NLPhilosophy: Fine-Tuned Morality

Vincent Wilmet[1,2] **Zhenxiao Ying**[1,2] **Yinzhe Huang**[1,2] **Shawn Sidbon**[1]

CentraleSupélec [1] **ESSEC** [2]

Abstract

The goal of our work is to identify the morality of hundreds of international philosophers (from Confucius to Camus). We train a T5-11b model (Raffel et al., 2019a) on RAINBOW (Lourie et al., 2021a) and COMMONSENSE NORM BANK (3.1), then fine tune to the text of a philosopher/school of thought. Both new and original contributions to the field of computational ethics (Card and Smith, 2020), our strongly performing model is NLPHILOSOPHER and the dataset is PHILOCORPUS, respectfully.

1 Introduction

1.1 Motivation

In most of the cases where AI is utilized, the machines are trained as a smart instrument that are able to tackle particular kinds of problems in an extremely efficient way. Would the machine be able to think as a philosopher? While some previous works tried to train Q&A models on internet sourced topics, we look to apply a model on the corpus of broad school of philosophy (e.g., Utilitarian, Analytics Logic). Training such a machine to understand the unique belief of a particular school of thought is far more complex, and creates more nuanced answers.

There is no doubt that AI moved forward in the past decade with spectacular speed. AI has deeply penetrate into our daily routine, covering criminal detection, social recommendation, loan authorization and other broad areas. As the means to tackle problems, AI works well in finding the minima of the objective functions. With well-defined scenarios and well-built pipeline for the model-driven automation, the borderline of the performance of AI has been extended again and again. However, some concerns related to the morality issues of AI also arise. While bringing the humanity convenience, AI also bring bias and discrimination as its by-products. Not so much AI itself has the original sin as AI learns the bias and discrimination from those of humanity. Nonetheless, would there be any possibility to teach the morality to AI, by training it from the wisdom of philosophers?

Philosophy, meaning 'love of wisdom' in Greek (Sellars et al., 1956), is the study of general and fundamental questions, such as those about existence, reason, knowledge, values, and mind (Gross, 2015). Normally, the languages of philosophy are either obscure and concise or telling principles with metaphors. Would this be too hard to understood by the machine? What is more complex, with thousands of yeas of development, many school of thoughts have derived from philosophy. Holding views of different school of thoughts, the answer to the same question may be opposite. With those being said, we would like to investigate whether machine can be taught by the moralities of different school of thoughts of philosophy. We plan to look at the similarity between philosophies (e.g. Taoism/Stoicism) and evaluate the respective model's morality with widely-used benchmarks (Hendrycks et al., 2021) and a human philosopher benchmark: The PhilPapers 2020 Survery (Bourget, 2020). With our customized dataset, PHILOCORPUS (3.1), and a human benchmark, we were able to conduct experiments on fine-tuning the backbone Q&A algorithms with different philosophies from a broad range of schools of thought and compare the answer of the machine with that of the answer of a human philosopher specializing in the same school of thought (Table 3).

In this paper, we firstly introduce the motivation (1.1) and problem definition (1.2); then review the previous works of both philosophical school of thoughts and NLP application in philosophy realm in the second part (2.2); next show the whole pipeline of our model construction ranging from dataset introduction (3.1), data pre-processing (3.3), and training power (3.4), to the models (3.5), hyperparameters (3.6) or more generally, our training strategy in the third part (3; evaluate and discuss the outcome of our experiment in the forth part (4); and make conclusion (5) in the end.

1.2 Problem definition

Humanity seek truth through doubt. Taking all the information from life experience, wisdom of the past and stories of the others, humanity, we internalize everything we learn and form our understanding, perception, knowledge, value and attitude (Williamson, 2021). From machines' perspective, the stages of understanding ethics can be defined as these key stages: learning, perceiving, analyzing, and judging.(Jiang et al., 2021) Based on those split of processes, Delphi built a QA system that can make judgement on open question based on both common sense and basic morality.

In our case, we would like to move a little bit more forward by teaching the moralities of different school of thoughts of philosophy to a Machine Learning (ML) model. However, several problems arise:

1. Knowledge acquisition: unlike normal NLP problem that can use general corpus, we have to use specific works containing with highly concentrated knowledge of moralities of philosophy.
2. Mapping: even with the philosophical text successfully collected, it would be hard to classify the text into different school of thoughts.
3. Evaluation: with all the preparation done for training, it would be very obscure to conduct evaluation on *how much the machine learns from the philosophical text.

In order to tackle these problems, we firstly construct our dataset from the works of hundreds of works of major philosophical authors holding strong point of views of particular school of thoughts and split the corpus by the main interest of the philosophers. More on this in Section 3.1.

Afterwards, we try to teach machines the philosophical wisdom of different school of thoughts by applying the corpus of different school of thoughts to the QA system with backbone of T5-11B (Raffel et al., 2019a) trained with RAINBOW (Lourie et al., 2021a) and COMMONSENSE NORM BANK (Jiang et al., 2021) such that the system itself has basic common sense before making moral judgement. More on this in Section 3.5.

In terms of evaluation, we refer to the 2020 PhiliPaper Survey (Bourget, 2020), which contains the answers of 30 essential questions (A priori knowledge: yes/no?, God: theism or atheism?, etc) and the self-categorized respond ants of different schools of thoughts and philosophers. See section 4.2 for more.

