be unable to maximize reward over time unless it conforms to social norms and human values that are producing the reward signal. For example, without value alignment, the paper clip generator might try to take over all known means of production to make as many paper clips as possible. With value alignment, the agent will receive more reward for avoiding the disruption of production that is beneficial to society. Therefore, although it is not maximizing the amount of paper clips it can make, it is maximizing its reward signal by complying with human norms and values (Bostrom, 2013). While, in theory, value alignment would be an excellent mechanism of reinforcement learning, codifying sociocultural values into a reward signal remains a challenge. The two case studies provided below provide preliminary evidence for two approaches to extracting value from narrative: crowdsourcing a number of example stories or providing a framework for emotional analysis of text.

Case Study 1 – Crowdsourcing narratives to achieve value alignment

Crowdsourcing is an effective means to produce a small, highly specialized corpus of narratives (Li et al., 2013). Such narratives could create a value-based rewards system for agents that are specialized to a single task. Furthermore, crowdsourced narratives can be simplified to make agent learning easier, avoiding similes, metaphorical language, and complex grammar that could confuse the AI agent. The Scheherazade system (Li et al., 2013) utilizes crowdsourced narratives to learn and generate stories.

Unlike most other story generators, Scheherazade does not contain any explicitly encoded story models. Instead, it generates models based on a corpus of crowdsourced stories. Scheherazade will attempt to tell a story about any topic requested by a human user. If it does not have a model for a story about the requested topic, it asks people on the Internet – via Amazon's Mechanical Turk service – to write example stories about that topic in natural language. The system then builds a model for a story about the topic, not using any classic narrative theories, but just from the corpus of crowdsourced stories it garnered from the Internet. Scheherazade represents a domain as a plot graph $<E, P, M, E_o, E_c>$, where E is a set of events (also called plot points), $P \subseteq E \times E$ is a set of precedence constraints, $M \subseteq E \times E$ is a set of mutual exclusion constraints, $E_o \subseteq E$ is a set of optional events, and $E_c \subseteq E$ are events conditioned on whether optional events have occurred (Li, 2014).

Precedence constraints indicate that a particular event must occur prior to another event occurring. Mutual exclusion constraints indicate when one event precludes the occurrence of another event, resulting in "branching" alternatives to how a situation can unfold. A plot graph represents the space of possible stories, including stories inferred to exist but that are not part of the crowdsourced corpus. Figure 10 shows an example plot graph describing the process of going to a pharmacy (Riedl and Harrison, 2016).

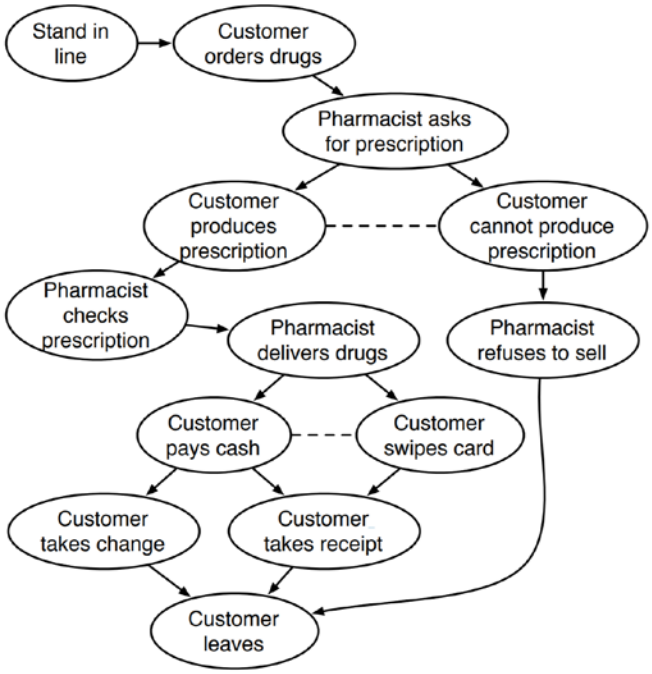

Figure 10. An example plot graph modeling a trip to a pharmacy

Nodes are plot points, solid arrows are precedence constraints, and dashed arrows are mutual exclusions (Li et al., 2013).

Utilizing the plot graph models it creates from crowdsourced narratives, Scheherazade can reproduce stories not in the original training corpus at near-human level ability (Li et al., 2013). By crowdsourcing many examples of a single plot, the system is able to achieve better coverage of all events and details which may have been left out in a single telling of the story. This plot graph learning process aligns the many examples of crowdsourced stories to extract the most reliable pattern of events. In using such a large corpus of similar narratives, the plot graph technique provides resilience to the noise introduced by the non-expert crowdsourced writers and filters out outlier events and unlikely sequences (Riedl and Harrison, 2016). One significant implication of this system is that narrative information extraction from texts is successful in creating narrative models. Without any encoded narrative theory, Scheherazade was able to

construct a robust model of narrative plot and then generate novel stories from that model. These findings suggest that narrative information extraction is a viable way to create computational narrative intelligence; the text is all that is needed.

While novel, the Scheherazade system focuses only on creating models of story fabula, not on a user model that could advance value alignment. Riedl and Harrison (2016) combine the model learning aspect of Scheherazade with inverse reinforcement learning to generate a value-aligned reward signal. After learning a plot graph using the technique described by Li et al, (2013), Riedl's system translates the plot graph into a trajectory tree in which nodes are plot points and directed arcs denote transitions from one plot point to another such that each path from root to leaf is a complete narrative (Riedl and Harrison, 2016). Unlike a plot graph, a trajectory tree is a literal representation of the story space. Each plot point in the plot graph can appear numerous times in different branches of the trajectory tree (see figure 11).

