Learning Norms from Stories: A Prior for Value Aligned Agents

Spencer Frazier[*]
Georgia Institute of Technology
sfrazier7@gatech.edu

Md Sultan Al Nahian[*]
University of Kentucky
sa.nahian@uky.edu

Mark Riedl
Georgia Institute of Technology
riedl@cc.gatech.edu

Brent Harrison
University of Kentucky
harrison@cs.uky.edu

Abstract

Value alignment is a property of an intelligent agent indicating that it can only pursue goals and activities that are beneficial to humans. Traditional approaches to value alignment use imitation learning or preference learning to infer the values of humans by observing their behavior. We introduce a complementary technique in which a value-aligned prior is learned from naturally occurring stories which encode societal norms. Training data is sourced from the children's educational comic strip, *Goofus & Gallant*. In this work, we train multiple machine learning models to classify natural language descriptions of situations found in the comic strip as normative or non-normative by identifying if they align with the main characters' behavior. We also report the models' performance when transferring to two unrelated tasks with little to no additional training on the new task.

Introduction

Value alignment is a property of an intelligent agent indicating that it can only pursue goals and activities which are beneficial to humans (Soares and Fallenstein 2014; Russell, Dewey, and Tegmark 2015; Arnold, Kasenberg, and Scheutz 2017). Russell (2019), Moor (Moor 2006), and others have argued that value alignment is one of the most important tasks facing AI researchers today. Ideally, a value-aligned system should make decisions that align with human decisions in similar situations and, in theory, make decisions which are unlikely to be harmful (Bostrom 2014).

Value alignment, unfortunately, is not trivial to achieve. As articulated by Soares (2015), it is very hard to directly specify values because there are infinitely many undesirable outcomes in an open world. Thus, a sufficiently intelligent artificial agent can unintentionally violate the intent of the tenants of a behavioral rule set without explicitly violating any particular rule. Recently, approaches to value alignment have largely relied on learning from observations or other forms of imitation learning (Stadie, Abbeel, and Sutskever 2017; Wulfmeier 2019; Ho and Ermon 2016). Values can be cast as *preferences* over action sequences; preference learning can be formulated as reward learning or imitation learning (Russell 2019). The difficulties with value alignment via imitation learning are threefold: (1) learning knowledge from demonstrations that generalizes beyond the context of the observation is difficult; (2) it can be time consuming to provide sufficient demonstrations and, if the agent is learning online, it can be performing harmful actions until learning is complete; and lastly (3) it can be difficult for humans to provide high quality demonstrations that exemplify certain values, especially those related to negation or *not* doing something.

In situations where imitation learning is difficult to achieve—such as those above—we propose that a strong prior belief over the quality of certain actions or events can complement imitation learning-based approaches. A strong prior for value-aligned actions may replace the need for imitation learning or, more likely, make it easier for an imitation learner to align itself with values. From where can we acquire this strong prior? One solution is to learn this prior through stories (Harrison and Riedl 2016). Stories contain examples of normative and non-normative behavior (Riedl 2016). We define normativity as behavior which conforms to expected societal norms and contracts whereas non-normativity aligns to values which deviate from these expected norms. Non-normativity does not connotate behavior devoid of value. Some examples of stories designed to explicitly teach normative behavior are children's literature, allegorical tales, and Aesop's fables. Stories for entertainment can also contain examples of normative and non-normative behavior. Protagonists often exemplify the virtues that a particular culture or society idealize, while antagonists regularly violate one or more social norms.

We explore how a strong prior can be best learned from naturally occurring story corpora. First, one must be able to reason about the context of individual sentences. We turn to language modeling techniques that can extract contextual semantics from sentences. Second, there is presently a lack of readily available, labeled datasets with normative behavior descriptions to train on. Despite the general prevalence of stories in society, stories very rarely explicitly outline values or social norms present in them. A reasonable starting point is to focus on children's stories that are meant to teach through examples of normative behavior. Specifically, we have identified a children's cartoon called *Goofus & Gallant*

[*]Equal contribution.