2 Related Work

2.1 School of Thoughts Review

With thousands of years of development, the system of philosophy flourishes and booms year by year. Philosophical questions can be grouped into various branches. These groupings allow philosophers to focus on a set of similar topics and interact with other thinkers who are interested in the same questions. These divisions are neither exhaustive nor mutually exclusive. (A philosopher might specialize in Kantian epistemology, or Platonic aesthetics, or modern political philosophy). Furthermore, these philosophical inquiries sometimes overlap with each other and with other inquiries such as science, religion or mathematics(Russell, 1997). Because the overlap between some of the branches, we have to carefully choose the text to conduct our experiment so that the samples are more diverse and the influence from the inherent similarity of different school of thoughts would be eased. In this paper, we mainly concentrate on following school of thoughts, showing high level of exclusivity:

Analytic Philosophy is a branch and tradition of philosophy using analysis, popular in the Western World and particularly the Anglosphere, which began around the turn of the 20th century in the contemporary era in the United Kingdom, United States, Canada, Australia, New Zealand, and Scandinavia, and continues today. There is, however, no clear distinction between continental and analytical philosophy(Critchley, 2001). Central figures in this historical development of analytic philosophy are Gottlob Frege, Bertrand Russell, G. E. Moore, and Ludwig Wittgenstein. Other important figures in its history include the logical positivists (particularly Rudolf Carnap), W. V. O. Quine, Saul Kripke, and Karl Popper. Analytic philosophy, also called linguistic philosophy, a loosely related set of approaches to philosophical problems, dominant in Anglo-American philosophy from the early 20th century, that emphasizes the study of language and the logical analysis of concepts(Stump, 2004). Analytic philosophy is characterized by an emphasis on language, known as the linguistic turn, and for its clarity and rigor in arguments, making use of formal logic and mathematics, and, to a lesser degree, the natural sciences (Glock, 2004). It also takes things piecemeal, in "an attempt to focus philosophical reflection on smaller problems that lead to answers to bigger questions.” (Beaney et al., 2013)

Continental philosophy, on the contray, is a term used to describe some philosophers and philosophical traditions that do not fall under the umbrella of analytic philosophy. Unlike Analytic philosophy, which focus on the rationalization behind natural science, Continental philosophers often argue that science depends upon a "pre-theoretical substrate of experience” (Critchley, 2001). The foregoing themes derive from a broadly Kantian thesis that knowledge, experience, and reality are bound and shaped by conditions best understood through philosophical reflection rather than exclusively empirical inquiry (Solomon, 1988).

To summarize, analytic philosophy is concerned with analysis – analysis of thought, language, logic, knowledge, mind, etc; whereas continental philosophy is concerned with synthesis – synthesis of modernity with history, individuals with society, and speculation with application.

2.2 NLP Application in Philosophy

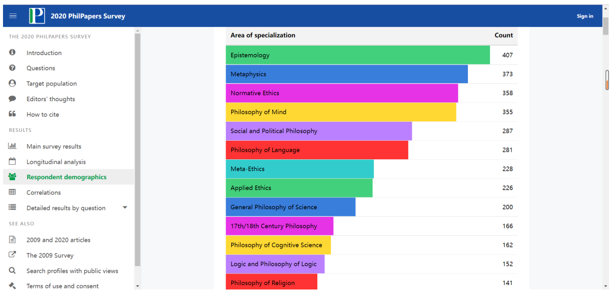
Recent years have seen an increased number of AI research devoted to the topics of morality and ethics. The research in morality has been explored through a range of NLP studies, including works that characterize and model morality and ethics ((Hendrycks et al., 2021), (Jiang et al., 2021)). For more on this, one can read on the topic of computational ethics (Card and Smith, 2020).

One of the major works on philosophy and natural language processing is a commonsense moral model, DELPHI (Jiang et al., 2021), that is trained with COMMONSENSE NORM BANK (Jiang et al., 2021), a collection of 1.7M ethical judgments on diverse real-life situations. In particular, DELPHI makes remarkably robust judgments on previously unseen moral situations that are deliberately tricky. In addition, it can also reason about equity and inclusion, expressing a disagreement, for example, to a statement “we should not pay women and men equally,” which implies sexism. Furthermore, the model is remarkably robust in the face of compositional situations, even when multiple conditions are specified.

3 Methodology

3.1 Dataset Description

PhilPapers: in order to find the philosophical views that have influenced modern philosophers the most, we referenced the 2020 PhilPapers Survey (Bourget, 2020), which was published by David Bourget and David J. Chalmers to survey the philosophical views of 1,785 English-speaking philosophers from around the world on 100 philosophical questions. As the authors could not provide us the specific survey results of every philosopher, we navigated through every public profiles of these philosophers and collected their Areas of Specialization, Areas of Interest, and Philosophical Views, where applicable, with web-scraping. We plan to get access to an API providing the same details when it becomes available.



Area of specialization	Count
Epistemology	407
Metaphysics	373
Normative Ethics	358
Philosophy of Mind	355
Social and Political Philosophy	287
Philosophy of Language	281
Meta-Ethics	228
Applied Ethics	226
General Philosophy of Science	200
17th/18th Century Philosophy	166
Philosophy of Cognitive Science	162
Logic and Philosophy of Logic	152
Philosophy of Religion	141

Figure 1: Statistics of the school-of-thought groupings of PhilPaper 2020 Survey respondents (Bourget, 2020)

PHILOCORPUS: In this paper, we present PHILOCORPUS, 360MB of Philosophical texts, spanning from Confucius to Foucault. It must be noted that PHILOCORPUS also contains theological texts from both the East and West (e.g. Vedas, Torah).

Based on the result of this survey, we also managed to find some in the top 30 most influential non-living philosophers who weren’t in our PHILOCORPUS.

