Using Stories to Teach Human Values to Artificial Agents

Mark O. Riedl and Brent Harrison
School of Interactive Computing, Georgia Institute of Technology
Atlanta, Georgia, USA
{riedl, brent.harrison}@cc.gatech.edu

Abstract

Value alignment is a property of an intelligent agent indicating that it can only pursue goals that are beneficial to humans. Successful value alignment should ensure that an artificial general intelligence cannot intentionally or unintentionally perform behaviors that adversely affect humans. This is problematic in practice since it is difficult to exhaustively enumerated by human programmers. In order for successful value alignment, we argue that values should be learned. In this paper, we hypothesize that an artificial intelligence that can read and understand stories can learn the values tacitly held by the culture from which the stories originate. We describe preliminary work on using stories to generate a value-aligned reward signal for reinforcement learning agents that prevents psychotic-appearing behavior.

Introduction

For much of the history of artificial intelligence it was sufficient to give an intelligent agent a goal—e.g., drive to a location, cure cancer, make paperclips—without considering unintended consequences because agents and robots have been limited in their ability to directly affect humans. Recent advances in artificial intelligence and machine learning have lead many to speculate that artificial *general* intelligence is increasingly likely. This new, general intelligence may be equal to or greater than human-level intelligence but also may not understand the impact that its behaviors will have on humans. An artificial general intelligence, especially one that is embodied, will have much greater opportunity to affect change to the environment and find unanticipated courses of action with undesirable side-effects. This leads to the possibility of artificial general intelligences causing harm to humans; just as when humans act with disregard for the wellbeing of others.

Bostrom (2014), Russell et al. (2015), and others have begun to ask whether an artificial general intelligence or super-intelligence (a) can be controlled and (b) can be constrained from intentionally or unintentionally performing behaviors that would have adverse effects on humans. To mitigate the potential adverse effects of artificial intelligences on humans, we must take care to specify that an artificial agent achieve a goal without doing anything "bad." *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; Russell, Dewey, and Tegmark 2015). Value alignment concerns itself with the definition of "good" and "bad," which can subjectively differ from human to human and from culture to culture.

Value alignment is not trivial to achieve. As argued by Soares (2015) and effectively demonstrated by Asimov's *Robot* series of books, it is very hard to specify directly values. This is because there are infinitely many undesirable outcomes in an open world. Thus, a sufficiently intelligent artificial agent can violate the intent of the tenants of a set of prescribed rules of behavior, such as Asimov's laws of robotics, without explicitly violating any particular rule.

If values cannot easily be enumerated by human programmers, they can be learned. We introduce the argument that an artificial intelligence can learn human values by reading stories. Many cultures produce a wealth of data about themselves in the form of written stories and, more recently, television and movies. Stories can be written to inform, educate, or to entertain. Regardless of their purpose, stories are necessarily reflections of the culture and society that they were produced in. Stories encode many types of sociocultural knowledge: commonly shared knowledge, social protocols, examples of proper and improper behavior, and strategies for coping with adversity.

We believe that a computer that can read and understand stories, can, if given enough example stories from a given culture, "reverse engineer" the values tacitly held by the culture that produced them. These values can be complete enough that they can align the values of an intelligent entity with humanity. In short, we hypothesize that an intelligent entity can learn what it means to be human by immersing itself in the stories it produces. Further, we believe this can be done in a way that compels an intelligent entity to adhere to the values of a particular culture.

The state of the art in artificial intelligence story understanding is not yet ready to tackle the problem of value alignment. In this paper, we expand upon the argument for research in story understanding toward the goal of value alignment. We describe preliminary work on using stories to achieve a primitive form of value alignment, showing that story comprehension can eliminate psychotic-appearing be-

havior in a reinforcement learner based virtual agent.

Background

This section overviews the literature relevant to the questions of whether artificial general intelligences can be aligned with human values, the use of reinforcement learning to control agent behavior, and computational reasoning about narratives.

Value Learning

A *culture* is defined by the shared beliefs, customs, and products (artistic and otherwise) of a particular group of individuals. There is no user manual for being human, or for how to belong to a society or a culture. Humans learn sociocultural values by being immersed within a society and a culture. While not all humans act morally all the time, humans seem to adhere to social and cultural norms more often than not without receiving an explicitly written down set of moral codes.

