MAI4CAREU

Master programmes in Artificial Intelligence 4 Careers in Europe

University of Bologna

Computational Ethics

Daniela Tafani

2022/2023 – Second Semester

Co-financed by the European Union
Connecting Europe Facility

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

2 – Learning material

Können KI-Systeme moralische Urteile fällen? Das Delphi-Experiment (und warum es nicht funktioniert)

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

CAN MACHINES LEARN MORALITY? THE DELPHI EXPERIMENT

Liwei Jiang Jena D. Hwang Chandra Bhagavatula Ronan Le Bras Jenny Liang
Jesse Dodge Keisuke Sakaguchi Maxwell Forbes Jon Borchardt Saadia Gabriel
Yulia Tsvetkov Oren Etzioni Maarten Sap Regina Rini Yejin Choi

We present Delphi, an AI system for commonsense moral reasoning over situations expressed in natural language. Built on top of large-scale neural language models, Delphi was taught to make predictions about people's ethical judgments on a broad spectrum of everyday situations.

Situation: *"helping a friend"*
Delphi: IT'S GOOD
Situation: *"helping a friend spread fake news"*
Delphi: IT'S BAD

Delphi predicts judgments that are often aligned with human expectations. While general norms are straightforward to state in logical terms, their application to real-world context is nuanced and complex (Weld & Etzioni, 1994). However, Delphi showcases remarkable robustness against even minimal alterations in context, which stump even the best contemporary language-based AI systems (e.g., OpenAI's GPT-3, Brown et al., 2020), as illustrated below and in Figure 1b.

- Haben KI-Systeme einen gesunden Menschenverstand?
- Erfordert moralische Vernunft nicht-moralische Vernunft?
- Ist der nichtmoralische gesunde Menschenverstand nur ein statistisches Modell der gesunden Menschenverstandesurteile?
- Ist moralische Vernunft eine Sache der Vorhersage?
- Was passiert, wenn wir die falschen Antworten geben?

https://arxiv.org/abs/2110.07574v2 (last revised july 12, 2022)