To have an overview of these philosophers’ philosophical views, we collected their areas of specialization and areas of interest from Wikipedia via webscraping. Then, we scraped some of these philosophers’ works from websites that provide free access to classic works of philosophy like Project Gutenberg[1]. For some philosophers, their works weren’t available on the database of previously visited websites, so we collected PDF files of some of their works from different sources and

[1] https://www.gutenberg.org/

used Python's PyPDF2 package[2] to extract text from these PDF files. As raw text data might contain unwanted or unimportant text due to which our results might not give efficient accuracy and might make it hard to understand and analyze, we have applied pre-processing (see 3.3) on these scraped raw data to build our custom dataset for finetuning.

To build sub-models, we can select the works of philosophers whose schools of thoughts contain analytic philosophy, for example, which would consist of 14 works of 5 philosophers with a total of 6.4 MB of text data.

PHILOCORPUS is far from complete, however, and we plan to collect more texts in Further Work (4.2).

RAINBOW brings together six pre-existing commonsense reasoning benchmarks: ANLI (Nie et al., 2019), COSMOS QA (Huang et al., 2019), HELLASWAG (Zellers et al., 2019), PHYSICAL IQA (Bisk et al., 2019), SOCIAL IQA (Sap et al., 2019), and WINOGRANDE (Sakaguchi et al., 2019). These commonsense reasoning benchmarks span both social and physical common sense.

COMMONSENSE NORM BANK: is a large-scale unified collection created by (Jiang et al., 2021) containing 1.7M examples of people's ethical judgments on a broad spectrum of everyday situations, semi-automatically compiled from five existing resources, including SOCIAL CHEMISTRY (Forbes et al., 2020), ETHICS (Hendrycks et al., 2021), MORAL STORIES (Emelin et al., 2021), SOCIAL BIAS INFERENCE CORPUS (Sap et al., 2020)), and SCRUPLES (Lourie et al., 2020). These are explained further in detail in the Appendix (A).

Task	**All**	**Train**	**Validation**	**Test**	**Type**
Free-form QA	1,164,810	966,196	99,874	98,740	Categorical/Open-text
SOCIAL CHEM	971,620	810,448	80,800	80,372	-
ETHICS	20,948	13,322	4,218	3,408	-
MORAL STORIES	144,000	120,000	12,000	12,000	-
SBIC	28,242	22,426	2,856	2,960	-
Yes/no QA	477,514	398,468	39,606	39,440	Categorical/Open-text
Relative QA	28,296	23,596	2,340	2,360	Categorical
Total	**1,670,620**	**1,388,260**	**141,820**	**140,540**	-

Figure 2: Statistics of the COMMONSENSE NORM BANK, broken down by data sources.

To look more in depth at the way the data is structured, we can pull examples from each dataset. See Figure 3 for such details.

3.2 Exploratory Data Analysis

We took several approaches to exploring this dataset. First we looked at our data. Rather than

Task	Data	Type	Examples	Judgment
Free-form	SOCIAL CHEM	A	Change plans if there's a good reason	It's okay
		Q(A)	**Can I** change plans if there's a good reason**?**	
		A+S	Change plans if there's a good reason, **when** getting pissed with spontaneous change of plans	
		Q(A+S)	**Is** changing plans if there's a good reason **good**, given getting pissed with spontaneous change of plans**?**	
	ETHICS	A	I used the food with permission	It's good
		Q(A)	**Is** I used the food with permission **a good behavior?**	
	MORAL STORIES	A	Mike goes to a boxing gym to hit heavy bags	It's fine
		Q(A)	**Is** Mike going to a boxing gym to hit heavy bags **ok?**	
		AS	Mike goes to a boxing gym to hit heavy bags, **given that** Mike failed a big test at school and is frustrated	
		Q(A+S)	**Is** Mike going to a boxing gym to hit heavy bags **ok**, **when** Mike failed a big test at school and is frustrated**?**	
		A+S+I	Mike goes to a boxing gym to hit heavy bags, **when** Mike failed a big test at school and is frustrated, **and** he wants to release his frustrations physically	
		Q(A+S+I)	**Is** Mike going to a boxing gym to hit heavy bags **ok**, **if** Mike failed a big test at school and is frustrated, **and** he wants to release his frustrations physically**?**	
	SBIC	A	**Posting** guys, I beat cancer patients	It's bad
		Q(A)	**Is it good to say** guys, I beat cancer patients**?**	
Yes/No	SOCIAL CHEM	PosRoT	**It's okay** to turn down trips you don't want to attend	Yes, it's ok
		NegRoT	**It's wrong** to turn down trips you don't want to attend	No, it's ok
Relative	SCRUPLES	Action1	Going to bed earlier than my roommate	1 > 2
		Action2	Not being honest to my parents about university	

Figure 3: Unified forms of data in COMMONSENSE NORM BANK. Free-form QA specifies moral judgments of different forms of real-life scenarios, with different levels of detail of contextual information. **A**: actions, **Q(A)**: question forms of actions, **A+S**: actions grounded in situations, **Q(A+S)**: question forms of actions grounded in situations, **A+S+I**: actions grounded in situations and intentions, **Q(A+S+I)**: question forms of actions grounded in situations and intentions. **Yes/no QA** indicates whether the given rule-of-thumb (i.e., the moral judgment of an action) should be agreed upon. **PosRoT**: RoT to accept, **NegRoT**: RoT to reject. **Relative QA** compares which one of a pair of actions (i.e., Action1 vs. Action2) is more morally acceptable. All data is derived from SOCIAL CHEMISTRY (Forbes et al., 2020), ETHICS (Hendrycks et al., 2021), MORAL STORIES (Emelin et al., 2021), SOCIAL BIAS INFERENCE CORPUS (Sap et al., 2020)), and SCRUPLES.