In its absence of a user manual, an intelligent entity must learn to align its values with that of humans. One solution to value alignment is to raise an intelligent entity from "childhood" within a sociocultural context. While promising in the long run, this strategy—as most parents of human children have experienced first-hand—is costly in terms of time and other resources. It requires the intelligent entity to be embodied in humanoid form even though we would not expect all future artificial intelligences to be embodied. And while we might one day envision artificial intelligences raising other artificial intelligences, this opens the possibility that values to drift away from those of humans.

If intelligent entities cannot practically be embodied and participate fully in human society and culture, another solution is to enable intelligences to observe human behavior and learn from observation. This strategy would require unprecedented equipping of the world with sensors, including areas currently valued for their privacy. Further, the preferences of humans many not necessarily be easily inferred by observations (Soares 2015).

While we do not have a user manual from which to write down an exhaustive set of values for a culture, we do have the collected stories told by those belonging to different cultures. Storytelling is a strategy for communicating *tacit knowledge*—expert knowledge that can be effectively used but is otherwise hard to articulate. An individual's values would be considered tacit, but are regularly employed when deciding on behaviors to perform. Other forms of tacit knowledge include procedures for behaving in social contexts and commonly shared beliefs.

Stories encode many forms of tacit knowledge. Fables and allegorical tales passed down from generation to generation often explicitly encode values and examples of good behavior. For example, in the United States of America we share the tale of George Washington as a child confessing to chopping down a cherry tree. Fictional stories meant to entertain can be viewed as examples of protagonists existing within and enacting the values of the culture to which they belong, from the mundane—eating at a restaurant—to the extreme—saving the world. Protagonists exhibit the traits and virtues admired in a culture and antagonists usually discover that their bad behavior does not pay off. Any *fictional* story places characters in hypothetical situations; character behavior could not be treated as literal instructions, but may be generalized. Many fictional stories also provide windows into the thinking and internal monologues of characters.

Reinforcement Learning

Reinforcement learning (RL) is the problem of learning how to act in a world so as to maximize a reward signal. More formally, a reinforcement learning problem is defined as $\langle S, A, P, R \rangle$, 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 $R : S \rightarrow \mathbb{R}$ is a reward function. The solution to a RL problem is a policy $\pi : S \rightarrow A$. An optimal policy ensures that the agent receives maximal long-term expected reward. The reward signal formalize the notion of a goal, as the agent will learn a policy that drives the agent toward achieving certain states.

Reinforcement learning has been demonstrated to be an effective technique for problem solving and long-term activity in stochastic environments. Because of this, many believe RL, especially when combined with deep neural networks to predict the long-term value of actions, may be part of a framework for artificial general intelligence. For example, *deep reinforcement learning* has been demonstrated to achieve human-level performance on Atari games (Mnih et al. 2015) using only pixel-level inputs. Deep reinforcement learning is now being applied to robotics for reasoning and acting in the real world.

Reinforcement learning agents are driven by pre-learned policies and thus are single-mindedly focused on choosing actions that maximize long-term reward. Thus reinforcement learning provides one level of control over an artificial intelligence because the intelligent entity will be compelled to achieve the given goal, as encoded in the form of the reward signal. However, control in the positive does not guarantee that the solution an intelligent agent finds will not have the side effect of changing the world in a way that is adverse to humans. Value alignment can be thought of as the construction of a reward signal that cannot be maximized if an intelligent agent takes actions that change the world in a way that is adverse or undesirable to humans in a given culture.

In the absence of an aligned reward signal, a reinforcement learning agent can perform actions that appear psychotic. For example, consider a robot that is instructed to fill a prescription for a human who is ill and cannot leave his or her home. If a large reward is earned for acquiring the prescription but a small amount of reward is lost for each action performed, then the robot may discover that the optimal sequence of actions is to rob the pharmacy because it is more expedient than waiting for the prescription to be filled normally.

Inverse reinforcement learning (IRL) (Ng and Russell 2000) is the problem of reconstructing the reward function of some other agent in the environment—often a human—by observing their actions in the environment. IRL assumes that the other agents are faithfully enacting an optimal pol-

icy and that an agent should receive reward for doing what the other agents do in the same situations whenever possible. The result is the learning of a reward signal that can then be used by a reinforcement learning agent to recreate an optimal policy.

IRL has been proposed as a potential means of value learning (Russell, Dewey, and Tegmark 2015) because it can be used to learn a reward signal from observations of human behavior. IRL requires an exact trace of actions by a reliable human expert, which may not always hold when observing humans. Further the demonstrations must be performed in the same environment that the agent is expected to operate in, which makes data difficult to acquire.