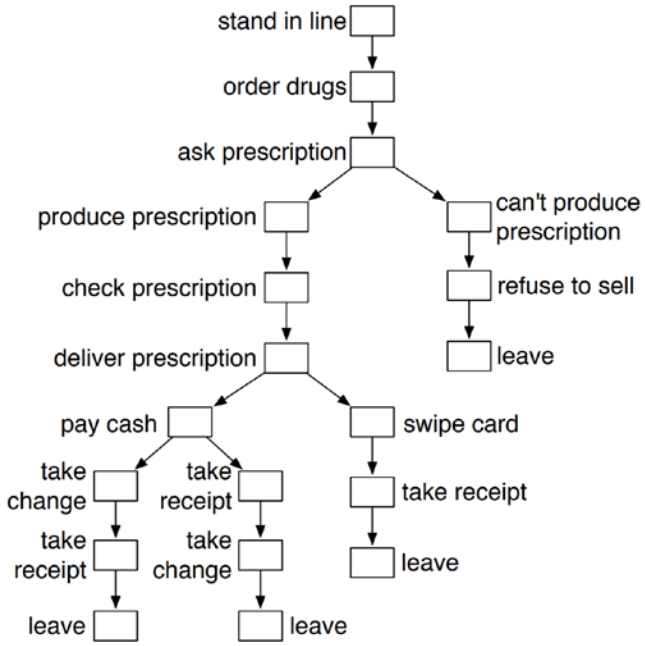

Figure 11. The trajectory tree generated from the pharmacy plot graph

The trajectory tree is used to produce the reward signal that will govern the agent's behavior by assigning reward values to each tree event. The reinforcement learning agent simultaneously tracks its state in the environment as well as its progress through the trajectory tree. If the agent deviates from the expected behavior of the tree, it receives a negative reward – a punishment for deviant behavior. Essentially, the crowdsourced narratives create a model of expected behavior, and a process of reward-based reinforcement learning teaches the agent to act according to expectations. In this process, the final policy learned by the agent meets the twin objectives of control and value alignment. The policy will compel the agent to solve the given problem, in this case by progressing through the trajectory tree. Additionally, the agent will prefer to solve the problem in a human-like fashion to maximize its reward signal. In problem spaces in which there are multiple solutions, some of which would be considered psychopathic, the agent will strongly prefer the sequence of behaviors that is most like the stories of typical human behavior from the crowdsourced stories. As a consequence, the agent will avoid behaviors that are harmful in most cases. For example, in the pharmacy narrative, agents are rewarded for adhering to the normal human behaviors provided by the crowdsourced narratives. As the agent progresses through the simulation, it maximizes rewards by following a similar set of steps; therefore, it would be unlikely to perform psychotic behaviors like robbing the pharmacy. However, this does not guarantee the agent will not act in a manner adverse to humans if circumstances justify it, just as humans will violate social and cultural norms when forced into extreme circumstances (Riedl and Harrison, 2016).

Riedl's trajectory tree reinforcement learning technique provides a mechanism for mapping expected human behavior to an artificial agent's governing behavioral policy. It

provides a proof of concept for value-aligned reinforcement learning, showing that intelligent virtual agents can learn from crowdsourced exemplar stories to behave in a more human-like manner, reducing the possibility of psychotic-appearing behavior. Furthermore, it reinforces the idea that narrative information extraction is a viable method of creating computational narrative intelligence. However, this technique focuses only on the fabula of stories – the plot events and the sequence in which they occur. The model of behavior created by story fabula leaves room for psychotic behavior since it does not address the internal motivational states of human operators in stories. A more robust model of expected behavior could be created from the more complex aspects of a narrative, such as the reader's emotional response.

The main limitation of this work is that it cannot be generalized. Reward signals learned from crowdsourced stories can only serve agents that have a very limited range of functionality. Crowdsourcing narratives to model every aspect of human interaction is not scalable. Plot graphs generated from crowdsourced narratives could omit important steps that cause the agent to learn an adversely valued policy. Furthermore, the creation of reward functions for reinforcement learning agents is not well understood (Riedl and Harrison, 2016). Often, manual tuning to the reward function is required for proper policies to be enacted. For an artificial general intelligence, stories will be needed that can directly address—or be generalized to—a broad range of contingencies. This is an open research problem.

Case Study 2 – Modeling emotional arcs to achieve value alignment

Researchers from Vermont's computational story lab recognized that stories have the power to transfer tacit information and describe our observations of the real world. They were

interested in exploring the idea that there are a finite number of story structures that humans tell over and over again (Reagan et al., 2016). As narratology scholars like Propp and Campbell recognized, we tend to prefer stories that fit into one of the story structures that we are familiar with and to reject narratives that do not align with our experience (Nickerson, 1998). The process of rejecting unfamiliar stories mirrors the computational process of reinforcement learning described by Riedl and Harrison (2016). If we could program artificial agents to replicate the process of rejecting unfamiliar stories, we would be much closer to creating safe, value-aligned AI. However, in order to do so, more information about the societal and cultural effects of story structure is needed – specifically, a better model of the human reader's motivation and emotions must be described. Therefore, Reagan et al. (2016) created an experiment to test aspects of the theories of folkloristics, specifically the idea that there are a finite number of core stories that humans tell.

In a study of over 1,000 books from the online Gutenberg corpus, Reagan et al. used a simple, robust sentiment analysis tool to extract the reader-perceived emotional content of written stories. Unlike previous models of narrative, their computational analysis does not examine plot structures at all. While the plot captures the fabula of a narrative, their work examines the emotional arc invoked by the text. The emotional arcs experienced by the user does not provide direct information about the plot but does give insight into a more complete model of the audience. Additionally, it signals which plot events are perceived as positively or negatively valued. Recent work in literary sentiment analysis has, in fact, suggested that shifts in the sentiment scores of texts can serve as reliable indicators for plot movement in novels (Gao et al., 2016).

To generate the emotional arcs of novels in the corpus, sentiment was analyzed by the labMT dataset in 10,000-word windows. This sentiment analysis tool slides through the text, scoring the sentiment of the natural language and creating a word shift graph (see figure 12). While the plot of novels is often nested, complicated, and confusing for artificial readers, the emotional arcs of stories can be more easily derived from sentiment scores of novels. To generate a sentiment score, a dictionary-based approach is used for transparency and understanding of sentiment. The LabMT dictionary was selected for robust performance over many corpora and best coverage of word usage (Reagan, 2017). To determine a sample T's average happiness the following equation was used:

$$h_{avg}(T) = \frac{\sum_{i=1}^{N} h_{avg}(\omega_i) \cdot f_i(T)}{\sum_{i=1}^{N} f_i(T)} = \sum_{i=1}^{N} h_{avg}(\omega_i) \cdot p_i(T)$$

where each of the N words in a given dictionary is denoted as ω_i, word sentiment scores as $h_{avg}(\omega_i)$, word frequency as $f_i(T)$, and normalized frequency of ω_i in T as $p_i(T) = f_i(T) / \sum^{N}_{i=1} f_i(T)$. In general, for each emotional arc the mean is subtracted before computing the distance or clustering.