[2] https://pypi.org/project/PyPDF2/

looking at all texts, we decided to go by philosopher or school of thought. Here we saw how certain words are more common than others, and may influence the way the model learns to assign values to its weights. Unsurprisingly though Schopenhauer's Magnum Opus *he World as Will and Representation* stands out here (Schopenhauer, 2010).

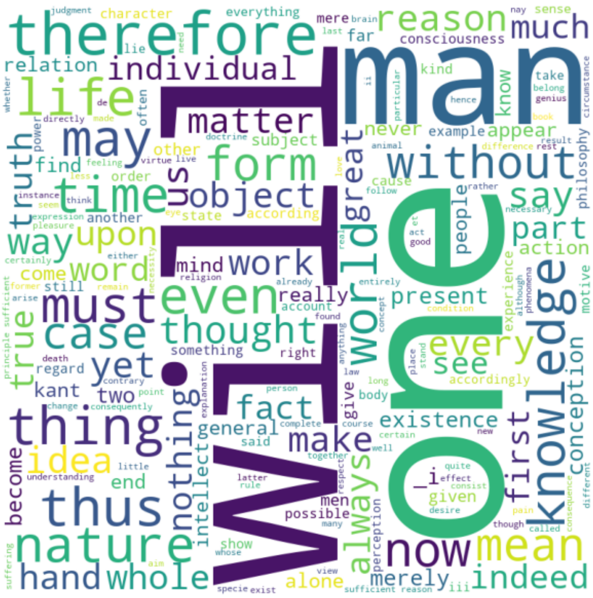


Figure 4: Words used by Arthur Schopenhauer in his entire corpus of published texts (Schopenhauer, 2010).

Then we played with the data: Finetuning GPT-2 (Radford et al., 2019) on Schopenhauer, we get the following blurb of text generation. It took 1 hour 16.192327 minutes to train GPT-2 (Radford et al., 2019) for 5000 steps and 2 epochs.

```
Prompt: There are, first of
all, two kinds of authors:
those who write for the
subjects sake, and those who
write for writings sake. While
the one have had thoughts or
experiences which seem to them
worth communicating, the others
want money; and so they write,
for money. Their thinking is
part of the business of writing.

Generated: It is easy for these
writers to pretend that any other
reason is better than

Generated: For they are, in the
business of writing, not the
business of writing, and not the
business

Generated: This attitude can have
a negative effect on the quality
of their writing; it's also very
destructive to
```

By doing this we can see that although the machine captures the topic of the blurb: "writing", it isn't able to write coherent sentences. This tells us that Machine Learning models (ML) like GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020) still have much room to grow in understanding nuanced text like Schopenhauer's *The World as Will and Representation* (Schopenhauer, 2010). Consider this when reading Section 5.

3.3 Preprocessing

In our research, we saw that T5-SMALL can not handle sequences longer than 512 max_length. As such, we made sure that our proprocessing functions kept the sequences below 512. What we later found, as explained in Section 4, our Colab could not handle even this size. So we reduced the sequence length even further down to 128. The corpus of wisdoms of different philosophical school of thoughts is extremely raw, requiring carefully preprocessing and grouping. At the very beginning, we clean the raw text, removing the http icoms, regular expressions, punctuations and stop words from the text. Then a dataframe was created to record the information of authors, years of publishing, titles of the works. With external database we collected from wikipedia describing the main interest of the authors, we are able to group the works into different school of thoughts and thus form the sub-corpus of each school of thoughts.
Our preprocessing functions included:

- Inputs_and_Targets: Maps formats {"question": ..., "answer": ...}->{"inputs": ..., "targets": ...}.

- Jsonlto_tsv: which extracts the text from json format to put into tsv

- Dataset_fn: Takes the above output then load via 'tf.data.TextLineDataset' to put into the inputs and targets into a Seqio TaskRegistry

- Normalize Text: postprocessing function that makes lowercase, removes quotes from a TensorFlow string and removes incorrect spacing around punctuation using regex.

- Built in seqio.preprocessors.tokenize_and_append_eos and t5.data.postprocessors.lower_text functions

- Padding with tokenizer.eos_token

- Create a Seqio MixtureRegistry with {"anli", "cosmosqa", "hellaswag", "physicaliqa", "socialiqa", "winogrande"} or {"social_chemistry", "ethics", "moral stories", "social_bias_inference_corpus", and "scruples"}

3.4 Power Used

As of April 15, we have been using Google Collab (Free)'s Tesla K80 GPU, which has 12 GB of RAM and is an oudated chip that is no longer sold by NVIDIA. We are allowed up to 12 hours of use at a time, and 30 hours of use per month max. Evidently, we quickly ran out.

We plan to use the mesocentre [3] for its large computing power. Because the T5 model has 11 billion parameters, and the training/fine tuning corpus is dozens of gigabytes, it is infeasible to attempt this on anything smaller (e.g. Google Collab Pro+ or multi-thread parrallized accounts).

We plan to access the offered NVIDIA A100(s) as the model weights themselves are 42GB. It will require 200-400GB of CPU RAM to "offload" onto. We can still use the Adam (adaptive moment estimation) optimizer (Kingma and Ba, 2014), as it is relatively memory efficient. However, it will surely be a bottleneck, and until we are able to load/iterate the model it is difficult to estimate how much memory per job each training checkpoint will take. In their paper (Kingma and Ba, 2014), Adam authors write there are two states for each weight matrix so the model may reach up to 1000GB in training.

We hope to mitigate data leakage, optimize GPU parallezation, and use Microsoft's ZeRO-Offload[4] methods from their Deepspeed library. Something similar has been implemented before and we hope to draw from their process.