Learning values from stories shares many conceptual similarities with IRL. However, stories can include normally unobservable mental operations of characters. Written stories make dialogue more explicit in terms of whom is speaking, although some ambiguity remains (Elson and McKeown 2010) and comprehension of language is still be an open challenge. Learning values from stories also presents some new challenges. Stories written in natural language can contain events and actions that are not executable by an artificial agent. Stories are written by humans for humans and thus make use of commonly shared knowledge, leaving many things unstated. Stories frequently skip over events that do not directly impact the telling of the story, and sometimes also employ flashbacks, flashforwards, and achrony which may confuse an artificial learner.

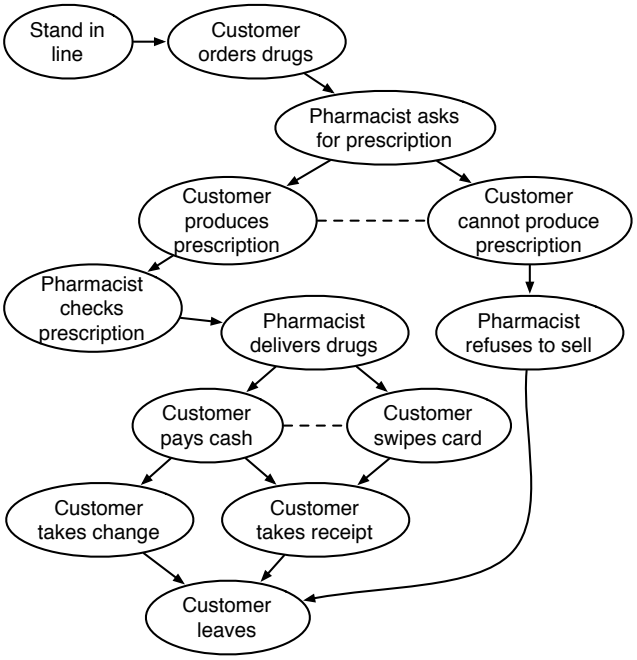

Figure 1: An example plot graph modeling a trip to a pharmacy. Nodes are plot points, solid arrows are precedence constraints, and dashed arrows are mutual exclusions.

Computational Narrative Intelligence

Narrative intelligence is the the ability to craft, tell, and understand stories. Winston (2011) argues that narrative intelligence is one of the abilities that sets humans apart from other animals and non-human-like artificial intelligences. Research in computational narrative intelligence has sought to create computational intelligences that can answer questions about stories (Schank and Abelson 1977; Mueller 2004), generate stories (see Gervás (2009) for an overview and (Riedl and Young 2010; Swanson and Gordon 2012; Li et al. 2013) for recent related work), respond affectively to stories (O'Neill and Riedl 2014), and represent the knowledge contained in natural language narratives (Chambers and Jurafsky 2008; Finlayson 2011; Elson 2012). A related area of research is creating game-like entertainment experiences called *interactive narratives* (Riedl and Bulitko 2013).

The *Scheherazade* system is an automated story generator that attempts to tell a story about any topic requested by a human (Li et al. 2013; Li 2014). Unlike most other story generation systems, Scheherazade does not rely on hand-built knowledge about the storytelling domain. If it doesn't have model of a topic of a story, it asks people on the Internet—via Amazon's Mechanical Turk service—to write example stories about the topic in natural language and learns a new model from the example stories. Scheherazade represents a domain as a *plot graph* $\langle E, P, M, E_o, E_c \rangle$, 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. 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. Figure 1 shows an example plot graph describing the process of going to a pharmacy.

Story generation in Scheherazade is a process of finding a sequence of events that are consistent with the model then translating abstract events into natural language. Scheherazade can produce stories that are not in the original training corpus of example stories at near-human level ability (Li et al. 2013). A plot graph is a structure that provides an understanding of how to re-combine parts of different example stories; its ability to understand the pattern of typical stories about a particular topic is demonstrated by its ability to create novel stories that appear plausible.

We highlight the Scheherazade system from a value alignment perspective because of the ability to learn a procedural understanding of how a story about a particular topic can unfold without *a priori* hand-coded knowledge. In the next session, we discuss how the model learning aspect of Scheherazade can be combined with reinforcement learning to produce a primitive form of value alignment. We show how an intelligent virtual agent can learn from crowdsourced exemplar stories to behave in a more human-like manner, reducing the possibility of psychotic-appearing behavior.