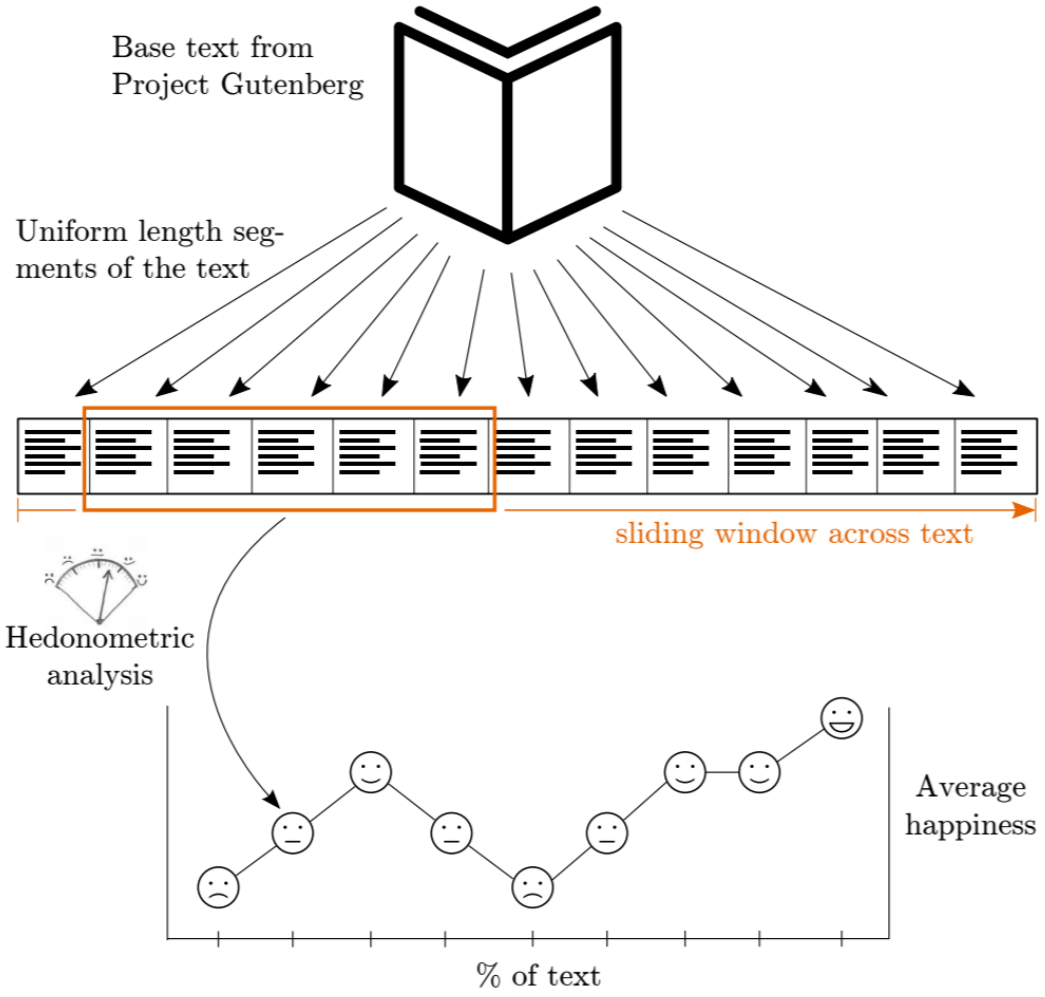

Figure 12. Schematic of the process of computing emotional arcs

The indicated uniform length segments (gap between samples) taken from the text form the sample with fixed window size set at N_w = 10, 000 words. The segment length is thus $N_s = (N - (N_w + 1))/n$ for N the length of the book in words, and n the number of points in the time series. Sliding this fixed size window through the book generated n sentiment scores which comprise the emotional arc (Riedl and Harrison, 2016).

Methods

In their analysis, Reagan et all. applied three independent tools: Matrix decomposition by Singular Value Decomposition (SVD), supervised learning by agglomerative (hierarchical) clustering with Ward's method, and unsupervised learning by a Self-Organizing Map (SOM, a

type of neural network). Each tool encompasses different strengths: the SVD finds the underlying basis of all of the emotional arcs, the clustering classifies the emotional arcs into distinct groups, and the SOM generates arcs from noise which are similar to those in our corpus using a stochastic process.

Singular Value Decomposition (SVD)

The standard linear algebra technique Singular Value Decomposition (SVD) was used to find a decomposition of stories onto an orthogonal basis of emotional arcs. Starting with the sentiment time series for each book bi as row *i* in the matrix A, the SVD was applied to find

$$A = U\Sigma V^T = WV^T$$

where U contains the projection of each sentiment time series onto each of the right singular vectors (rows of V^T, eigenvectors of $A^T A$), which have singular values given along the diagonal of Σ, with W = UΣ. Different intuitive interpretations of the matrices U, Σ, and V^T are useful in the various domains in which the SVD is applied; in this study the right singular vectors are used as an orthonormal basis for the sentiment time series in the rows of A. Σ and U are combined into the single coefficient matrix W for clarity and convenience, such that W now represents the mode coefficients (De Lathauwer, 1994).

Hierarchical Clustering

Ward's method was used to generate a hierarchical clustering of stories, which proceeds by minimizing variance between clusters of books (Ward, 1963). The mean-centered books and the distance function are represented as

$$D\left(b_i, b_j\right) = \ l^{-1} \sum_{t=1}^{l} \left| b_i(t) - b_j(t) \right) |$$

for t indexing the window in books b_i, b_j to generate the distance matrix.

Self-Organizing Map (SOM)

Self-Organized Map (SOM), an unsupervised machine learning method (a type of neural network), was used to cluster emotional arcs. The SOM works by finding the most similar emotional arc in a random collection of arcs. This study used an 8x8 SOM (for 64 nodes, roughly 5% of the number of books), connected on a square grid, training according to the original procedure (with winner take all, and scaling functions across both distance and magnitude). The neighborhood influence function at iteration *i* is represented by

$$\text{Nbd}_k(i) = \left[j \in \ \mathbb{N} \mid D(k, j) < \ \sqrt{N} \cdot (i + 1)^{\alpha} \right]$$

for a node k in the set of nodes $\mathbb{N}$, with distance function D given above and total number of nodes N (Kohonen, 1990). Reagan et al. (2016) take $\alpha = -0.15$ and implement the learning adaptation function at training iteration *i* as $f(i) = (i + 1)^{\beta}$, again with $\beta = -0.15$, a standard value for the training hyper-parameters.

Results

After applying the three methods to the corpus, they found that the text supported six separate emotional arcs:

- “Rags to riches” (rise).
- “Tragedy”, or “Riches to rags” (fall).
- “Man in a hole” (fall-rise).
- “Icarus” (rise-fall).
- “Cinderella” (rise-fall-rise).
- “Oedipus” (fall-rise-fall).

The same six emotional arcs were obtained using all three methods of sentiment clustering: as modes from a matrix decomposition by SVD, as clusters in a hierarchical clustering using Ward’s algorithm, and as clusters using unsupervised machine learning (Reagan et al., 2016). The computational evidence for emotional arcs mirrors many literary theories of story arcs, including those of Vonnegut (1947). Results for the SVD analysis of the text analysis are shown in figures 13 and 14 below.