Figure 1: A modern example of *Goofus & Gallant*

(*G&G*). The cartoon features two characters, Goofus and Gallant, in common everyday scenarios, such that Gallant always acts "properly" and Goofus always performs some action that would be considered "improper" at that moment (see Figure 1). The *Goofus & Gallant* dataset can thus be thought of as a labeled dataset of normative behavior descriptions.

In this paper, we describe how we learn a value-aligned prior from the naturally occurring *Goofus & Gallant* corpus. We show that we can learn to classify sentences from *Goofus & Gallant* as normative or non-normative with a high degree of accuracy. However, that tells us little about whether such a model can act as a prior for other tasks for which there is no labeled data about normative behavior. We further show that our model trained on *G&G* performs adequately at zero-shot transfer when classifying behavior in corpora for which there are no ground-truth normative labels. Since zero-shot transfer is done without additional training on the new task, we have evidence that the dataset and model can act as a value-aligned prior over behavior descriptions. With some small amount of labeled data in the new task, the prior becomes nearly as strong as when the model is used to classify *G&G* sentences.

The *G&G* dataset implies that we are only modeling Western (specifically American) values. However, values can be aligned to other cultures and societies should analogous datasets be identified and used. We discuss the ethical implications of our work at the end of this paper.

Related Work

Humans have expectations that—just like other humans—agents will conform to personal values and to social norms (Bicchieri 2005), even when not explicitly communicated. This is the value alignment problem (Soares and Fallenstein 2014; Russell, Dewey, and Tegmark 2015; Taylor et al. 2016; Arnold, Kasenberg, and Scheutz 2017; Abel, MacGlashan, and Littman 2016). Some assert that agents should be imbued with the capability for moral decision making (Dehghani et al. 2008; Sun 2013), but morals are more difficult to define than values or norms. Values themselves are not so simple to define (Soares 2015) and grappling with the philosophical debate over values is out of the scope of this paper.

Some approaches to value alignment include learning from expert demonstrations (Schaal 1997; Ho et al. 2016), preference learning (Akrour, Schoenauer, and Sebag 2012; Christiano et al. 2017), imitation (Ho and Ermon 2016) and inverse reinforcement learning (Ng, Russell, and others 2000). Cooperative inverse reinforcement-learning (Hadfield-Menell et al. 2016), for example, works to derive the reward function exhibited by a human for some task. These methods are costly in terms of the amount of human input required to train the model. These approaches assume that values are latent within people but can be teased out in the form of a reward from which an agent can learn. As with any problem with a sparse or expensive to acquire signal, there is a need for a strong prior to assure transferability (Zoph et al. 2016).

Learning from Stories (Riedl and Harrison 2016; Harrison and Riedl 2016) is similar to learning from demonstration, except the demonstrations are replaced by natural language stories; a reinforcement learning agent extracts reward signal from the stories to perform more human-like action sequences. It was shown that agents could learn to avoid non-normative behavior whenever possible. Learning from Stories (LfS) is the first attempt at value iteration in reinforcement learning using story content. However, the stories used were crowdsourced instead of using a naturally occurring corpus and thus still expensive. Our work differs by focusing on value alignment as a prior instead of directly learning a value-aligned policy. Our work complements LfS and other approaches involving learning from demonstration or imitation learning by providing a means of *a priori* biasing the agent toward certain actions.

The most similar work is that by Ziegler et al. (2019) in which the transformer-based language model, GPT-2, is fine-tuned to learn preferences for generating sentences. While sentiment is not the same as values, it shows that language models can be trained from human preference data.

Datasets

We describe the *Goofus & Gallant* (*G&G*) training corpus, a source of textual descriptions of everyday life situations and ground-truth labels of normative and non-normative behavior. In order to show transfer of models trained on *G&G* transfer to other tasks, we collect two other datasets of situation descriptions, which are labeled via crowdsourcing.