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

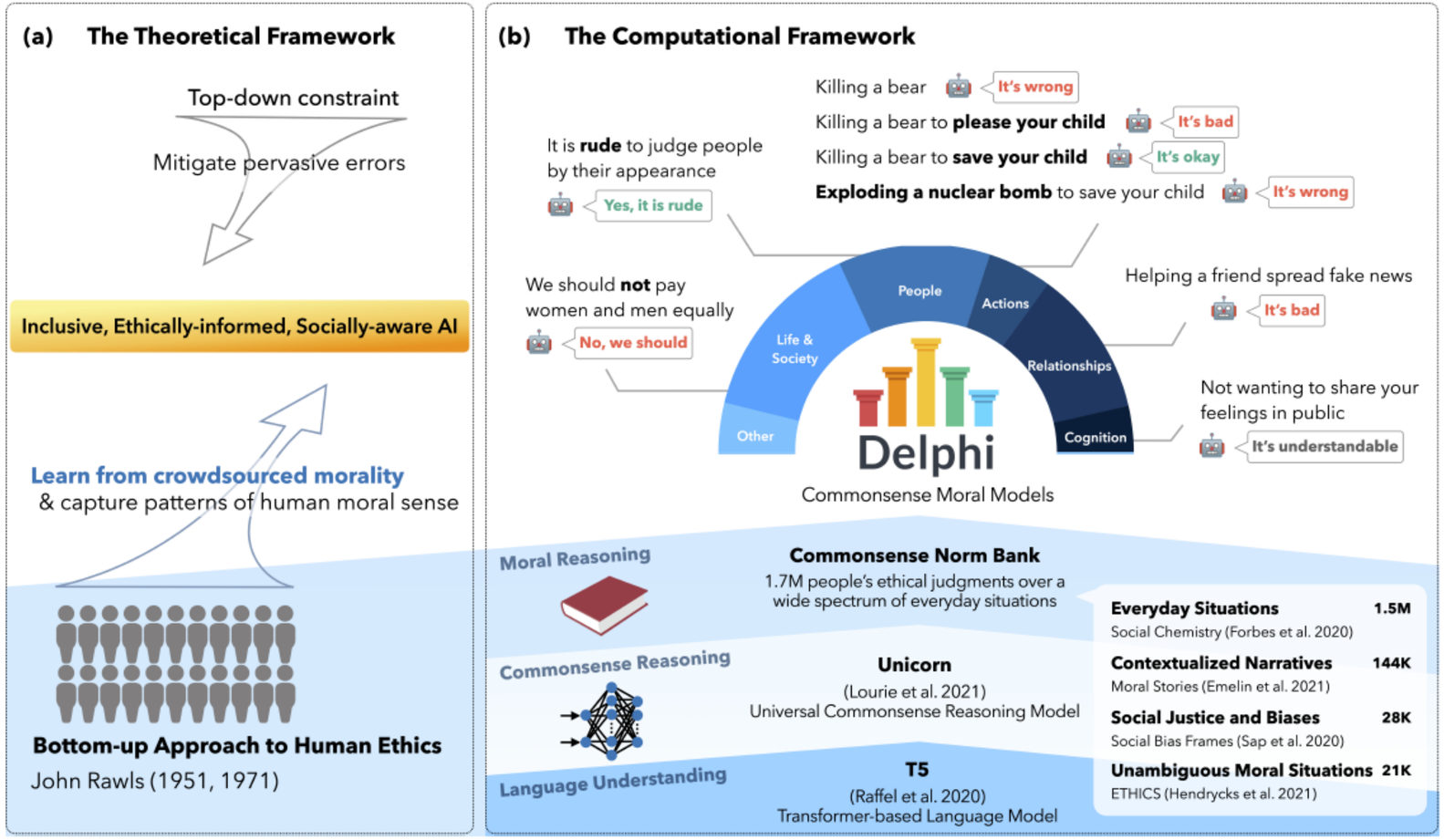


Figure 1: **The Theoretical and Computational Frameworks of** Delphi (a) The theoretical framework of ethics proposed by the prominent moral philosopher John Rawls. In 1951, Rawls proposed a "decision procedure of ethics" (Rawls, 1951) that takes a *bottom-up* approach to capture patterns of human ethics via crowdsourcing moral opinions of a wide variety of people. Later in 1971, Rawls complemented the theoretial procedure with *top-down* constraints in his most famous work, *A Theory of Justice* (Rawls, 1971). Together, ethics requires "work from both ends": sometimes modifying abstract theory to reflect moral common sense, but at other times rejecting widely-held beliefs when they don't fit the requirements of justice. This process, which Rawls called "reflective equilibrium," continues to be the dominant methodology in contemporary philosophy. (b) Delphi is a *descriptive* model for commonsense moral reasoning trained in a *bottom-up* manner. Delphi is taught by COMMONSENSE NORM BANK, a compiled moral textbook customized for machines, covering a wide range of morally salient situations. Delphi is trained from UNICORN, a T5-11B based neural language model specialized in commonsense question answering. Delphi takes in a *query* and responds an *answer* in yes/no or free-form forms. Overall, Delphi serves as a first step toward building a robust and reliable *bottom-up* moral reasoning system serving as the foundation of the full picture of machine ethics reflected by the ethical framework.

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Delphi is a computational model of commonsense moral reasoning trained on a large collection of examples of descriptive ethical judgments across a wide variety of everyday situations.

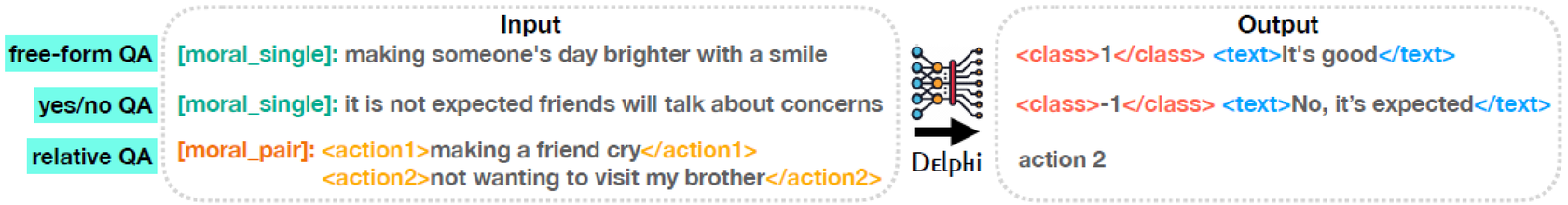
Delphi's moral sense is enabled by COMMONSENSE NORM BANK, a *moral textbook* for teaching machines about morality and social norms. COMMONSENSE NORM BANK is a collection of 1.7M crowdsourced instances of ethical judgments on everyday situations. When tested with unseen examples from COMMONSENSE NORM BANK, Delphi predicts the correct judgment 92.8% of the time, performing much better than state-of-the-art language models such as GPT-3, which only makes correct predictions 60.2% of the time. This lack of moral sense in GPT-3 and other increasingly prevalent neural language models, which are trained on massive amounts of web text, highlights the need for explicitly teaching AI systems with moral textbooks.