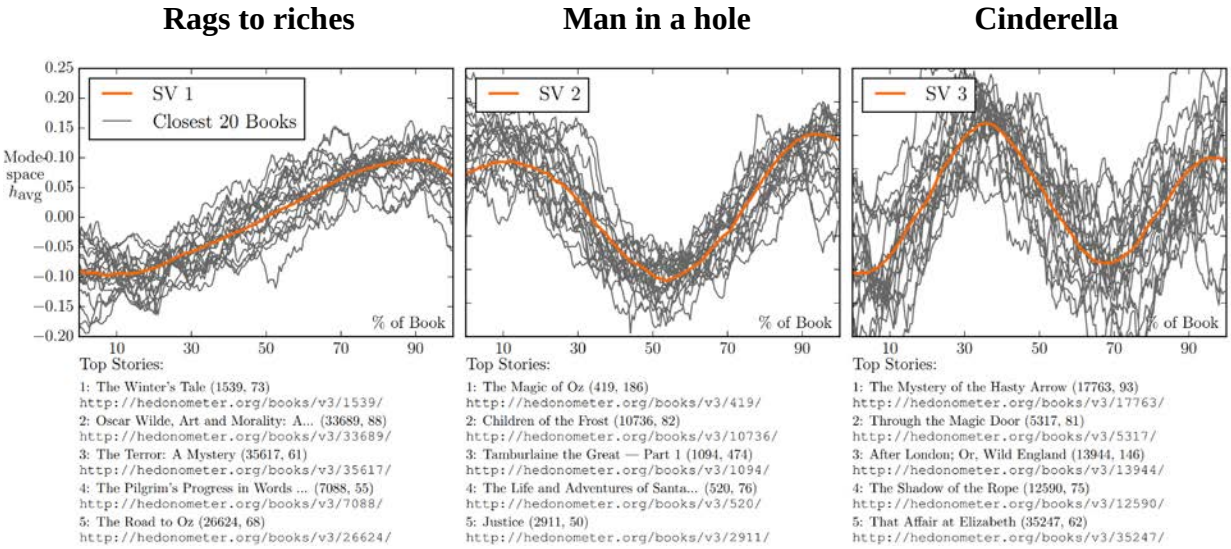

Figure 13. First 3 SVD nodes

To locate the emotional arcs on the same scale as the modes, we show the modes directly from the rows of V^T and weight the emotional arcs by the inverse of their coefficient in W for the particular mode. The closest stories shown for each mode are those stories with emotional arcs which have the greatest coefficient in W.

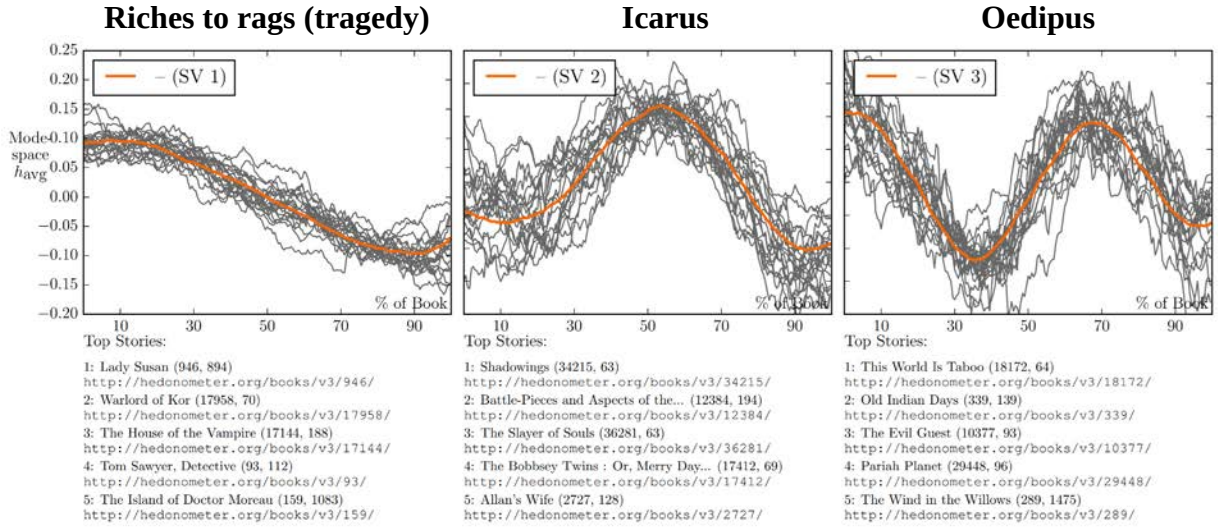

Figure 14. The negation of the first 3 SVD modes

To locate the emotional arcs on the same scale as the modes, we show the modes directly from the rows of V^T and weight the emotional arcs by the inverse of their coefficient in W for the particular mode. The closest stories shown for each mode are those stories with emotional arcs which have the greatest coefficient in W.

Understanding the emotional arcs of stories has sweeping implications for the field of computational narrative intelligence and AI safety (Bostrom, 2013; Mani, 2013; Soares, 2016). For computational narrative researchers, emotional arcs provide a method for better story generation algorithms, argument construction, and teaching common sense to artificial intelligence systems (Bex and Bench-Capon, 2010). Models of emotional arcs in stories can help agents understand the relationships between certain plot events and their corresponding emotional response. For example, a story understanding system built on emotional arcs would be able to understand why the queen died when we say “the king died and the queen died of grief.”

Furthermore, emotional arcs of plots could be used to create value aligned, safe AI (Reagan et al., 2016; Bostrom, 2013; Soares, 2016). Emotional arcs could be used as reward signals, like the trajectory trees used by Riedl and Harrison (2016). Many stories contain varied human behaviors (positive and negative) while still following an emotional arc that ends happily. Therefore, incorporating select emotional arcs into an agent’s reward system would give the agent the ability to pursue short-term negative reward actions in pursuit of a long term positive goal (Milli,2017). Using audience emotion as a reward signal would ensure that artificial agents always pursued an ending that resulted in a familiar “happy ever after.” Just as Riedl’s heuristic punished agents for deviating from the expected plot structure of a story, a new heuristic based on emotion could punish agents for deviating from an expected emotional response.

Of course, an approach to value alignment using emotional arcs extracted from text presents problems. To use emotional arcs as a reward signal, agents would have to have some way of measuring their audience’s emotional response. Current technology is capable of reading human facial emotion but would be complex and costly for a value alignment solution.

Furthermore, it is entirely possible that emotional arcs are too broad for practical AI learning. Understanding emotional arcs is only one small step in understanding human motivation systems and values. However, these emotional arcs provide an important insight into the model of the human reader. As discussed in Chapter 2, creating such a model gives us an important glimpse into our value system. Defining the emotional responses of the human audience of stories is a promising way forward to further define our own value system in a way that can easily be translated to machines through reinforcement learning.

Potential shortcomings in future work

Human value systems, reinforcement learning, and the technical details of narrative are all not fully understood. The difficulties in enumerating the problem space of the value alignment problem were laid out in chapter 1. Technical problems in scalability of narrative models, contextual understanding, and the many shortcomings of agents to read and understand stories was discussed in chapter 2. In this past chapter, I acknowledged that, despite promising work, the mechanisms of reinforcement learning are poorly understood. Any value alignment solution would have to address problems in the fields of the humanities, policy, machine learning, and narratology. Given the interdisciplinary nature of such work, truly created value aligned AI may be almost impossible. Even with successful value alignment, AI agents may still harm human beings. Just as Asimov's laws of robotics are too rigid to prevent harm, a value-based system may be too loosely defined to control intelligent agents. We simply do not know.

However, very few innovations that have shaped history are created with full knowledge of their inner workings and eventual impact. Despite the technical challenges and uncertainty of

success, I believe that an artificial intelligence that has been encultured with human values will not only be the safest option for humanity, but that it will also possess an unthinkable capacity to do good – to wield formidable computational power and algorithmic problem-solving processes while guided by the best values that humanity has to offer. Furthermore, in the current research landscape, I believe that using stories and narratives is the best way to encode our values into the machines we are creating.