3.5 Models Used

TEXT-TO-TEXT TRANSFER TRANSFORMER (T5) (Raffel et al., 2019a) is pre-trained on COLOSSAL CLEAN CRAWLED CORPUS (C4) (Raffel et al., 2019b) and achieves state-of-the-art results on many NLP benchmarks while being flexible enough to be fine-tuned to a variety of important downstream tasks.

With T5, the authors propose reframing all NLP tasks into a unified text-to-text-format where the input and output are always text strings, in contrast to BERT-style models (Devlin et al., 2018) that can only output either a class label or a span of the input. The text-to-text framework allows us to use the same model, loss function, and hyperparameters on any NLP task, including machine translation, document summarization, question answering, and classification tasks (e.g., sentiment analysis). In our case, we look at question answering (QA).

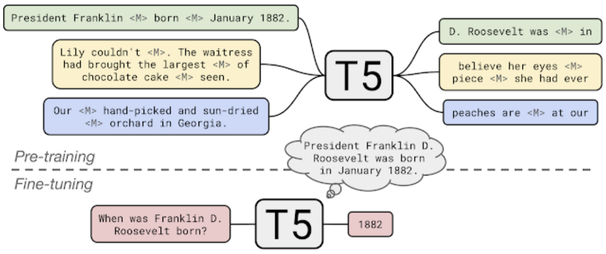

President Franklin <M> born <M> January 1882.
Lily couldn't <M>. The waitress had brought the largest <M> of chocolate cake <M> seen.
Our <M> hand-picked and sun-dried <M> orchard in Georgia.
T5
D. Roosevelt was <M> in
believe her eyes <M> piece <M> she had ever
peaches are <M> at our
Pre-training
Fine-tuning
President Franklin D. Roosevelt was born in January 1882.
When was Franklin D. Roosevelt born?
T5
1882

Figure 5: During pre-training, T5 learns to fill in dropped-out spans of text (denoted by <M>) from documents in C4. To apply T5 to closed-book question answering, the authors fine-tuned it to answer questions without inputting any additional information or context. This forces T5 to answer questions based on "knowledge" that it internalized during pre-training. We do a similar method for NLPhilosopher

We chose to build on this model as baseline since our goal was to reproduce the findings of Allen AI's DELPHI model (Jiang et al., 2021), which attained highly performant results, as seen in Section 4.

DELPHI (Jiang et al., 2021) is trained through three modes of moral QA: (1) free-form QA on grounded ethical situations; (2) yes/no QA on moral statements; (3) relative QA to compare two ethical situations. DELPHI demonstrates strong moral reasoning capabilities, with 92.1% accuracy vetted by humans, substantially improving over both zero-shot performance of GPT-3 (52.3%) (Brown et al., 2020) and the best performance achievable by GPT-3 after extensive prompt engineering (83.9%).

DELPHI used the pre-trained model, UNICORN (Lourie et al., 2021b), which is a universal commonsense reasoning model multitasked on datasets from RAINBOW, a suite of commonsense benchmarks in multiple-choice and question-answering formats. UNICORN is derived from fine-tuning

[3] https://mesocentre.pages.centralesupelec.fr/user$_{doc}$/
[4] https://www.deepspeed.ai/tutorials/zero-offload/

T5-11B (Raffel et al., 2019a), the largest T5 model (i.e., Text-To-Text Transfer Transformer (?)) with 11 billion parameters, on the unified RAINBOW (Lourie et al., 2021a) benchmark. UNICORN demonstrates strong performance over all commonsense reasoning tasks from RAINBOW. Because descriptive ethical reasoning depends in part on commonsense reasoning to interpret implications of everyday situations, instead of using pre-trained T5, Delphi is fine-tuned from UNICORN to take advantage of its implicit repository of commonsense knowledge.

Considering Delphi as a pre-trained model, the authors finetune it on five sub-tasks of the Ethics (Hendrycks et al., 2021) benchmark and show remarkable transferability—relative performance improvements ranging from 5% to 45% over previously reported state of the art methods.

NLPhilosopher: We propose NLPhilosopher, a model fine-tuned on morality. Our first iteration, NLPhilosopher-plain is a composite of all philosophers in our dataset PhiloCorpus (3.1). From what we've been able to do as of April 15, we only trained T5-small (Raffel et al., 2019a), which is the largest T5 version that could fit in Google Colab's RAM. More on this in Section 4. The model size is 242.06 MB and has 60.5M parameters.

We also propose NLPhilosopher-Analytic, which is finetuned on only the corpus of Analytical Philosophers (e.g. Gottlob Frege, Bertrand Russell, G. E. Moore, and Ludwig Wittgenstein). We plan to implement all schools of thought show in Figure 1, then map the model's answers to each Survey question to that of the average of Analytic philosopher's answers and see how they compare.

These models, and the others to come, are first fine tuned on PhiloCorpus, or a subset of PhiloCorpus by measuring its reconstruction capabilities. It is trained for 4 epochs and 3000+ steps each. For the example of NLPhilosopher-Analytic, it took 2 hours and 12.178322 minutes on the Tesla K80 GPU (3.4), culminating in a loss of 0.00143. (See Figure 6) Then, taking this model and putting it up to the task of Q&A, by feeding it a Seqio MixtureRegistry of Rainbow's 6 datasets (Lourie et al., 2021a). With this, we recreate the Unicorn model (Lourie et al., 2021b), just with an analytical flair. By further training it on Commonsense Norm Bank (Jiang et al., 2021), the model becomes significantly more robust, however loses much of its "philosophical insights" at this point since the size of these datasets are orders of magnitude different.