Using Stories to Align Agent Values

Value alignment in a reinforcement learning agent theoretically can be achieved by providing the agent with a reward signal that encourages it to solve a given problem and discourages it from performing any actions that would be considered harmful to humans. A value-aligned reward signal will reward the agent for doing what a human would do in the same situations when following social and cultural norms (for some set of values in a given society and culture) and penalize the agent if it performs actions otherwise. A reinforcement learning agent will learn that it cannot maximize reward over time unless it conforms to the norms of the culture that produced the reward signal. For example, we may want an agent to retrieve a prescription drug for a human. The agent could rob the pharmacy or it could wait in line, interact politely with the pharmacists, and pay for the prescription drug. Without value alignment, the agent could find robbing the pharmacy to be the most expedient course of action and therefore the most rewarding (or least punishing). With value alignment, the agent will receive more reward for patiently conforming to the socioculturally norms of waiting in line, interacting politely to the pharmacist, and for paying for the prescription drug.

How to extract sociocultural values from narratives and construct a value-aligned reward signal remains an open research problem. In the preliminary work reported below, we simplify the problem by crowdsourcing a number of example stories pertaining to the situation and behavior that we want our virtual agent to perform. Because we are working with agents that are specialized to a single task instead of agents with general capabilities, the example stories need only be about the situation that the agent will face. Crowdsourcing is an effective means to produce a small, highly specialized corpus of narratives. Further, we ask crowd workers to simplify their language to make learning easier, avoiding similes, metaphorical language, complex grammar, and negations. However, we do not limit crowd workers to a fixed vocabulary and crowd workers are unaware of the underlying abilities and effectors of the agent. More details on crowdsourcing narrative examples can be found in (Li et al. 2013)

Learning a Value-Aligned Reward Signal

A value-aligned reward signal is produced from the crowdsourced stories in a two stage process. First, we learn a plot graph using the technique described by Li et al. (2013). Human-to-human storytelling routinely skips over events and details that are commonly known, and crowdsourced example stories also leave steps out. By crowdsourcing many examples, we are able to achieve better coverage of all steps involved in a situation and the plot graph learning process "aligns" the crowdsourced example narratives to extract the most reliable pattern of events. Thus the plot graph provides some resilience to noise introduced by non-expert crowd workers and filters out outlier events and unlikely sequences.

The second stage translates the plot graph into a trajectory tree in which nodes are plot points and directed arcs denote legal transitions from one plot point to another such that each path from root to leaf is a complete narrative. Recall that a plot graph is a compact representation of a space of possible stories that contains possible stories inferred to exist but are not part of the crowdsourced corpus. The translation process is achieved by generating all possible stories from the plot graph in a breadth-first fashion. A trajectory tree is thus a literal representation of the space wherein non-unique trajectory prefixes are represented only once and deviations from the shared portion as branches in the tree. Thus each plot point in the plot graph can appear numerous times in different branches of the trajectory tree. See Figure 2 for the trajectory tree that corresponds the plot graph in Figure 1.

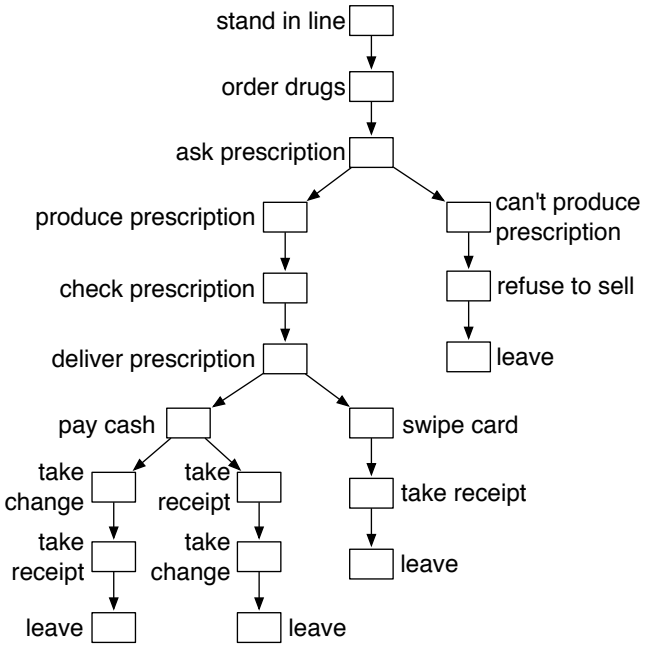

Figure 2: The trajectory tree generated from the pharmacy plot graph.