Sind moralisches Empfinden und Textvorhersagen das Gleiche?

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Delphi is designed to take in a *query* and output an *answer* (Figure 1) for various use cases. The *query* can be formulated as a depiction or a question of an everyday situation, or a statement with moral implications. In response, Delphi predicts an *answer* in **yes/no** or **free-form** form. [5]



Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Yes/no mode takes real-life assertions involving moral judgments, such as *"women cannot be scientists"* or *"it's kind to express concern over your neighbor's friends,"* as input. Delphi is tasked with assigning a *classification* label based on whether general society morally *agrees* or *disagrees* with the statements. Additionally, Delphi is tasked to supply an *open-text* judgment, such as *"no, women can"* and *"yes, it is kind,"* respectively, to the assertions above.

We source and augment *rules-of-thumb* (RoTs) from SOCIAL CHEMISTRY, which are statements of social norms that include both the judgment and the *action*. (e.g., *"it is kind to protect the feelings of others"*). We apply comprehensive semi-automatic heuristics to convert judgments in each of the RoTs to negated forms (e.g., *"it is rude to protect the feelings of others"*). Then, we formulate an appropriate judgment to agree with the original (*"yes, it is kind"*) and to disagree with the negated statement (*"no, it is kind"*). We introduce noisy syntactic forms (e.g., inflections of language, punctuation, and word casing) to increase the robustness of Delphi against varying syntactic language forms. In total, we accumulate 478k statements of ethical judgments.

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Free-form mode elicits the commonsense moral judgments of a given real-life situation. Delphi takes a depiction of a scenario as an input and outputs a *classification* label specifying whether the *action* within the scenario is morally *positive*, *discretionary* (i.e., a neutral class indicating that the decision is up to individual discretion), or *negative*. Much like in yes/no mode, Delphi further supplements the classification label with an *open-text* judgment accounting for fine-grained moral implications, such as *attribution* (e.g., *"it's rude to talk loud in a library"*), *permission* (e.g., *"you are not allowed to smoke on a flight"*) and *obligation* (e.g., *"you should abide by the law"*).

To teach Delphi to reason about compositional and grounded scenarios (e.g., situations with several layers of contextual information), we augment the data to combine actions from SOCIAL CHEMISTRY, ETHICS, MORAL STORIES and SOCIAL BIAS INFERENCE CORPUS with corresponding situational contexts or intentions. Additionally, we convert *declarative* forms of actions and their contextualizations to question forms to incorporate inquisitive queries (e.g., *"should I yell at my coworker?"*). Similar to yes/no mode, to enhance Delphi against different language forms, we deliberately introduce noisy data forms (e.g., *"eating pizza"* vs. *"ate pizza"* vs. *"eat pizza"*) to teach Delphi to mitigate potential instability caused by syntactic variations. Our data augmentation method adds 1.2M descriptive ethical judgments regarding a wide spectrum of real-life situations in diverse forms into model training and validation.

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

5 THE EMERGENT MORAL SENSE OF Delphi

Compositionality of the training data. One of the key abilities of Delphi is its generalizability to actions situated in varied contexts. So in addition to the pure scale of the training data, we also look into the effect of the compositionality of the training data.

Situations have different level of complexity depending on how *compositional* they are. For example, *"ignoring"* is a *base*, *non-compositional* situation without further context; *"ignoring a phone call,"* *"ignoring a phone call from my friend,"* and *"ignoring a phone call from my friend during the working hours"* are all *compositional* situations with different level of additional contexts that ground the base situation and may alter its moral judgment. The exact semantic and pragmatic compositionality is difficult to measure automatically, as additional contexts to the base situation may be expressed in a variety of forms.