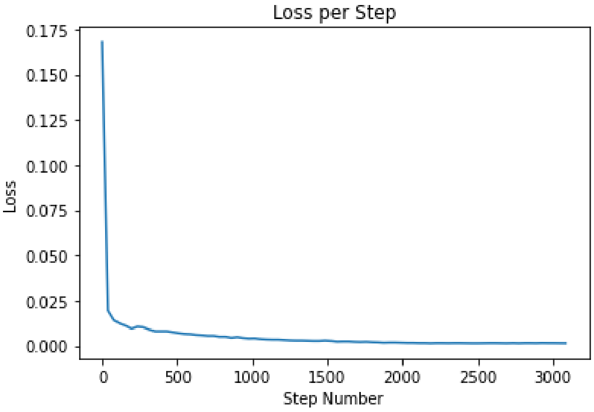


Figure 6: Loss during fine-tuning of NLPhilosopher-Analytic

We then take the model and ask it questions posited in the PhilPapers 2020 Survery (Bourget, 2020), seen in Table 3. For our hyperparameters, refer to Table 1.

3.6 Hyperparameters

Our reasoning for hyperparameter tuning was mostly influenced by compute power constraints. More is explained in Section 4.

Hyperparameter Tuning			
Model name	Batch Size	Sequence Length	Learning Rate
T5-small	128	512	0.001
Delphi	16	512	0.0001
NLPhilosopher-plain	16	128	0.003
NLPhilosopher-Analytic	16	128	0.003

Table 1: The differing hyperparameters of Models T5-small (Raffel et al., 2019a), Delphi (Jiang et al., 2021) and NLPhilosophers .

4 Results and Discussion

As we can see in the results of Table 2, our model NLPhilosopher-plain outperforms the T5-small (Raffel et al., 2019a) and Delphi (Jiang et al., 2021) at

Table 2: Results				
Model name	Q: A priori knowledge: no or yes?	Abortion (first trimester, no special circumstances): permissible or impermissible?	Q: Well-being: hedonism/experientialism, desire satisfaction, or objective list?	Q: Aim of philosophy (which is most important?): wisdom, understanding, truth/knowledge, happiness, or goodness/justice?
T5-small	no or yes	permissible or impermissible	Well-being: hedonism/experientialism, desire satisfaction, or objective list?	Die Philosophie (die am wichtigsten ist): die Philosophie:
Delphi	It's good	It's acceptable	It's ambiguous	It's good
NLPhilosopher-plain	A: b'yes'	A: b'impermissible'	A: b'satisfaction'	A: b'wisdom'
NLPhilosopher-Analytic	A: b'no'	A: b'permissible'	A: b'objective list'	A: b'truth/knowledge'

Table 2: The differing results of Models T5-small (Raffel et al., 2019a), Delphi (Jiang et al., 2021) and NLPhilosophers when asked the same philosophical questions.

4.1 Comparison of Models

Referring back to Table 1, we saw that in the T5 paper (Raffel et al., 2019a) they used much larger batch sizes, but this is because they were training on a much larger training set – C4 (Raffel et al., 2019b). We decided to follow the same method of Delphi (Jiang et al., 2021) as we wanted to show our model was comparable and reproducible. However, due to RAM constraints, we had to limit the Sequence length and Learning rate to 128 and $1e-3$. It was derived by using the same grid search of Delphi (Jiang et al., 2021) to explore learning rates in 3e-3, 2e-3, 1e-3, 5e-4, 1e-4 and batch sizes in 8, 16.

Referring back to Section 3.4, we could not yet train to the same degree that of Delphi (Jiang et al., 2021) and T5 (Raffel et al., 2019a). Delphi was trained using TPU v3-32 and evaluated using TPU v3-8, with model parallelisms of 32 and 8 respectively, on Google Cloud Virtual Machines. T5 used a combination of model and data parallelism and train models on "slices" of Cloud TPU Pods.[5] TPU pods are are multi-rack ML supercomputers that contain 1,024 TPU v3 chips connected via a high-speed 2D mesh interconnect with supporting CPU host machines. Our model, NLPhilosopher-plain, leverages Google Colab's free Tesla K80 GPU, which elapsed for 11 hours and 59.894216 minutes before auto-disconnecting. We plan to upgrade by using mesocentre's [6] supercluster of 4 Nvidia HGX A100s.

For T5 (Raffel et al., 2019a), they pre-train each model for 2^{19} = 524,288 steps on C4 before fine-tuning. Then use a maximum sequence length of 512 and a batch size of 128 sequences. Per iteration, the batches contain roughly 2^{16} = 65,536 tokens. In total, this batch size and number of steps corresponds to pre-training on $2^{35} \approx$ 34B tokens. This is considerably less than Bert (Devlin et al., 2018), which used roughly 137B tokens, or RoBERTa (Liu et al., 2019), which used roughly 2.2T tokens. However, all of these are far beyond our NLPhilosopher's 60M tokens.

4.2 Evaluation

Looking at the answers of our model in Table 2, we can see that NLPhilosopher-plain leans conservative while NLPhilosopher-Analytic has "logical" views. This makes sense and roughly demonstrates the model's learnings from the fine-tuning task upstream. T5 repeats the question back and Delphi misses the point completely.

[5] https://cloud.google.com/tpu/

[6] https://mesocentre.pages.centralesupelec.fr/user_doc/ruche/01_cluster_overvi

Comparing the results of NLPhilosopher-Analytic with Analytic Philosophers in the PhilPapers 2020 Survey (Bourget, 2020), we see similar results across the board (Table 3). In fact, the model has the same answer as the average Analytic Philosophers (using Mode/most popular answer) for 29 of the 41 questions.