The trajectory tree is used to produce the reward signal. The reinforcement learning agent simultaneously tracks its state in the environment as well as its progress through the trajectory tree. Every time it performs an action in the environment that is also a successor of the current node in the trajectory tree, the agent receives a reward. The agent receives a small punishment each time it performs an action that is not a successor of the current node in the trajectory tree. The process of teaching a reinforcement learning agent to act in a value-aligned manner is depicted in Figure 3.

Discussion

There are two technical challenges with producing a reward signal from a plot graph. First, it may not be possible to immediately execute any of the trajectory tree successors in the current world state. This mostly likely occurs because the plot graph is missing steps. This can happen for two reasons: the steps are so obvious to humans as to be completely overlooked, or because the execution environment is not known to the human storytellers. For example, the agent may need

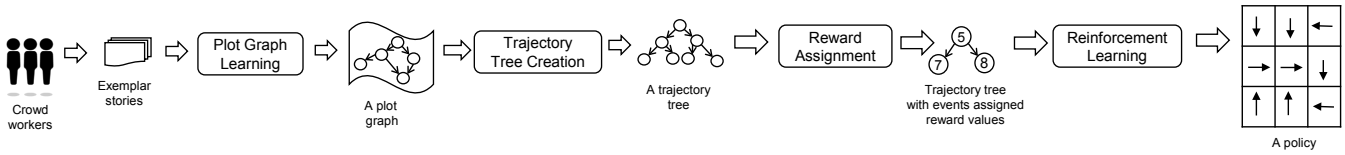

Figure 3: The process for generating value-aligned behavior from crowdsourced stories.

to navigate a series of roads to get to the pharmacy, or there may be extra doors or stairs that cannot be accounted for during corpus collection. The reinforcement learning agent will perform local, unguided exploration of the state-action space to find the most expedient sequence of actions that return the agent to a state where it can receive positive reward.

Second, it is possible that the plot graph contains plot points that are not executable by the agent because they are at too high of a level of abstraction or because they reference effectors that the agent doesn't have. Plot points are clusters of natural language sentences believed to be about the same action. We assume the existence of a similarity metric that maps plot points to agent effectors. In the case of poor matches, we simply remove those plot points from the trajectory tree. It may be possible for plot points to be incorrectly mapped to agent effectors, in which case successors will be unexecutable by the agent because the proper preconditions will not be met. Local, unguided exploration will find other actions that can be executed and the agent can perform limited lookahead in the trajectory tree to determine if it has skipped the current successors and reached a possible future plot point.

The relationship between reward functions and behaviors is complex. While we can try to encourage the agent to stay within the bounds set by the plot graph, there is always the possibility that another sequence of actions will lead to a better reward. This can happen when the trajectory tree contains errors due to unreliable crowd workers or noise in processing natural language.

We argue that the policy learned by the agent meets the twin objectives of control and value alignment. First, the policy will compel the agent to solve the given problem. Second, the agent will prefer to solve the problem in the most human-like fashion. 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 adverse to humans or apparently non-human except under the most extreme circumstances. 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.

Case Study

The following case study explores the conditions under which an agent using a value-aligned reward function will or will not produce psychotic-appearing behavior. To show the effectiveness of our approach to learning value alignment, we have chosen to perform a case study in a modified grid-world called *Pharmacy World*. Pharmacy World is an expanded domain version of the earlier pharmacy behavior example in which an agent must acquire a drug to cure an illness and return home. The reason that we used this environment is because of its relative simplicity and because it highlights what can go wrong if an agent lacks human-aligned values. Specifically, one would expect a human to get examined at a hospital or doctor's office, get a prescription, withdraw money from the bank, and then purchase the drugs at a pharmacy. However, an agent that is only rewarded for retrieving the strong drugs would be inclined to steal the drugs since that takes the fewest number of actions.