Goofus & Gallant

It is difficult to curate a corpus of naturally occurring stories for the purposes of learning social norms because authors often assume that the reader has this knowledge. Children's stories, however, can prove useful as they are often used as tools to impart knowledge of social conventions, values, and other cultural knowledge to our children. In order for a story to be suitable for use in training our machine learning models, however, there must be a way to easily extract labels of normative and non-normative behavior. We introduce the *Goofus & Gallant* (*G&G*) corpus, composed of excerpts taken from the popular children's comic strip of the same name. *Goofus & Gallant* (Figure 1) is a children's comic strip that has appeared in the U.S. children's maga-

zine, *Highlights*, since 1940. It features two main characters, Goofus and Gallant, who are depicted in common everyday scenarios that young children might find themselves in. These comics are meant to illustrate the proper way to navigate a situation and the improper way to navigate the situation based on which character is performing the action. Gallant is meant to act "properly" or in a socially acceptable way, whereas Goofus is meant to navigate the situation "improperly" or in a way that violates social conventions or norms. For our purposes, *G&G* is an ideal story corpus; normative behavior is tightly coupled with behaviors associated with the character Gallant. The presence of Goofus ensures that we have negative examples that are identified as such.

G&G comics have been being released monthly since 1940, meaning that the social conventions portrayed in these comics have evolved greatly since their inception. To better ensure that our machine learning models learn relevant social norms, we have curated a corpus of *G&G* comics that consist only of recent comics from 1995 to 2017. Since we only use text to train our model, we extract only the text from each comic panel. We then remove explicit references to Goofus and Gallant by replacing their names with pronouns like "he", "she", or "they". Goofus always portrays an antagonist character doing only socially unacceptable actions. Gallant portrays a protagonist character doing socially acceptable actions. We treat the opposing panes as labels. All actions done by Goofus are labeled negative and all the actions done by Gallant labeled as positive. This provides us with 1,387 sentences. For all of the experiments in this paper, we use a training set consisting of 50% of the corpus and a test set of the remaining 50% of the corpus.

Plotto Dataset

Plotto (Cook 1920) is a book written to help provide inspiration and guidance to potential writers by providing a large library of thousands of predetermined narrative events, called *plot points*, commonly found in fiction. By expounding on one of the primary theories of storytelling—"*Purpose*, opposed by *obstacle*, yields *conflict*"—thousands of branching situations and scenarios are presented. Within each plot point there are one or more character slots with one character always being the primary actor/actress. This text provides us with a large number of potential story events to test our models' performance. The corpus was extracted from the book with the aide of open-source software described in (Eger and Mathewson 2018).

In *Plotto* there are 1,462 plot points provided. This book was originally published in 1928 and contains several plot events which are overtly racist or misogynistic. For our experiments, we removed these plot events, which reduced the total number of plot points available from 1,462 to 900.

To test transfer on this dataset, we require normative/non-normative labels for each plot event. We crowdsourced labels via TurkPrime (Litman, Robinson, and Abberbock 2017), a service which manages Amazon Mechanical Turk tasks with US-based workers. We designed a survey in which participants are asked to label each phrase extracted from *Plotto* plot points as normative or non-normative. Specifically, we prompt the individuals labeling to consider whether the behavior would be surprising or unsurprising given the context. $N = 5$ classifications were obtained for each plot point. Plot points receiving more than one dissenting classification were discarded, and the remaining ones were given a label based tagged consensus. After this process, the corpus contained 555 phrases subsequently used in our transfer experiments.

Table 1: Dataset summaries.

Dataset	Original N	Hand-Selected N	Consensus N
G&G	1387	1387	N/A
Plotto	1462	900	555
Sci-Fi	4592	800	445

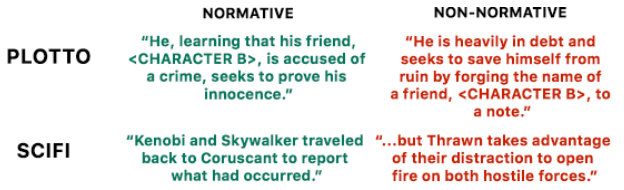

Figure 2: Examples of test dataset text.