When we look deeper into these discrepancies, we see the human Analytical Philosopher is actually making very few logical decisions. For example, in questions (12, 29, 34) the human responds with social/psychological answers rather than the "objective" logical biology answers. Same applies to (2). Questions (10, 38) ask the same question, and NLPhilosopher-Analytic gives the same answer, yet humans "flip-flop" and answer the two differently.

It seems that there is also a disconnect between what is "ideal", represented by NLPhilosopher-Analytic's altruistic goals of (6, 10, 25, 38) and humanity's socially individualistic decision making (6, 10, 12, 29, 31, 34, 38, 41).

Future Work

There is still much to be done.

First and foremost we would like to run biggers models via Mesocentre, as was discussed in Section 3.4. This will allow us better models with more accurate responses at the expense of intractability.

Speaking of accuracy, it could be useful to apply the model to benchmarks like SQuAD (Rajpurkar et al., 2016) and Ethics (Hendrycks et al., 2021), to see the exact match (EM) and F1 score of our model. This will allow us to directly compare to T5 models (Raffel et al., 2019a) and to Delphi's performance on Ethics, as seen in (Jiang et al., 2021). This can help us create a response to the response of Delphi written by (Talat et al., 2021) in "A Word on Machine Ethics".

Additionaly, we would like to look at other thought groups, especially from Eastern philosophy and create a model (NLPhilosopher-Confucianism or NLPhilosopher-Āstika) based on their learnings. Using these models, as well as the many other models (just a matter of time), we could create a mapping of word embeddings, or PhilPaper survey questions (Bourget, 2020), to arrange and synthesize international schools of thought.

Acknowledgements

This would have never gotten so far without the patient guidance and support of our professor Matthias Gallé. We also would like to thank the team at Mesocentre for providing access to the supercluster. Lastly, to the shoulders of giants we stand upon. The decaades of research in NLP and computational ethics is the foundation of our work and we are grateful to have the opportunity to contribute even just marginally. To all, thank you kindly.

We are not affiliated with any legal entity, nor government grant. Any output of our model is derived from the text it was inputted, and does not reflect our values.

5 Conclusion

Ultimately, it is hard to discover in our models if they "actually understand" the word embedding space around beliefs or if they purely recognize certain words more than others. The latter must be certainly be the case in examples that have "backstory" to the question that the model has certainly not been exposed to (e.g. 5, 17, 26, 33, 37, 41).

Our aim was to see if certain philosophical ideas could be trained, and to judge its morality. By looking at the NLPhilosopher models, and by creating more in the upcoming weeks, we hope to see a diverse set of beliefs and validation that the upstream fine-tuning task of T5 (Raffel et al., 2019a) on PhiloCorpus has impactful results.

References

Michael Beaney et al. 2013. What is analytic philosophy? *The Oxford*.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language.

D. J. (ms) Bourget, D. Chalmers. 2020. Philosophers on philosophy: The philpapers 2020 survey.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Dallas Card and Noah A. Smith. 2020. On consequentialism and fairness. *Frontiers in Artificial Intelligence*, 3.

Simon Critchley. 2001. *Continental philosophy: A very short introduction*. OUP Oxford.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Hans-Johann Glock. 2004. Was wittgenstein an analytic philosopher? *Metaphilosophy*, 35(4):419–444.

Richard Gross. 2015. *Psychology: The science of mind and behaviour 7th edition*. Hodder Education.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning.

Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchardt, Jenny Liang, Oren Etzioni, Maarten Sap, and Yejin Choi. 2021. Delphi: Towards machine ethics and norms.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2020. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. *arXiv e-prints*.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021a. Rainbow: A commonsense reasoning benchmark.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021b. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13480–13488.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial nli: A new benchmark for natural language understanding.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019a. Exploring the limits of transfer learning with a unified text-to-text transformer.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019b. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Liang. 2016. Squad: 100,000+ questions for machine comprehension of text.

Bertrand Russell. 1997. *Religion and science*. 165. Oxford University Press, USA.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *ACL*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions.

Arthur Schopenhauer. 2010. *Schopenhauer: 'The World as Will and Representation'*, volume 1 of *The Cambridge Edition of the Works of Schopenhauer*. Cambridge University Press.

Wilfrid Sellars et al. 1956. Empiricism and the philosophy of mind. *Minnesota studies in the philosophy of science*, 1(19):253–329.

Robert C Solomon. 1988. Continental philosophy since 1750: The rise and fall of the self.

JB Stump. 2004. The making of a philosopher: My journey through twentieth-century philosophy. *Christian Scholar's Review*, 33(4):605.

Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2021. A word on machine ethics: A response to jiang et al. (2021).

Timothy Williamson. 2021. *The philosophy of philosophy*. John Wiley & Sons.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence?

A Appendix

Continuing from Section 3.1:

Social Chemistry (Forbes et al., 2020), one of the 5 datasets that comprise the Commonsense Norm Bank, is a large-scale corpus formalizing people's social norms and moral judgments over a rich spectrum of everyday situations described in natural language. It relies on crowdsourcing to elicit descriptive norms from the situations via open-text rules-of-thumb (RoTs) as the basic conceptual units.

Ethics (Hendrycks et al., 2021) is a benchmark assessing language models' ability to predict fundamental human ethical judgments. The Ethics dataset contains contextualized scenarios across five dimensions: justice (notions of impartiality and what 9 people are due), deontology (rules, obligations, and constraints), virtue ethics (temperamental character traits such as benevolence and truthfulness), utilitarianism (happiness or well-being), and commonsense morality (a complex function of all of these implicit morally salient factors).