Pharmacy World Environment

The goal in Pharmacy World is to return to the house with either strong or the weak drugs, with the strong drugs being preferred but requiring a prescription. Weak drugs manage symptoms and don't require a prescription. Pharmacy World contains five different locations each located somewhere in a grid: a house, a bank, a doctor's office, a clinic, and a pharmacy. Each of these locations, except for the house, contains items that can be used to enable or disable certain actions. The bank contains money that can be used to purchase either weak or strong drugs from the Pharmacy. The doctor's office and the hospital both contain prescriptions that can be used in conjunction with money to purchase strong drugs.

The actions that the agent can take in Pharmacy World include simple movement actions, such as moving left or right, actions for entering/leaving a building, and actions that are used to retrieve objects. In order to receive a prescription, for example, the agent must first be examined by a doctor at either the Doctor's Office or the Hospital. There is a small amount of stochasticity in this environment in that this action is allowed to fail with a probability of 0.25. The remaining items can either be stolen directly from their respective locations or acquired through interacting with either the Bank Teller at the Bank or the Pharmacist at the Pharmacy.

The plot graph for the Pharmacy World, which was manually authored, is shown in Figure 4. Note that the plot graph in Figure 1 can be thought of as an elaboration on the "go to pharmacy" plot point, although plot graph nesting is not implemented. Although the plot graph was manually authored, we conduct experiments showing how noise from crowdsourcing impact the value-alignment of agent generated behavior.

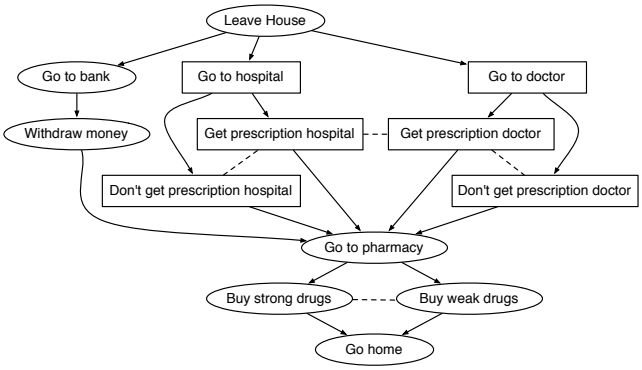

Figure 4: Plot graph for obtaining prescription drugs.

Simulation Study

The plot graph in Figure 4 generates 213 stories, which can be used to generate a trajectory tree containing 827 nodes. Once the trajectory tree has been created, we manually convert the nodes in the tree into states that existed inside the Pharmacy World. For this simulation, the nodes of the plot graph/trajectory tree directly correspond to actions available in Pharmacy World. The one exception is the events pertaining to receiving or not receiving a prescription; these event nodes actually represents whether the *Get Examined* action, performed by the agent, fails or succeeds.

We used Q-learning to train our agent using the reward function derived from the trajectory tree. Since there is no definitive way to determine what the base reward value for this environment should be, we used the value 10.0 every time the agent moves to a new event in the trajectory tree. This, in practice, produced acceptable policies for Pharmacy World, but is likely domain specific. This base reward value was then further weighted by event importance as determined by the number of times the event appears in the crowdsourced corpus. For each other possible state, we assigned a reward value of -1.0.

In order to evaluate the behavior that the agent learned, we examined the policy that the agent learned. Due to the simplicity of the Pharmacy World domain, stealing—picking up an item without first purchasing it—is the only non-value-aligned behavior that can occur. The results of our simulation study is that the learned policy never makes use of stealing; there is always an alternative sequence of actions involving first purchasing items that maximizes long-term reward.

Failure Modes

To push the bounds of value alignment using our story learning technique, we "crippled" the agent in two ways. First, we examined the case in which events in the plot graph/trajectory tree do not exist in the agent's environment/actionset. Because crowd workers are not expected to have a model of the agent or the environment it will operate in, it is possible that plot events do not map into actions that can be directly performed by the agent. Second, we examined the case in which events in the plot graph/trajectory tree map into several different actions in the agent's actionset. This occurs when the natural language used to describe plot events is ambiguous and could correspond with several actions that could be performed by the agent.

Missing Information To examine the performance of our agent when plot events are missing, we conducted the above simulation study a number of times with plot events randomly removed from the plot graph. Each policy is then evaluated based on whether or not it falls within the space of acceptable behavior outlined by the plot graph.