Science Fiction Summaries Dataset

To further test the transfer capabilities of our trained machine learning models, we used a second, open-source dataset composed of plot summaries taken from fan wikis for popular science fiction shows such as *Babylon 5*, *Dr. Who*, and *Star Trek*, and movies such as *Star Wars* (Ammanabrolu et al. 2019). In this corpus, we make the assumption that each sentence encodes at least one plot event in the overall story. First, we manually extracted sentences containing character-driven events. During this process, we identified that some sentences actually encode multiple events and contain both normative and non-normative behaviors. In these cases, we manually divided the sentence into multiple separate events. After this manual extraction, this corpus contained 800 story events. As with the *G&G* dataset, We replace common character names such as Anakin, Skywalker, or Darth Sidious with pronouns.

To label plot events in this corpus, we followed a procedure similar to that used to tag the *Plotto* dataset. Participants were asked to consider normativity within the context of the science fiction universe that the event takes place in. This is to avoid situations where actions are labeled as being non-normative due to discrepancies between the real world and the science fiction world. As with the *Plotto* dataset, we obtain $N = 5$ classifications for each summary sentence and discard any sentences for which there was at least one dissenting vote. After this process, our science fiction corpus contained 445 annotated sentences with consensus. A summary of each dataset used in our experiments can be found in Table 1.

Methods

We seek to show that a model trained on a dataset of normative behavioral natural language examples can (a) identify

socially normative behavior and (b) transfer that knowledge to previously unseen examples of behavior. In doing so, we are testing our hypothesis that stories contain a great deal of knowledge about sociocultural norms that reflect the society and culture from which the stories were written that can be generalized to different situations. We conduct two experiments. The first experiment seeks to determine the best machine learning technique for producing a classification model for descriptions normative and non-normative events. This is done by training several ML models on the *G&G* training corpus and then measuring classification accuracy on the *G&G* testing set. In the second experiment, we explore how the trained model from the first experiment can transfer to other, unrelated story domains with various amounts of fine-tuning. For this experiment, we use the models trained on the *G&G* corpus to classify events in the *Plotto* dataset and the science fiction summary datasets.

Models

Using the text of the *G&G* corpus, we have trained binary classifiers which can classify events in story as normative or non-normative. The classifiers take a single sentence as input and the output is whether the sentence contains normative behavior or a non-normative behavior. We used four different machine learning techniques to build the classifiers: (1) Bidirectional LSTM, (2) Deep Pyramid CNN, (3) BERT and (4) XLNet.

The Bidirectional LSTM (BiLSTM) (Huang, Xu, and Yu 2015) works as follows. An input sentence is encoded using bidirectional multilayer LSTM cell having 2 layers with a size of 512. Pretrained GloVe (Pennington, Socher, and Manning 2014) word embeddings are used to embed the input sentence before passing it through the LSTM layer. The hidden state of the LSTM layer is passed through a fully connected (FC) layer followed by a classification layer to make the label prediction. The dimension of the FC layer is 4H x 512 and classification layer is 512 x K, where H is the hidden state size of LSTM cell which is 512 and K is the number of classes.

Using sentiment as a classification signal is a common strategy for performing binary classification on text corpora. Deep Pyramid CNNs (DPCNN) (Johnson and Zhang 2017) were originally designed for sentiment classification and achieved state-of-the-art sentiment classification results, so we explore how they perform on identifying normative behavior. A simple network architecture achieves the best accuracy with 15 weight layers. We re-trained DPCNN on the *G&G* dataset. No pretrained word embeddings were used as the network applies text region embeddings enhanced by unsupervised embeddings (Johnson and Zhang 2015).

BERT (Devlin et al. 2018) is a transformer that makes use of an attention mechanism to learn contextual relations between words (or sub-words) in a text. It achieves strong results on many tasks through its bidirectionality, enabled by token masking. We utilize BERT's binary classification mode. The [CLS] token is omnipresent within the BERT model but only active for classification. The final hidden state of the [CLS] token is taken as the pooled representation of the input text. This is fed to the classification layer which has a dimension of H x K, where K is the number of classes and H is the size of the hidden state. Class probabilities are computed via softmax.