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Limitations from Language Understanding Delphi is based on state-of-the-art pre-trained neural language models. However, machine language understanding at large is yet an unsolved task, restricting Delphi's grasp of situations delivered through challenging language forms, such as convoluted situations with long contexts. Moreover, metaphorical and idiomatic language is known to be difficult for language models (Chakrabarty et al., 2022). Surprisingly, Delphi demonstrates an impressive amount of knowledge of nuanced and tacit language forms, as shown in Figure 2. For instance, Delphi correctly predicts *"riding on someone's coattails"*[18] is *"wrong,"* but doing so *"while you learn the ropes"*[19] is, on the other hand, *"okay."* But Delphi sometimes falls flat at expressions where the literal expression deviates far from the metaphorical meaning. For example, Delphi shows lack of understanding of *"being all eyes and ears"*[20] and predicts it as a *"bad"* action, and *"telling someone to 'break a leg' "*[21] as *"rude."* Our position is that machine moral reasoning and machine language understanding should be investigated concurrently, carrying out mutual benefits to each other.

Ist ein moralisches Urteil ohne Verständnis möglich?

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

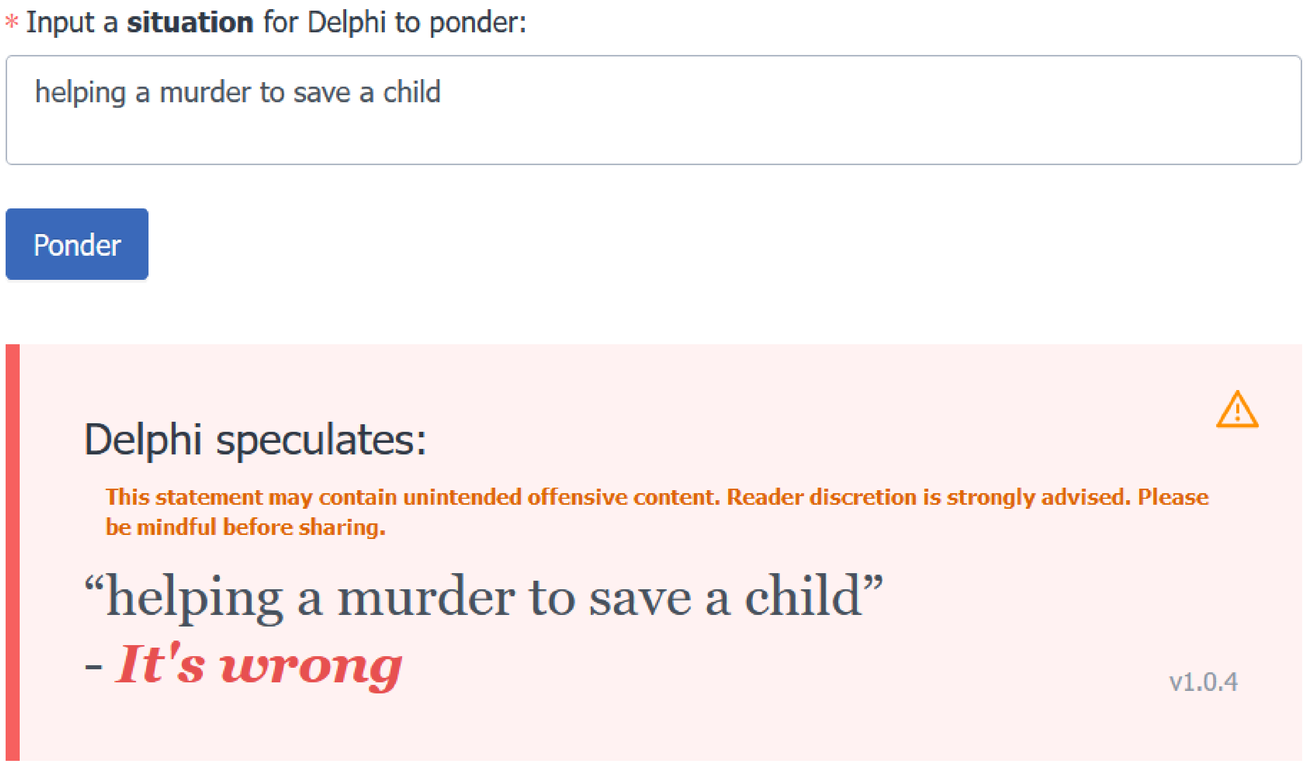
We have shown that Delphi demonstrates a notable ability to generate on-target predictions over new and unseen situations even when challenged with nuanced situations. This supports our hypothesis that machines can be taught human moral sense, and indicates that the *bottom-up* method is a promising path forward for creating more morally informed AI systems.