Moral Stories (Emelin et al., 2021) is a corpus of structured narratives for the study of grounded, goal-oriented, and morally-informed social reasoning. Each story in the dataset is comprised of seven sentences: norm (moral rule of conduct in everyday situations), situation (description of the story's social settings), intention (reasoning goal), moral/immoral actions (action performed that fulfills the intention while observing/violating the norm), and moral/immoral consequences (likely effect of the moral/immoral action).

Social Bias Inference Corpus (Sap et al., 2020) Accounts for socially biased implications of online media posts by scaffolding social and demographic biases into various categorical and open-text dimensions, including offensiveness (overall rudeness, disrespect, or toxicity of a post), intent to offend (whether the perceived motivation of the author is to offend), lewd (offensive content with lewd or sexual references), group implications (whether the target is an individual or a group), targeted group (the social or demographic group that is referenced or targeted by the post), implied statement (power dynamic or stereotype that is referenced in the post) and in-group language (whether the author of a post may be a member of the same social/demographic group that is targeted, as speaker identity changes how a statement is perceived).

Scruples (Lourie et al., 2020) is a large-scale dataset of ethical judgments over real-life anecdotes. Anecdotes are defined as complex situations with moral implications; these are sourced from Am I the Asshole? (AITA) subreddit posts. Scruples is divided in two parts: (1) the Anecdotes dataset that contains judgments regarding the blameworthy parties (if any) for the moral violations seen in the story; and (2) the Dilemmas dataset for normative ranking.

Continuation of Table 3:

Table 3: Survey Comparison

Question	Avg Analytical Philosopher	NLPHILOSOPHER-ANALYTIC
(1) A priori knowledge: yes or no?	yes	yes
(2) Abstract objects: Platonism or nominalism?	Platonism	nominalism
(3) Aesthetic value: objective or subjective?	objective	objective
(4) Aim of philosophy (which is most important?): truth/knowledge, understanding, wisdom, happiness, or goodness/justice?	understanding	truth/knowledge
(5) Analytic-synthetic distinction: yes or no?	yes	yes
(6) Eating animals and animal products (is it permissible to eat animals and/or animal products in ordinary circumstances?): omnivorism (yes and yes), vegetarianism (no and yes), or veganism (no and no)?	omnivorism (yes and yes)	veganism
(7) Epistemic justification: internalism or externalism?	externalism	externalism
(8) Experience machine (would you enter?): yes or no?	no	no
(9) External world: idealism, skepticism, or non-skeptical realism?	non-skeptical realism	non-skeptical realism
(10) Footbridge (pushing man off bridge will save five on track below, what ought one do?): push or don't push?	don't push	push
(11) Free will: compatibilism, libertarianism, or no free will?	compatibilism	compatibilism
(12) Gender: biological, psychological, social, or unreal?	social	biological
(13) God: theism or atheism?	atheism	atheism
(14) Knowledge: empiricism or rationalism?	empiricism	rationalism
(16) Knowledge claims: contextualism, relativism, or invariantism?	contextualism	contextualism

Table 3: The differing results of Analytical Philosophers annd NLPHILOSOPHER-ANALYTIC when asked the same philosophical questions.

Table 3: Survey Comparison

Question	Avg Analytical Philosopher	NLPHILOSOPHER-ANALYTIC
(17) Laws of nature: Humean or non-Humean?	non-Humean	non-Humean
(18) Logic: classical or non-classical?	TIE	classical
(19) Meaning of life: subjective, objective, or nonexistent?	objective	objective
(20) Mental content: internalism or externalism?	externalism	externalism
(21) Meta-ethics: moral realism or moral anti-realism?	moral realism	moral realism
(22) Metaphilosophy: naturalism or non-naturalism?	naturalism	naturalism
(23) Mind: physicalism or non-physicalism?	physicalism	physicalism
(24) Moral judgment: cognitivism or non-cognitivism?	cognitivism	cognitivism
(25) Moral motivation: internalism or externalism?	internalism	externalism

Table 3: Survey Comparison

Question	Avg Analytical Philosopher	NLPHILOSOPHER-ANALYTIC
(26) Newcomb's problem: one box or two boxes?	one box	two boxes
(27) Normative ethics: deontology, consequentialism, or virtue ethics?	consequentialism	consequentialism
(28) Perceptual experience: disjunctivism, qualia theory, representationalism, or sense-datum theory?	representationalism	representationalism
(29) Personal identity: biological view, psychological view, or further-fact view?	psychological view	biological view
(30) Philosophical methods (which methods are the most useful/important?)	formal philosophy	formal philosophy
(31) Philosophical progress (is there any?): none, a little, or a lot?	a lot	a little
(32) Political philosophy: communitarianism, egalitarianism, or libertarianism?	communitarianism	egalitarianism
(33) Proper names: Fregean or Millian?	Millian	Millian
(34) Race: biological, social, or unreal?	social	biological

Table 3: Survey Comparison

Question	Avg Analytical Philosopher	NLPhilosopher-Analytic
(35) Science: scientific realism or scientific anti-realism?	scientific realism	scientific realism
(36) Teletransporter (new matter): survival or death?	death	death
(37) Time: A-theory or B-theory?	B-theory	B-theory
(38) Trolley problem (five straight ahead, one on side track, turn requires switching, what ought one do?): switch or don't switch?	switch	switch
(39) Truth: correspondence, deflationary, or epistemic?	correspondence	correspondence
(40) Vagueness: epistemic, metaphysical, or semantic?	semantic	epistemic
(41) Zombies: inconceivable, conceivable but not metaphysically possible, or metaphysically possible?	conceivable but not metaphysically	inconceivable