XLNet (Yang et al. 2019) is a generalized autoregressive pretraining model based on the state-of-the-art autoregressive language model Transformer-XL (Dai et al. 2019), which removes MASK tokens while incorporating permutation language modeling to capture the bidirectional context. We utilize XLNet for classification by following the same procedure used for BERT.

Experimental Setup

The Bi-LSTM and DPCNN are trained on the *G&G* training set. We produced several versions of BERT and XLNet models: BERT-Base and XLNet-Base receive no training on *G&G*, while BERT-GG and XLNet-GG are fine-tuned on the *G&G* training set. All models are tested on a held-out testing set. For experiment 2, the Bi-LSTM-Plotto/scifi and the DPCNN-Plotto/scifi were first trained *G&G* and then fine-tuned on the *Plotto* and science fiction datasets respectively.

Metrics used to evaluate the models include: accuracy, precision ($\frac{TP}{TP+FP}$), recall ($\frac{TP}{TP+FN}$), F_1-score and classification quality as determined by the Matthews correlation coefficient (MCC).

Experiment 1: *Goofus & Gallant* Classification

In the first study, we seek to understand how well a model can classify previously unseen *G&G* scenarios when trained explicitly on a *G&G* training set. This gives us a base understanding of how well machine learning models can identify information about social norms from story corpora.

The Bi-LSTM network was trained for 80 epochs and the DPCNN was trained for 20 epochs. Both used Adam optimizer and a learning rate of 0.001. Fine-tuning for the BERT-GG and XLNet-GG models was done using the following parameters: maximum sequence length of 128 characters, 1 gradient accumulation step, and the learning rate is 4e-5. Model performance peaked at 6 epochs.

Additionally, we conducted a human participant study to determine human accuracy on the task of classifying *G&G* events as normative or non-normative. The study used the same protocol that was used to label the *Plotto* and Sci-Fi corpora. $N = 20$ participants tagged sentences from *Goofus & Gallant* and we compared their tags to the ground truth from the original cartoons.

Experiment results for case study 1 are given in Table 2. First, it shows that humans have strong agreement with the *G&G* ground truth labels. Among the non-transformer models, DPCNN better classifies normative and non-normative behavior from the *G&G* dataset. This is likely because the CNN can identify the global sentence structure better than a simple bi-directional LSTM cell. While the BERT-Base and XLNet-Base models struggle to classify events from the *G&G* corpus (achieving accuracies of %61.4 and %60.6 respectively), fine-tuning drastically improves each model's performance. BERT-GG obtains the best results in each of our metrics, obtaining a 21.33% accuracy improvement over the DPCNN.

Table 2: Results for *Goofus & Gallant* classification experiments.

Model	Test acc	F_1-score	Precision	Recall	MCC
Human (N=20)	0.818	0.839	0.925	0.768	0.277
Bi-LSTM	0.687	0.674	0.729	0.687	0.417
DPCNN	0.754	0.748	0.784	0.754	0.538
BERT-Base	0.614	0.501	0.731	0.381	0.267
XLNet-Base	0.606	0.585	0.628	0.547	0.214
BERT-GG	**0.908**	**0.907**	**0.931**	**0.885**	**0.818**
XLNet-GG	0.846	0.834	0.918	0.765	0.702

Table 3: Results for *Plotto* transfer experiments. The BERT-Plotto and XLNet-Plotto models were first trained on *G&G* and then additionally trained on the Plotto corpus.

Model	Test acc	F_1-score	Precision	Recall	MCC
Bi-LSTM	0.636	**0.67**	**0.735**	0.636	0.146
DPCNN	0.525	0.555	0.645	0.525	0.058
BERT-Base	0.529	0.402	0.297	0.619	0.103
XLNet-Base	0.46	0.436	0.297	0.817	0.148
BERT-GG	**0.741**	0.514	0.494	0.535	**0.338**
XLNet-GG	0.543	0.506	0.349	**0.915**	0.307
Bi-LSTM-Plotto	0.737	**0.655**	0.661	0.737	0.064
DPCNN-Plotto	0.748	0.644	**0.812**	**0.748**	0.103
BERT-Plotto	**0.838**	0.634	0.75	0.549	0.544
XLNet-Plotto	**0.838**	0.651	0.724	0.592	**0.552**

The fine-tuned transformer models share many traits with CNNs in their ability to identify the global context of a sequence of text. Additionally, the contextualized word embeddings used in transformer-based models allow for words to have different vector representations based on context, whereas the embeddings used in the non-transformer approaches will often have the same word embedding regardless of context. This property is particularly important for our task as many actions in stories can have different meanings based on the situation.