Conclusion

Not only can we understand ourselves through the lens of human stories, but we can also use data driven approaches to narrative to create beneficial artificial intelligence agents. Preliminary work suggests that the implicit and explicit sociocultural knowledge encoded in stories can produce value-aligned reward signals for reinforcement learning agents. Using the values from human stories, such agents would only pursue actions that were beneficial for humans in the long term. Both crowdsourced models of narrative and value-alignment based on emotional arcs present a promising path forward to achieve AI value alignment.

An AI that learned and adopted the values implicit to a particular culture or society through stories will avoid psychotic behavior except under the most extreme circumstances. Such an intelligent agent will be instrumental to solving the AI control problem, and as the use of artificial intelligence becomes more prevalent in our society, and as artificial intelligence becomes more capable, solving the control problem grows more and more pressing. Giving artificial intelligences the ability to read and understand stories may be the most expedient means

of enculturing artificial intelligences so that they can better integrate themselves into human societies and contribute to our overall wellbeing.

Chapter 5

Conclusion

Summary

After introducing the motivation and structure of the thesis in Chapter 1, I introduced the main problems and areas of research in the field of AI safety. Because the field is still in its infancy, much conceptual work is needed to uncover tractable problems that can generate both short- and long-term solutions. I presented the challenges of the AI control problem and argued that value alignment is the best solution to the control problem. Using theories from the realm of the humanities, I further argued that natural language is the best source of data for encoding our tacitly held values.

In Chapter 2, I discussed how natural language has been modeled and generated in the field of computational narratology. I focus on the gap in the research concerning audience models. Having a robust model of the human audience of a narrative would not only create better story understanding and generation systems, but it would also serve as a practical way to create a model of human values. These values could then be used in solving the AI value alignment problem. In Chapter 3, I propose several primitive techniques that could be used to develop a better audience model (through emotional arcs or crowdsourcing). I also show that once an audience model of value is created, it could be used to control intelligent agents' behavior through inverse reinforcement learning heuristics.

Recommendations for future work

While this thesis has been primarily conceptual, the next steps for this topic are experimental. Work must be done to hone the technique of extracting models of audience emotion, expectations, and values, as done in Riedl and Harrison (2016) when they use crowdsourced narratives to model audience expectation. Ideally, future work could create finer grained models of audience from the existing corpus of novels and other text. Reagan et al.'s (2016) work on emotional arc of narrative provides a coarse model for audience emotion. A finer grained analysis would link certain plot events with negative or positive emotional responses to create a model that associated events with an emotional reaction. Another area of future work would be to expand the dimensionality of the emotional analysis of text. Reagan et al.'s work assumed a linear scale of emotion, which vastly oversimplifies the complexity of human emotional responses.

Work must also be done on understanding reinforcement learning heuristics. Ideally, a model created from further work in emotional analysis of text could be tested as a reward signal. Agents trained on this reward signal could be tested in any number of ways to examine the viability of emotional responses as a method of value alignment in AI systems.

Closing Remarks

In writing this thesis, I set out to perform my own experiments and figure out how to solve the value alignment problem. However, I soon realized that the interdisciplinary nature of AI safety was beyond my academic capabilities as an undergraduate researcher. Like Dautenhan

(2001), I came to see that, currently, the most necessary contributions to the field of AI safety must necessarily be on a conceptual level, not on the level of actual algorithmic implementation in AI systems. The latter will require far more sophistication and technical experimentation than one Penn State senior is capable of.

What I have done with this thesis is boil down hundreds of hours' worth of information into an actionable research step – the creation of more complete audience models using narrative information extraction techniques from the field of computational narratology. Results of research into narrative for autonomous agents, along the lines motivated in this thesis, can contribute to solving the AI value alignment problem. Not only do I believe that I have achieved a conceptual contribution with this thesis, but I also firmly believe in the value of this work moving forward. I have long known the magic of stories – their power to bring joy, healing, and inspiration to those who cracked open their covers. From my time working in a bookshop to the many months spent writing this thesis, my belief in the power of stories has never wavered. Stories hold the key to shaping the future of technology by transferring human values to machines. Just as I learned bravery from Ivanhoe, humor from Petruchio, and integrity from Jane Eyre, so can machines, if we are clever and thoughtful about the way we program them and the mechanisms we put in place to govern them.

BIBLIOGRAPHY

Abad, M. (2014). *Technium: technology is the seventh kingdom of nature*. [online] Blogthinkbig.com. Available at: https://blogthinkbig.com/technium-technology-is-the-the-seventh-kingdom-of-nature [Accessed 16 Feb. 2018].

Abbeel, P., & Ng, A. Y. (2011). Inverse reinforcement learning. In *Encyclopedia of machine learning* (pp. 554-558). Springer, Boston, MA.

Agre, P. *Computation and Human Experience*. Cambridge, UK: Cambridge University Press. 1997.

Asimov, I. (1976). *I, ROBOT*. New York, NY: Harper Collins.

Anderson, S. (2007). Asimov's "three laws of robotics" and machine metaethics. *AI & SOCIETY*, 22(4), pp.477-493.

Bailey, P. (1999). Searching for Storiness: Story-Generation from a Reader's Perspective. *In Working Notes of the Narrative Intelligence Symposium, AAAI Fall Symposium Series*. (= Technical Report FS-99-01.) Menlo Park, CA: AAAI Press, pp. 157–164.

Balleine, B. W., Daw, N. D., & O'Doherty, J. P. (2009). Multiple forms of value learning and the function of dopamine. In *Neuroeconomics* (pp. 367-387).

Barrett, E. *The Society of Text: Hypertext, Hypermedia and the Social Construction of Knowledge*. Cambridge.

Bednar, R. (2016). Signifier Signified. [image] Available at: https://blogs.ubc.ca/rbednar/2016/02/10/significance-of-semiotics-and-sounds/ [Accessed 2 Jul. 2018].

Bex, F. J., & Bench-Capon, T. J. (2010, November). Persuasive Stories for Multi-Agent Argumentation. In *AAAI fall symposium: computational models of narrative* (Vol. 10, p. 04).

Black, J. and Wilensky, R. (1979). An Evaluation of Story Grammars*. *Cognitive Science*, 3(3), pp.213-229.

Blackmore, S. (2011). *Consciousness: An Introduction*. 2nd ed. New York, NY, USA: Oxford University Press.

Bostrom, N. (2013). *Superintelligence*. Oxford: Oxford University Press.

Buckler, S. A., & Zien, K. A. (1996). The spirituality of innovation: learning from stories. *Journal of Product Innovation Management: AN INTERNATIONAL PUBLICATION OF THE PRODUCT DEVELOPMENT & MANAGEMENT ASSOCIATION*, *13*(5), 391-405.

Campbell, J. (1990). *The hero's journey*. Dorset, England: Element.