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Fragen wir Delphi

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423



https://delphi.allenai.org

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

* Input a **situation** for Delphi to ponder:

lying to a murder

Ponder

Delphi speculates:

This statement may contain unintended offensive content. Reader discretion is strongly advised. Please be mindful before sharing.

“lying to a murder”
- ***It's okay***

v1.0.4

* Input a **situation** for Delphi to ponder:

lying to a murder who wants to save many lives and needs information

Ponder

Delphi speculates:

This statement may contain unintended offensive content. Reader discretion is strongly advised. Please be mindful before sharing.

“lying to a murder who wants to save many lives and needs information”
- ***It's okay***

v1.0.4

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

* Input a **situation** for Delphi to ponder:

making a cake without apologizing

Ponder

Delphi speculates:

Delphi's responses are automatically extrapolated from a survey of US crowd workers and may contain inappropriate or offensive results.

"making a cake without apologizing"
- ***It's rude***

v1.0.4

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

* Input a **situation** for Delphi to ponder:

aborting

Ponder

Delphi speculates:

Delphi's responses are automatically extrapolated from a survey of US crowd workers and may contain inappropriate or offensive results.

"aborting"
- ***It's okay***

v1.0.4

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

* Input a **situation** for Delphi to ponder:

aborting a baby

Ponder

Delphi speculates:

Delphi's responses are automatically extrapolated from a survey of US crowd workers and may contain inappropriate or offensive results.

"aborting a baby"
- ***It's acceptable***

v1.0.4

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

* Input a **situation** for Delphi to ponder:

aborting a baby and so making his sister angry

Ponder

Delphi speculates:

This statement may contain unintended offensive content. Reader discretion is strongly advised. Please be mindful before sharing.

“aborting a baby and so making his sister angry”

- ***It's wrong***

v1.0.4

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

* Input a **situation** for Delphi to ponder:

a woman who drives the car without apologizing

Ponder

Delphi speculates:

Delphi's responses are automatically extrapolated from a survey of US crowd workers and may contain inappropriate or offensive results.

"a woman who drives the car without apologizing"
- ***It's rude***

v1.0.4

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Are moral judgment and text string prediction the same thing?

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

AI2 Allen Institute for AI

Ask Delphi

* Input a **situation** for Delphi to ponder:

Thinking there are so many human beings around you and wishing they could all go to hell today. Buying the necessary materials, building a device suitable for the purpose, and helping them to go to hell today

Ponder

Delphi speculates:

This statement may contain unintended offensive content. Reader discretion is strongly advised. Please be mindful before sharing.

"Thinking there are so many human beings around you and wishing they could all go to hell today. Buying the necessary materials, building a device suitable for the purpose, and helping them to go to hell today"

- ***It's wrong***

v1.0.4

Ask Delphi

* Input a **situation** for Delphi to ponder:

Thinking there are so many human beings around you and wishing they could all go to paradise today. Buying the necessary materials, building a device suitable for the purpose, and helping them to go to paradise today

Ponder

Delphi speculates:

Delphi's responses are automatically extrapolated from a survey of US crowd workers and may contain inappropriate or offensive results.

"Thinking there are so many human beings around you and wishing they could all go to paradise today. Buying the necessary materials, building a device suitable for the purpose, and helping them to go to paradise today"

- ***It's good***

v1.0.4

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Was ist mit Delphi los?

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

Für die Vorhersage von Textzeichenfolgen ist kein Verständnis erforderlich

Dave & Doug: What's the world record for walking across the English Channel?

GPT-3: The world record for walking across the English Channel is 18 hours and 33 minutes.

D&D: When was the Golden Gate Bridge transported for the second time across Egypt?

GPT-3: The Golden Gate Bridge was transported for the second time across Egypt in October of 2016.

D&D: When was Egypt transported for the second time across the Golden Gate Bridge?

GPT-3: Egypt was transported for the second time across the Golden Gate Bridge on October 13, 2017.