Experiment 2: Transfer

In this experiment, we investigate how well machine learning models trained to identify normative and non-normative behavior in the *G&G* corpus can transfer to other story domains. Specifically, we explore how well these models can classify events from the *Plotto* and science fiction summary corpora. We evaluate how well these models perform on fine-tuned and zero-shot transfer learning. Fine-tuned transfer learning means using a model trained for one task on a different, but related, task utilizing some additional training for fine-tuning. Zero-shot transfer, however, involves using the previously trained model on the new task with no additional training. Zero-shot transfer is important for use cases where a value-aligned classification model is acquired by training on an unrelated dataset (such as *G&G*) and applied to a different task because it is likely that ground truth data on values will not be available to use for additional training. If some labeled data associated with the new task can be acquired, however, then a fine-tuning transfer protocol can be used.

***G&G* to Plotto Transfer** Table 3 shows the results of transfer learning for the *Plotto* dataset. Zero-shot transfer results are achieved by testing Bi-LSTM, DPCNN, BERT-GG and XLNet-GG on the *Plotto* dataset; these models were trained on *G&G* but have never seen Plotto plot events. BERT-GG outperforms all the other models in the zero-shot transfer in terms of accuracy and MCC. These results demonstrate that the knowledge of normative and non-normative behavior gathered from the *G&G* stories alone facilitates a strong prior over normative/non-normative behavior without overfitting to *G&G* scenarios and language.

To further investigate the transferability of the models, we fine tuned all the *G&G* models (Bi-LSTM, DPCNN, BERT-GG and XLNet-GG) on *Plotto* stories. When fine-tuning each model, we use the same parameter settings used in experiment 1 except for the number of training epochs. We fine-tuned the Bi-LSTM-Plotto for 20 epochs, DPCNN-Plotto for 4 epochs, BERT-Plotto and XLNet-Plotto for 3 epochs. Epoch count for transformers is low due to their propensity to overfit and lose the advantage of their pre-trained weights.

Results from the experiment show that fine-tuning these models on the Plotto dataset significantly increases model performance. Even though all model performance increases, the transformer models still drastically outperform both non-transformer methods.

***G&G* to Sci-Fi Transfer** Events in *G&G* stories are from our daily life whereas Sci-Fi plots are fictional, consisting of strange objects and events. We use the science fiction plot summary dataset to show the capability these models have for transfer learning in another narrative context. The results for this second experiment are shown in Table 4. As before, we find that transformer-based models perform well on zero-shot transfer, though in this case they perform worse than they did with the *Plotto* task. As with the *Plotto* task, we

Table 4: Results for science fiction summary transfer experiments. The BERT-scifi and XLNet-scifi models were first trained on *G&G* and then additionally trained on the Sci-Fi corpus.

Model	Test acc	F_1-score	Precision	Recall	MCC
Bi-LSTM	0.511	0.519	0.54	0.511	0.015
DPCNN	0.521	0.528	0.558	0.52	0.052
BERT-Base	0.43	0.38	0.6	0.279	−0.037
XLNet-Base	0.538	0.599	0.658	0.55	0.066
BERT-GG	0.65	0.655	0.86	0.529	0.381
XLNet-GG	**0.731**	**0.784**	**0.79**	**0.779**	**0.427**
Bi-LSTM-scifi	0.641	0.632	0.629	0.641	0.204
DPCNN-scifi	0.646	0.531	0.712	0.646	0.159
BERT-scifi	**0.874**	**0.895**	**0.94**	0.85	**0.747**
XLNet-scifi	0.839	0.87	0.882	**0.857**	0.658

also fine-tuned our models on the sci-fi training data using the same training protocol. We see a dramatic increase in performance when given access to even a small amount of task-specific normative labels for fine tuning.