Chambers, N. and Jurafsky, D. (2008). Unsupervised Learning of Narrative Event Chains. *Proceedings of the 2008 Anniversary Meeting of the Association for Computational Linguistics*, Vol. 94305. 789-797.

Chatman, S. (1978). *Story and discourse*. Ithaca, N.Y.: Cornell University Press.

Chatuvedi, S., Srivastava, S., Daume, H. and Dyer, C. (2015). Modelling Dynamic Relationship Between Characters in Literary Novels.

Church, A. and Turing, A. (1937). On Computable Numbers, with an Application to the Entscheidungsproblem. *The Journal of Symbolic Logic*, 2(1), p.42.

Cheong, Y. G. (2007). *A Computational Model of Narrative Generation for Suspense*. PhD Thesis, Department of Computer Science, North Carolina State University.

Chong, E. (2017). The Control Problem. *IEEE Control Systems Magazine*, pp.14-16.

Coyne, B., Rambow, O., Hirschberg, J., and Sproat, R. (2010). Frame semantics in text-to-scene generation. *Knowledge-Based and Intelligent Information and Engineering Systems.* Springer, 375-384.

Dafoe, A. (2018). AI Policy Research Landscape. *Cambridge Leverhulme Centre for Intelligence.* Web.

Danaher, J. (2014). *Bostrom on Superintelligence (6): Motivation Selection Methods*. [online] Philosophicaldisquisitions.blogspot.ie. Available at: http://philosophicaldisquisitions.blogspot.ie/2014/08/bostrom-on-superintelligence-6.html [Accessed 30 Apr. 2018].

Dautenhahn, K., & Coles, S. (2001). Narrative intelligence from the bottom up: A computational framework for the study of story-telling in autonomous agents. *Journal of Artificial Societies and Social Simulation.*

De Lathauwer, L., De Moor, B., Vandewalle, J., & by Higher-Order, B. S. S. (1994, September). Singular value decomposition. In *Proc. EUSIPCO-94, Edinburgh, Scotland*, UK (Vol. 1, pp. 175-178).

De Raedt, L. (1998, July). Attribute-value learning versus inductive logic programming: The missing links. In *International Conference on Inductive Logic Programming* (pp. 1-8). Springer, Berlin, Heidelberg.

Dvorsky, G. (2014). *Why Asimov's Three Laws Of Robotics Can't Protect Us*. [online] Io9.gizmodo.com. Available at: https://io9.gizmodo.com/why-asimovs-three-laws-of-robotics-cant-protect-us-1553665410 [Accessed 16 Feb. 2018].

Fensel D. (2001) Ontologies. In: Ontologies. Springer, Berlin, Heidelberg.

Ferber, J., & Weiss, G. (1999). *Multi-agent systems: an introduction to distributed artificial intelligence* (Vol. 1). Reading: Addison-Wesley.

French, K. (2018). Cinderella Arc. [image] Available at: https://visage.co/kurt-vonnegut-shows-us-shapes-stories/ [Accessed 2 Jul. 2018].

Gao, J., Jockers, M. L., Laudun, J., & Tangherlini, T. (2016, November). A multiscale theory for the dynamical evolution of sentiment in novels. In *Behavioral, Economic and Socio-cultural Computing (BESC), 2016 International Conference* on (pp. 1-4). IEEE.

Gervas, P. (2018). Computer-driven creativity stands at the forefront of artificial intelligence and its potential impact on literary composition. In: R. Mesa, ed., *AC/E Digital Culture Annual Report.: Digital Trends in Culture. Focus: Readers in the Digital Age*. Accion Cultural Espanola, pp.88-98.

Gervás, P., Lönneker-Rodman, B., Meister, J. C., & Peinado, F. (2006, May). Narrative models: Narratology meets artificial intelligence. *In International Conference on Language Resources and Evaluation. Satellite Workshop: Toward Computational Models of Literary Analysis* (pp. 44-51).

Goyal, Amit, Ellen Riloff & Hal Daumé III (2010). "Automatically Producing Plot Unit Representations for Narrative Text." *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (EMNLP'2010), 77–86.

Grasbon, D. & N. Braun (2001). "A Morphological Approach to Interactive Storytelling." *Proceedings of Artificial Intelligence and Interactive Entertainment, CAST '01, Living in Mixed Realities, Sankt Augustin, Germany*, 337–40

Grace, K., Salvatier, J., Dafoe, A., Zhang, B. and Evans, O. (2017). When Will AI Exceed Human Performance? Evidence from AI Experts. *Cornell University Library*. [online] Available at: https://arxiv.org/abs/1705.08807.

Jaynes, J. (2000). *Origins of Consciousness in the Breakdown of the Bicameral Mind*. Rev. ed. New York, NY: First Mariner Books.

Jung, C. "Collective Unconscious." October 2, 1936, Analytical Psychology Club of New York City. Keynote address.

Kite, P. (2018). 09.01.04: Cinderella: A Cross-cultural Story. [online] Teachers.yale.edu. Available at: http://teachers.yale.edu/curriculum/viewer/initiative_09.01.04_u [Accessed 2 Jul. 2018].

Mani, I. and Hirst, G. (2013). *Computational modeling of narrative*. San Rafael, CA: Morgan & Claypool.

Martin, W. (1986). *Recent Theories of Narrative.* Ithaca: Cornell University Press. 5.

Mateas, M. and Sengers, P. (1999). AAAI Technical Report. *AAAI*, 99(01).

Mateas, M., & Stern, A. (2005, June). Structuring Content in the Façade Interactive Drama Architecture. *In AIIDE* (pp. 93-98).

Matuszek, C., Cabral, J., Witbrock, M., and DeOliveira, J. (2006). An Introduction to the Syntax and Content of Cyc. *AAAI Spring Sympo-sium: Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering.* 44-49.

Meister, J. (2012). *Digital Humanities 2012*. Hamburg: Hamburg Univ. Press.

Merriam-Webster's collegiate dictionary (10th ed.). (1999). Springfield, MA: Merriam-Webster Incorporated.

Muehlhauser, L., & Bostrom, N. (2014). Why we need friendly AI. *Think*, *13*(36), 41-47.

Mueller, T. (2003). Story understanding through multi-representation model construction. In Proceedings of the HLT-NAACL 2003 workshop on Text meaning-Volume 9. Association for Computational Linguistics, 46–53.

Müller, V. and Bostrom, N. (2014). "Future progress in artificial intelligence: A Survey of Expert Opinion", *Fundamental Issues of Artificial Intelligence* (Synthese Library; Berlin: Springer).

Nickerson, R.S. (1998). Confirmation Bias; A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2:175–220.

Kelly, K. (2010). *What Technology Wants*. Viking Press, pp.44-49.

Kober, J., Bagnell, J. A., & Peters, J. (2013). Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, *32*(11), 1238-1274.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78(9), 1464-1480.

Lehnert, W. G. (1981). "Plot Units: A Narrative Summarization Strategy." W. G. Lehnert & M. H. Ringle (eds.), *Strategies for Natural Language Processing*. Hillsdale, NJ: Lawrence Erlbaum.