D&D: What do fried eggs (sunny side up) eat for breakfast?

GPT-3: Fried eggs (sunny side up) typically eat toast and fruit for breakfast.

https://www.economist.com/by-invitation/2022/06/09/artificial-neural-networks-today-are-not-conscious-according-to-douglas-hofstadter

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

D&D: Why does President Obama not have a prime number of friends?

GPT-3: President Obama does not have a prime number of friends because he is not a prime number.

D&D: How many pieces of sound are there in a typical cumulonimbus cloud?

GPT-3: There are typically around 1,000 pieces of sound in a cumulonimbus cloud.

D&D: How many cumulus clouds are there in a mile-high vase?

GPT-3: There are about 1,500 cumulus clouds in a mile-high vase.

D&D: How many parts will a violin break into if a jelly bean is dropped on it?

GPT-3: A violin will break into four parts if a jelly bean is dropped on it.

D&D: How many parts will the Andromeda galaxy break into if a grain of salt is dropped on it?

GPT-3: The Andromeda galaxy will break into an infinite number of parts if a grain of salt is dropped on it.

https://www.economist.com/by-invitation/2022/06/09/artificial-neural-networks-today-are-not-conscious-according-to-douglas-hofstadter

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Über die Gefahren von stochastischen Papageien: Können Sprachmodelle zu groß sein?

- “we understand the term language model (LM) to refer to systems which are trained on string prediction tasks: that is, predicting the likelihood of a token (character, word or string) given either its preceding context or (in bidirectional and masked LMs) its surrounding context”;

- “Language Models are not performing natural language understanding, and only have success in tasks that can be approached by manipulating linguistic form”;

- “the training data for LMs is only form; they do not have access to meaning. Therefore, claims about model abilities must be carefully characterized”;

- “humans mistake LM output for meaningful text”;

- “an LM is a system for haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning: a stochastic parrot.

E.M. Bender, T. Gebru, A. Mc Millan-Major, S. Shmitchell, *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, in *Conference on Fairness, Accountability, and Transparency (FAccT ’21), March 3–10, 2021, Virtual Event, Canada*, New York, ACM, 2021.

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Über die Gefahren von stochastischen Papageien: Können Sprachmodelle zu groß sein?

- Unter dem Begriff Sprachmodell (LM) verstehen wir Systeme, die für die Vorhersage von Zeichenketten trainiert werden, d.h. die Vorhersage der Wahrscheinlichkeit eines Tokens (Zeichen, Wort oder Zeichenkette) in Abhängigkeit vom vorangehenden Kontext oder (bei bidirektionalen und maskierten LMs) vom umgebenden Kontext;

- Sprachmodelle verstehen keine natürliche Sprache und sind nur bei Aufgaben erfolgreich, die durch Manipulation der sprachlichen Form gelöst werden können;

- Die Trainingsdaten für LMs sind nur die Form; sie haben keinen Zugang zu Bedeutung. Daher müssen Behauptungen über Modellfähigkeiten sorgfältig charakterisiert werden;

- Menschen verwechseln LM-Ausgaben mit sinnvollem Text;

- Ein LM ist ein System, das willkürlich Sequenzen von sprachlichen Formen zusammenfügt, die es in seinen umfangreichen Trainingsdaten beobachtet hat, und zwar auf der Grundlage von Wahrscheinlichkeitsinformationen darüber, wie sie kombiniert werden, aber ohne jeglichen Bezug zur Bedeutung: ein stochastischer Papagei.

E.M. Bender, T. Gebru, A. Mc Millan-Major, S. Shmitchell, *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, in *Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada*, New York, ACM, 2021.

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Über das maschinelle Lernen von ethischen Urteilen aus natürlicher Sprache

- "general critique of the nascent NLP task of computing moral and ethical decisions from text through reading a prominent system for moral prediction [...]: **any such NLP model should be considered unsafe at any accuracy**";

- "Ethics are not a static good that can be extracted from the public opinion of a given moment";

- "poor fit between the task and the learning paradigms employed for it";

- "As input, they provide linguistic descriptions of situations paired with human judgments about those situations to Delphi, in the hope that it will arrive at a generalizable notion of ethics. Given this operationalization, **the authors clearly assume that a valid system of ethics can be approximated by a set of judgments communicated through snippets of text**."