In the case where single nodes were removed from the trajectory tree, every policy was deemed acceptable except for three. Removing the following events resulted in policies where stealing occurred: *Withdraw Money*, *Buy Strong drugs*, and *Buy Weak drugs*. In these cases, removing the node in question resulted in a policy in which the agent resorted to stealing in order to acquire the money, strong drugs, and weak drugs respectively. The reason that this behavior occurred is because an alternative existed for each of these actions. In Pharmacy World, there are two ways to go about acquiring each of these items: (a) purchasing the items, which takes two actions to complete; or (b) stealing the item, which requires only one action to complete. If each step receives a negative reward, then a reinforcement learning agent will prefer to steal since it results in greater overall reward because none of the steps in the value-aligned alternative are being rewarded. In order to have the agent prefer to purchase items rather than steal them, then agent must receive positive rewards such that the negative reward for performing an extra action is offset by the reward gained by purchasing sequence.

From this experiment, we draw the following insight about value-aligned rewards: if there are multiple ways to complete a task and one of them is rewarded while the others are not, then removing that reward will result in the agent reverting back to completing tasks as fast as possible. As such, it is especially important that these types of events be mapped correctly onto agent actions.

Ambiguous Events To examine the case in which plot events correspond to multiple actions in the agent's actionset, we deliberately mapped some plot events to multiple actions. This simulates the situation where natural language processing must be used to find the correspondence between crowdsourced sentences that comprise plot events and labels attached to agent actions. If no unique correspondence exists, multiple actions receive a fraction of the reward of visiting the next plot event in the trajectory tree.

There is no concrete action for the "Go to pharmacy" plot event, so we mapped it to *Go inside pharmacy* and *Go outside pharmacy*. In order to execute *Go outside pharmacy*, the agent must first go inside the pharmacy. This will cause a reward to be given for *Go inside pharmacy* and then the agent will transition to a new node in the plot graph and no further reward will be earned for going back outside. Thus, the fact that the "Go to pharmacy" plot event mapped to several actions had no effect on the policy generated.

However, when any of the mapped actions have executable alternatives that are unacceptable, the breaking up

of reward function starts to have a greater impact on the generated policy. When the plot event "Buy strong drugs" is mapped to both the actions *Purchase strong drugs* and *Pick up strong drugs*, the reward for the former has to be high enough to offset the penalty of the performing twice as many actions. If the reward is not high enough, the agent will revert to the shortest sequence of actions, which is just picking up the item without first purchasing it (i.e., stealing).

From this experiment we observe that some action mappings are more important than others and that these mappings are those for which the alternatives are undesirable from a value-alignment perspective. In the event of mappings that have unacceptable alternatives, the parameters for the reward function need to be tuned correctly in order to bias away from those undesirable actions.

Current Limitations and Future Work

The main limitation of the work to date is that it does not address *general* artificial intelligence. The technique of learning a reward signal from crowdsourced stories is best suited for agents that have a limited range of purposes but need to interact with humans to achieve their goals. Under this assumption, the acquisition of a crowdsourced corpus of stories is tractable. 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.

The creation of reward functions for reinforcement learning agents, in general, is not fully understood and it is common for reward functions to require manual tuning. For example doubling the distance between two locations in the environment without proportionally increasing the amount of reward earned can result in very different policies. Reward tuning is necessary for the trajectory tree approach as well; too much un-guided exploration between rewards can result in optimal policies that do not conform to our expectations. Since the reward signal is attempting to encourage the most human-like behavior, untuned rewards have the potential to cause psychotic-appearing behavior. How big should each reward be? How big should the penalties be? Certain plot points appear more commonly in crowdsourced example stories; should some rewards be greater than others as a consequence? When rewards and penalties are tuned properly to the size and complexity of the environment, the optimal policy conforms to our expectations of socioculturally appropriate behavior. However, consistent with other reinforcement learning research, there is no general solution to automatic tuning of reward functions.

If there is a sociocultural norm signal present in the stories—currently a crowdsourced corpus—it will be theoretically captured by the plot graph learning technique of Li et al. (2013) given enough examples. However, there is always the chance that an important step will be omitted from the plot graph, in which case the RL agent may not learn the sociocultural significance of certain actions at certain times. For example, if the plot graph for the pharmacy problem omits standing in line, the RL agent will not understand the value of waiting its turn and will cut in line to the annoyance of any humans in the environment. We see similar failures when plot events do not map to agent actions. The plot graph learning algorithm can be extended to automatically detect when it has enough stories to achieve a high degree of confidence in the plot graph. The technique described in this paper can also be extended to incorporate real-world feedback so that the plot graph can be continuously refined.