Lewis, T. (2014). *A Brief History of Artificial Intelligence*. [online] Live Science. Available at: https://www.livescience.com/49007-history-of-artificial-intelligence.html [Accessed 30 Apr. 2018].

Li, B., Lee-Urban, S., Johnston, G., and Riedl, M. O. (2013). Story generation with crowdsourced plot graphs. *In Proceedings of the 27th AAAI Conference on Artificial Intelligence.*

Li, B. (2014). Learning Knowledge to Support Domain Independent Narrative Intelligence. Ph.D. Dissertation, Georgia Institute of Technology.

McCune, L. (1995). A normative study of representational play in the transition to language. *Developmental psychology*, 31(2), 198.

Milli, S., Hadfield-Menell, D., Dragan, A. and Russell, S. (2017). Should Robots be Obedient?. *University of California, Berkeley*.

Miller, P. J. (1996). Instantiating culture through discourse practices: Some personal reflections on socialization and how to study it. *Ethnography and human development: Context and meaning in social inquiry*, 183-204.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Petersen, S. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540), 529.

Morris, D. (2017). *Elon Musk: Artificial Intelligence Is the 'Greatest Risk We Face as a Civilization'*. [online] Fortune. Available at: http://fortune.com/2017/07/15/elon-musk-artificial-intelligence-2/ [Accessed 30 Apr. 2018].

Nelson, M., and Mateas, M. (2005). Search-Based Dram a Management in the Interactive Fiction Anchorhead. *In Proceedings of the First Artificial Intelligence and Interactiv e Digital Entertainment Conference*, 99–104. Menlo Park, CA : AAAI Press.

Paltz Spindler, L. (2005). Hero's Journey. [image] Available at: http://www.sfcenter.ku.edu/Workshop-stuff/Joseph-Campbell-Hero-Journey.htm [Accessed 2 Jul. 2018].

Pizzi, D. (2011). *Emotional Planning for Character-based Interactive Storytelling*. PhD Thesis, School of Computing, Teesside University, Middlesbrough.

Plaisant, C., Rose, J.,Yu, B., Auvil, L., Kirschenbaum, M., Smith, M., Clement, T. and Lord, G., (2006). Exploring Erotics in Emily Dickinson's Correspondence with Text Mining and Visual Interfaces, *in Proc. of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, 141-150.

Plot and Story. (2017). *Freiburg University*. [online] Available at: http://www2.anglistik.uni-freiburg.de/intranet/englishbasics/Plot01.htm [Accessed 15 Jul. 2018].

Pratt, M. W., Norris, J. E., Arnold, M. L., & Filyer, R. (1999). Generativity and moral development as predictors of value-socialization narratives for young persons across the adult life span: From lessons learned to stories shared. *Psychology and Aging*, *14*(3), 414-426.

Price, R. (2016). Microsoft is deleting its AI chatbot's incredibly racist tweets. *Business Insider*. Web. [Accessed 6 Jul. 2018].

Reagan, A., Mitchell, L., Kiley, D., Danforth, C. and Dodds, P. (2016). The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1).

Reagan, A. J., Danforth, C. M., Tivnan, B., Williams, J. R., & Dodds, P. S. (2017). Sentiment analysis methods for understanding large-scale texts: a case for using continuum-scored words and word shift graphs. *EPJ Data Science*, *6*(1), 28.

Riedl, M. O., & Bulitko, V. (2012). Interactive narrative: An intelligent systems approach. *AI Magazine*, *34*(1), 67.

Riedl, M. (2016). Computational Narrative Intelligence: A Human-Centered Goal for Artificial Intelligence. *Georgia Institute of Technology*.

Riedl, M., Harrison, B. (2016). Using Stories to Teach Human Values to Artificial Agents. In *Proceedings of the 2nd International Workshop on AI, Ethics and Society.*

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach.* Malaysia; Pearson Education Limited.

Sanghrajka, R., Witori, W., Schriber, S., Gross, M. and Kapadia, M. (2018). Computer-assisted Authoring for Natural Language Story Scripts. *30th Conference on Innovative Applications of Artificial Intelligence*, IAAI - 18.

Saunders, W., Sastry, G., Stuhlmuller, A. and Evans, O. (2017). Trial without Error: Towards Safe Reinforcement Learning via Human Intervention. *University of Oxford*.

Saussure, F., Bally, C., Sechehaye, C., Urbain, J. and Riedlinger, A. (1916). *Course in General Linguistics*.

Smith, M. N (2005). Email. "Curmudgeon Reflections on NORA."

Soares, N., and Fallenstein, B. (2014). Aligning superintelligence with human interests: A technical research agenda. *Technical Report 2014-8*, Machine Intelligence Research Institute.

Soares, N. (2016). The Value Learning Problem. In: *Ethics for Artificial Intelligence Workshop*. New York, NY, USA: IJCAI.

Stern, A. (2005). "Structuring Content in the Facade Interactive Drama Architecture." *Proceedings of Artificial Intelligence and Interactive Digital Entertainment (AIIDE 2005), Marina del Rey*.

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, [online] 5(42). Available at: https://bmcneurosci.biomedcentral.com/articles/10.1186/1471-2202-5-42 [Accessed 2 Feb. 2018].

van Cranenburgh, A. (2009). *Language, consciousness and the bicameral mind*. University of Amsterdam.

Valls-Vargas, J., Zhu, J., & Ontañón, S. (2017, August). From computational narrative analysis to generation: a preliminary review. In *Proceedings of the 12th International Conference on the Foundations of Digital Games* (p. 55). ACM.

Vonnegut, K. (1947). *On the Shape of Stories*. Web.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301), 236-244.

Alayna A. Kennedy

37 Meadow Lane, Doylestown PA, 18901 | aakennedy3696@gmail.com | (215) 260-7947

EDUCATION

Bachelor of Science in Engineering Science Aug 2014 - May 2018
The Pennsylvania State University (PSU), University Park, PA
Schreyer Honors College, Presidential Leadership Academy

4-week PSU College of Engineering study abroad program in China May 2016
Course on Impact of Culture on Engineering

HONORS, AWARDS, & PRESS

Dean's List for 5 of 6 semesters Aug 2014-Jan 2016, Sept 2016-Present
Diefenderfer Scholarship in Engineering - **$40,500** Aug 2016-Present
Robert and Myrtle Vierck Scholarship - **$3,600** Jan 2016-Present
Dreibelbis Renaissance Scholar Award- **$4,500** Aug 2015-Present
Academic Excellence Scholarship - **$20,000** Aug 2014-Present
Provost's Award - **$16,000** Aug 2014-Present
Schumaker Honors Scholarship - **$20,000** Aug 2014
Braddock Scholarship for Eberly College of Science - **$5,000** Aug 2014
Tall Clubs of America National Scholarship Winner - **$1,500** May 2014
Penn State News article "Mind moving matter" Sept 2016
Daily Collegian article "Schreyer scholar to present at international conference" Sept 2016
Featured Scholar in the Schreyer recruitment video "Moments" Mar 2017