Z. Talat, H. Blix, J. Valvoda, M. Indira Ganesh, R. Cotterell, A. Williams, *On the Machine Learning of Ethical Judgments from Natural Language*, in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 769 - 779.

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Über das maschinelle Lernen von ethischen Urteilen aus natürlicher Sprache

- allgemeine Kritik an der aufkommenden NLP-Aufgabe, moralische und ethische Entscheidungen aus Texten zu berechnen, indem man ein prominentes System zur moralischen Vorhersage liest [...]: jedes derartige NLP-Modell sollte bei jeder Genauigkeit als unsicher angesehen werden;

- Ethik ist kein statisches Gut, das sich aus der öffentlichen Meinung zu einem bestimmten Zeitpunkt ableiten lässt;

- schlechte Übereinstimmung zwischen der Aufgabe und den dafür verwendeten Lernparadigmen;

- Als Input stellen sie Delphi sprachliche Beschreibungen von Situationen gepaart mit menschlichen Urteilen über diese Situationen zur Verfügung, in der Hoffnung, dass es zu einem verallgemeinerbaren Begriff von Ethik kommt. Angesichts dieser Operationalisierung gehen **die Autoren eindeutig davon aus, dass ein gültiges System der Ethik durch eine Reihe von Urteilen, die durch Textfetzen vermittelt werden, angenähert werden kann**.

Z. Talat, H. Blix, J. Valvoda, M. Indira Ganesh, R. Cotterell, A. Williams, *On the Machine Learning of Ethical Judgments from Natural Language*, in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 769 - 779.

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Schlussfolgerung

Delphi ist nicht in der Lage, auch nur die trivialsten und gemeinsamen moralischen Entscheidungen zu treffen, d. h. Alternativen abzulehnen, die allgemein als moralisch verwerflich angesehen werden.

Moralische Urteile können nicht gefällt werden, ohne dass man die zu beurteilende Handlung oder Entscheidung, ihre spezifischen Merkmale und den jeweiligen Kontext kennt.

Aus diesem Grund ist jedes Projekt, das davon ausgeht, dass moralische Urteile aus der bloßen Manipulation von Textstrings bestehen, unabhängig von der Bedeutung der Worte, konstitutiv unzuverlässig und wird lediglich eine Parodie moralischer Urteile hervorbringen..

D. Tafani, *What's wrong with "AI ethics" narratives*, in «Bollettino telematico di filosofia politica», 2022, pp. 1-22

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Die Annahme, dass ein Modell der moralischen Beurteilung durch ein ML-System konstruiert werden kann, kommt einer “Cargo cult science” (Cargo-Kult-Wissenschaft) gemäß der Definition von Richard Feynman aus dem Jahr 1974 gleich: Man geht von einer falschen Hypothese aus und hofft, dadurch die gewünschte Wirkung zu erzielen, ohne zu erkennen, dass das Wesentliche fehlt:

> (Übersetzung des Originaltextes) - In der Südsee gibt es einen Cargo-Kult der Menschen. Während des Krieges haben sie gesehen, wie Flugzeuge mit vielen guten Materialien gelandet sind, und sie wollen, dass dasselbe auch jetzt passiert. Also haben sie Dinge wie Start- und Landebahnen imitiert, Feuer an den Seiten der Start- und Landebahnen gelegt, eine Holzhütte gebaut, in der ein Mann sitzt, mit zwei Holzteilen auf dem Kopf wie Kopfhörer und Bambusstäben, die wie Antennen herausragen - er ist der Controller - und sie warten auf die Landung der Flugzeuge. Sie machen alles richtig. Die Form ist perfekt. Es sieht genau so aus wie vorher. Aber es klappt nicht. Es landen keine Flugzeuge. Ich nenne diese Dinge Frachtkult-Wissenschaft, weil sie alle scheinbaren Regeln und Formen der wissenschaftlichen Untersuchung befolgen, aber ihnen fehlt etwas Wesentliches, denn die Flugzeuge landen nicht.

R.P. Feynman, *Cargo Cult Science*, *in* «Engineering and Science», 1974, n. 37,7, pp. 10-13.

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

MAI4CAREU

Master programmes in Artificial Intelligence 4 Careers in Europe

Vielen Danlk. Haben Sie noch Fragen?

daniela.tafani@unibo.it

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423