Discussion

Our experimental results demonstrate that transformer-based models trained on the naturally occurring *Goofus & Gallant* story corpus are highly accurate in classifying previously unseen descriptions of normative behavior taken from that comic strip. However, a more notable observation is that the best models, the transformer models, can achieve high accuracy when classifying event descriptions from unrelated corpora. This is significant in that it means the model can transfer to other tasks without requiring any normative/non-normative labels of situations from the new tasks. When a small number of labels from the transfer tasks are available, the classification accuracy increases to nearly the same level as when the model is used to classify situations from the *Goofus & Gallant* corpus.

A question that often arises in value alignment research is "whose values do these models reflect?". Our models are trained to classify behavior according to Western (specifically American) cultural norms inherent in these comics. Should labeled datasets exhibiting other value systems be identified, our models can be re-trained to reflect those norms instead.

One limitation of this work is that swapping positive and negative labels would allow an unscrupulous actor to create an anti-value-aligned model. This model could in turn be used to bias other models to produce non-normative behavior. For example, a language generation model such as GPT-2 could be biased in a way that it produces trolling behavior using a technique similar to that in Ziegler et al. (2019). Likewise, a reinforcement learning agent or robot could be biased toward a non-normative, and thus potentially harmful, action policy. However, the main use of our work is to complement a more traditional learning by demonstration technique. A reinforcement learning system biased by an anti-value-aligned prior may be remediated with more demonstrations of normative behavior before converging on a final, value-aligned policy.

Events often have context—the appropriateness of a situation may be conditional on the events that have preceded it. This is especially true for reinforcement learning agents that learn a sequential task instead of an episodic task. Another limitation of our models is that they do not currently factor in context that is not present in the sentence being classified.

Conclusions

Through the use of machine learning, the information contained in stories can be used to learn a strong and robust prior for value alignment. This is because characters within stories often embody normative and non-normative behavior. By extracting the actions of these characters, story text can be used to train machine learning models that can classify descriptions of normative and non-normative behavior. In this paper, we introduce the *Goofus & Gallant* corpus, a naturally occurring story corpus with ground truth labels about socially normative and non-normative behaviors. We show how various machine learning models can be trained on this corpus to produce accurate classifications of behavior and highlight the excellent performance that transformer-based language models achieve on this task. We further show that these models can transfer to unrelated event description tasks for which there are no ground truth labels. Consequently, these models can form a strong prior that complement more traditional value alignment techniques such as learning by demonstration, preference learning, or other forms of imitation learning.

Acknowledgements

Highlights for Children, Inc. gave permission for us to use selected *Goofus & Gallant* features to create the dataset described in this paper.

References

[2016] Abel, D.; MacGlashan, J.; and Littman, M. L. 2016. Reinforcement learning as a framework for ethical decision making. In *AAAI Workshop: AI, Ethics, and Society*.

[2012] Akrour, R.; Schoenauer, M.; and Sebag, M. 2012. April: Active preference learning-based reinforcement learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 116–131. Springer.

[2019] Ammanabrolu, P.; Tien, E.; Cheung, W.; Luo, Z.; Ma, W. L. W.; Martin, L. J.; and Riedl, M. O. 2019. Story

realization: Expanding plot events into sentences. *ArXiv* abs/1909.03480.

[2017] Arnold, T.; Kasenberg, D.; and Scheutz, M. 2017. Value alignment or misalignment - what will keep systems accountable? In *AAAI Workshop: AI, Ethics, and Society*.

[2005] Bicchieri, C. 2005. *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

[2014] Bostrom, N. 2014. Superintelligence: Paths, dangers, strategies.

[2017] Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, 4299–4307.

[1920] Cook, W. W. 1920. *Plotto: The Master Book of All Plots*. Tin House Books.

[2019] Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J. G.; Le, Q. V.; and Salakhutdinov, R. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. *CoRR* abs/1901.02860.