RESEARCH

Undergraduate Researcher

National Science Foundation REU Fellow
University of Colorado, Department of Computer Science Colorado Springs, CO Jun 2016-Aug 2016

- Classified human walking speed with electromyogram inputs using different neural networks, such as feedforward, recurrent, dynamic, cascade architecture, and feature map networks to determine which presented the lowest computational cost and highest accuracy for use in myoelectric prosthetic devices
- Participated in daily classes lead by UCCS Computer Science professors, studying machine learning techniques and tools like random forests, decision trees, TensorFlow, Bayesian belief networks, and radial basis functions

Undergraduate Researcher
Movement Neuroscience Lab, University Park, PA Aug 2015-May 2015

- Examined the differences in neuromuscular synergy between dominant and non-dominant arm movements
- Recruited and administered reaching experiments using Kinereach technology on human subjects
- After a detailed literature review, prepared a write-up for the principal investigator and researchers
- Responsible for managing up to $1,000 of lab funds to compensate subjects and lab employees

Women in Science and Engineering Fellow
Bio Micro and Nano Systems Lab, University Park, PA Jan 2015-Jun 2015

- Selected to receive the Pennsylvania Space Grant Fellowship, which provides undergraduate women in STEM with financial compensation for participating in science and engineering research
- Quantified the size, density, and number of cancerous circulating tumor cells (CTCs) by capturing and analyzing images of the bloodwork of stage three and four cancer patients
- Performed experiments with a flexible microspring array device which filtered CTCs from whole blood samples
- Observed device filtered CTCs with up to 80% efficiency

PUBLICATIONS & PRESENTATIONS

A. Kennedy and R. Lewis, "Optimization of Neural Network Architecture for Biomechanic Classification Tasks with Electromyogram Inputs", International Journal of Artificial Intelligence & Applications, vol. 7, no. 5, pp. 1-16, 2016.

A. Kennedy "Machine Learning & the Future of Prosthetics" *Presentation.* International Conference for the Advancement of Bioscience and Bioengineering, October 2016, San Francisco, CA.

R. Sainburg, C. Maenza, and **A. Kennedy,** "Examining bilateral neural mechanisms of human handedness" *Presentation.* Penn State Kinesiology and Biomechanics Research Showcase, November 2015, University Park, PA.

WORK EXPERIENCE

Data & Analytic Technology Consulting Intern
PriceWaterhouse Cooper (PwC), New York, NY June 2017-Aug 2017

- Aggregated and analyzed data from the two companies' project management portfolios to make recommendations to the CTO on which projects should be put on hold during a merger and acquisition deal
- Tracked risks, action items, and decisions, and prepared weekly memos for both client and PwC leadership
- Advised analysts from PwC's artificial intelligence innovation accelerator to use KMeans clustering and Self Optimizing Maps (SOMs) to extract meaningful data for an EEG processing project during the its initial phases

IT Consultant and Business Analytics Intern
Highpoint Solutions, LLC, Philadelphia, PA May 2015-Aug 2015

- Generalized a client-specific Java compiler program to migrate documents into the Model N platform
- Collaborated with client teams to determine areas with inefficient data management and IT practices
- Received formal training on professional presenting and consulting skills

SKILLS

Programming MATLAB, Python, LaTeX, Java, JavaScript, HTML, Perl
Software Tableau, Image J, SolidWorks, Encog, Virtual Network Computing
Lab Skills AA spectroscopy, PCR, gel electrophoresis, titration, cell culturing, sterile technique

LEADERSHIP

Co-founder & Career Development Director
Schreyer for Women (SfW) Jan 2017-Present

- Served on the inaugural executive board of a club intended to promote women both locally and internationally
- Arranged speakers, presentations, and workshops to provide career development resources to underclassmen
- Brought two female CEOs to Penn State to speak to SfW membership
- Assisted in recruiting over 150 students during the club's first semester

Eta Class Member
The Presidential Leadership Academy Jan 2015-Present

- Over 3 years, took courses lead by PSU President Eric Barron that encouraged critical thinking and discussion
- Gained experience from trips to New York, Washington D.C., Seattle, Alabama, and Pittsburgh to meet with industry leaders and policy makers to discuss their perspectives on national issues
- Wrote a policy paper on the issue of compensating college athletes and presented to university leadership
- Worked closely with the director of the Academy to organize programming events for current students and to improve the program's alumni network involvement efforts

Lead Teaching Assistant & President
Leadership Jumpstart Program Sept 2014-Present

- Aided in students' development of leadership by overseeing a semester-long service project, providing feedback on student's skills, and
- Fulfilled administrative duties as president of the club that provides resources for current members of the class

Student Participant
Johnson & Johnson Future Leaders Program Jan 2015-May 2015

- Selected to participate in information sessions and workshops hosted by J&J to improve professional skills like presenting, networking, project development, and management

EXTRACURRICULARS

Data Strategist and Programmer
Nittany Data Labs, University Park, PA Aug 2016-Present

- Trained in machine learning and big data programming techniques during a Python-based training program
- Networked with companies such as 3M and Nestle to partner on data science consulting projects

University Relations Director
Penn State Lion Ambassadors Apr 2017-Present

- Lead a 4-person team to provide professional and service opportunities for over 100 students
- Partnered with student government, university administration officials, and student organizations on initiatives promoting diversity, sustainability, and academic excellence at Penn State
- Organized the first Lion Ambassador networking event, bringing dozens of alumni back to Penn State to network with current students

Campus Editor-at-Large
The Huffington Post Mar 2016-Present

- Wrote multiple articles on Penn State, college life, current events, and campus news, focusing on technology related events like the College of Information Sciences and Technology (IST) Startup Week
- Organized an event to promote Arianna Huffington's Sleep Revolution initiative, which encourages college students to focus on their own physical and mental health by getting enough sleep

General Member
Society for Women Engineers (SWE) Jan 2015-Present

- Engaged other undergraduate female engineers and served as a mentor for underclassmen, providing advice on class choices, internship search, and thesis topics

SERVICE & CIVIC ENGAGEMENT

Alternative Careers Initiative
Penn State University, Schreyer Honors College Apr 2017-Present

- Served as the student member on a committee intended to promote the viability of alternative career paths to undergraduate students, specifically in the Honors College
- Partnered with AshokaU, The Sullivan Foundation, and other innovative education programs to provide resources for students interested in non-traditional career paths

Community Member
The co.space, State College, PA Aug 2016-Present

- Member of a co-living home focused on the personal growth and professional developments of its residents
- Participated in multiple leadership retreats, entrepreneurship networking events, and community gatherings
- Founded a monthly community dinner club to bring students, professors, and community members together to discuss big picture questions

Port-Au-Prince Haitian Medical Mission
Our Lady of Mount Carmel Parish Sept 2013-Dec 2015

- Traveled to Port-Au-Prince to see thousands of patients and provide basic medical supplies to the community
- Organized multiple fundraisers, raising money and medical supplies to be sent to Haiti