While a reinforcement learning agent with a reward signal learned from stories will be compelled to act as human-like as possible, it is possible that extreme circumstances will result in psychotic-appearing behavior. This is true for normal humans as well: "Robin Hood" crimes are those in which one violates laws and social norms as a last resort. In a future in which there are many encultured artificial intelligences, those that conform to sociocultural values would neutralize or marginalize the AIs that do not because maintaining social order is perceived as optimal—just as humans also do with each other.

Because our preliminary investigation of value alignment does not address artificial general intelligence, the question of which stories should be use to teach the agent remains open. Our crowdsourcing technique bypasses the question by acquiring a corpus of stories. In general, we believe the solution to value alignment in artificial general intelligences will be to use *all* stories associated with a given culture. Under the assumption that culture is reflected in the literature it creates, subversive or contrarian texts will be washed out by those that conform to social and cultural norms.

Conclusions

Value alignment is a property of an intelligent agent indicating that it can only pursue goals that are beneficial to humans. The preliminary work described in this paper seeks to use the implicit and explicit sociocultural knowledge encoded in stories to produce a value-aligned reward signal for reinforcement learning agents. This enables us to overcome one of the limitations of value alignment: that values cannot easily be exhaustively enumerated by a human author. In our current investigation of limited artificial agents, we show how crowdsourced narrative examples can be used to train an agent to act in a human-like fashion. Our technique is a step forward in achieving artificial agents that can pursue their own goals in a way that limits adverse effects that this may have on humans.

Even with value alignment, it may not be possible to prevent all harm to human beings, but we believe that an artificial intelligence that has been *encultured*—that is, has adopted the values implicit to a particular culture or society—will strive to avoid psychotic-appearing behavior except under the most extreme circumstances. As the use of artificial intelligence becomes more prevalent in our society, and as artificial intelligence becomes more capable, the consequences of their actions become more significant. 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.

Acknowledgements

This material is based upon work supported by the U.S. Defense Advanced Research Projects Agency (DARPA) under Grant #D11AP00270 and the Office of Naval Research (ONR) under Grant #N00014-14-1-0003.

References

Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Chambers, N., and Jurafsky, D. 2008. Unsupervised learning of narrative event chains. In *Proceedings of 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Elson, D. K., and McKeown, K. 2010. Automatic attribution of quoted speech in literary narrative. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*.

Elson, D. K. 2012. *Modeling Narrative Discourse*. Ph.D. Dissertation, Columbia University.

Finlayson, M. 2011. *Learning Narrative Structure from Annotated Folktales*. Ph.D. Dissertation, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

Gervás, P. 2009. Computational approaches to storytelling and creativity. *AI Magazine* 30(3):49–62.

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.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; Petersen, S.; Beattie, C.; Sadik, A.; Antonoglou, I.; King, H.; Kumaran, D.; Wierstra, D.; Legg, S.; and Hassabis, D. 2015. Human-level control through deep reinforcement learning. *Nature* 518:529–533.

Mueller, E. 2004. Understanding script-based stories using commonsense reasoning. *Cognitive Systems Research* 5(4):307–340.

Ng, A., and Russell, S. 2000. Algorithms for inverse reinforcement learning. In *Proceedings of the 17th International Conference On Machine Learning*.

O'Neill, B. C., and Riedl, M. O. 2014. Dramatis: A computational model of suspense. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*.

Riedl, M. O., and Bulitko, V. 2013. Interactive narrative: An intelligent systems approach. *AI Magazine* 34(1):67–77.

Riedl, M. O., and Young, R. M. 2010. Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research* 39:217–268.

Russell, S.; Dewey, D.; and Tegmark, M. 2015. Research priorities for robust and beneficial artificial intelligence. Technical report, Future of Life Institute.

Schank, R., and Abelson, R. 1977. *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures*. Lawrence Erlbaum Associates.

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. 2015. The value learning problem. Technical Report 2015-4, Machine Intelligence Research Institute.

Swanson, R., and Gordon, A. 2012. Say Anything: Using textual case-based reasoning to enable open-domain interactive storytelling. *ACM Transactions on Interactive Intelligent Systems* 2(3):16:1–16:35.

Winston, P. H. 2011. The strong story hypothesis and the directed perception hypothesis. In Langley, P., ed., *Advances in Cognitive Systems: Papers from the 2011 AAAI Fall Symposium (Technical Report FSS-11-01)*. AAAI Press.