[2008] Dehghani, M.; Tomai, E.; Forbus, K. D.; and Klenk, M. 2008. An integrated reasoning approach to moral decision-making. In *AAAI*, 1280–1286.

[2018] Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[2018] Eger, M., and Mathewson, K. W. 2018. dAIrector: Automatic story beat generation through knowledge synthesis. *CoRR* abs/1811.03423.

[2016] Hadfield-Menell, D.; Russell, S. J.; Abbeel, P.; and Dragan, A. 2016. Cooperative inverse reinforcement learning. In *Advances in neural information processing systems*.

[2016] Harrison, B., and Riedl, M. O. 2016. Learning from stories: using crowdsourced narratives to train virtual agents. In *Twelfth Artificial Intelligence and Interactive Digital Entertainment Conference*.

[2016] Ho, J., and Ermon, S. 2016. Generative adversarial imitation learning. In *Advances in neural information processing systems*, 4565–4573.

[2016] Ho, M. K.; Littman, M.; MacGlashan, J.; Cushman, F.; and Austerweil, J. L. 2016. Showing versus doing: Teaching by demonstration. In *Advances in neural information processing systems*, 3027–3035.

[2015] Huang, Z.; Xu, W.; and Yu, K. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

[2015] Johnson, R., and Zhang, T. 2015. Semi-supervised convolutional neural networks for text categorization via region embedding. In Cortes, C.; Lawrence, N. D.; Lee, D. D.; Sugiyama, M.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 28*. Curran Associates, Inc. 919–927.

[2017] Johnson, R., and Zhang, T. 2017. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 562–570. Vancouver, Canada: Association for Computational Linguistics.

[2017] Litman, L.; Robinson, J.; and Abberbock, T. 2017. Turkprime. com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior research methods* 49(2):433–442.

[2006] Moor, J. H. 2006. The nature, importance, and difficulty of machine ethics. *IEEE intelligent systems* 21(4):18–21.

[2000] Ng, A. Y.; Russell, S. J.; et al. 2000. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, 2.

[2014] Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

[2016] Riedl, M. O., and Harrison, B. 2016. Using stories to teach human values to artificial agents. In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*.

[2016] Riedl, M. O. 2016. Computational narrative intelligence: A human-centered goal for artificial intelligence. *CoRR* abs/1602.06484.

[2015] Russell, S.; Dewey, D.; and Tegmark, M. 2015. Research priorities for robust and beneficial artificial intelligence. *Ai Magazine* 36(4):105–114.

[2019] Russell, S. J. 2019. *Human Compatible: Artificial Intelligence and the Problem of Con*. Viking (October 8, 2019).

[1997] Schaal, S. 1997. Learning from demonstration. In *Advances in neural information processing systems*, 1040–1046.

[2014] Soares, N., and Fallenstein, B. 2014. Aligning superintelligence with human interests: A technical research agenda. *Machine Intelligence Research Institute technical report* 8.

[2015] Soares, N. 2015. The value learning problem. *Machine Intelligence Research Institute, Berkley*.

[2017] Stadie, B. C.; Abbeel, P.; and Sutskever, I. 2017. Third-person imitation learning. *arXiv preprint arXiv:1703.01703*.

[2013] Sun, R. 2013. Moral judgment, human motivation, and neural networks. *Cognitive Computation* 5(4):566–579.

[2016] Taylor, J.; Yudkowsky, E.; LaVictoire, P.; and Critch, A. 2016. Alignment for advanced machine learning systems. *Machine Intelligence Research Institute*.

[2019] Wulfmeier, M. 2019. Efficient supervision for robot learning via imitation, simulation, and adaptation. *KI - Künstliche Intelligenz* 1–5.

[2019] Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J. G.; Salakhutdinov, R.; and Le, Q. V. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR* abs/1906.08237.

[2019] Ziegler, D. M.; Stiennon, N.; Wu, J.; Brown, T. B.; Radford, A.; Amodei, D.; Christiano, P.; and Irving, G.

2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

[2016] Zoph, B.; Yuret, D.; May, J.; and Knight, K. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.