AI & Robotics

Manuel Eiter

January 7, 2024

Abstract

Recent advancements in natural language processing and AI as a whole can be interpreted as the beginning of an AI revolution. With AI getting ground in more areas, guaranteeing ethical standards is essential. This paper gives a brief introduction into natural language processing and robots enhanced with it. It includes ethical considerations regarding both. And states why current ethical guidelines fall short of its object to regulate the development of AI and why regulations introduced via law might be a better approach (AI Act).

Introduction

With the latest breakthroughs in AI and specifically natural language processing (NLP) it is only a question of time until powerful models will be integrated into robots. Not only giving them language capabilities, but also the ability for automated decision-making.

Like every new invention, this will not come without any risk. In fact, AI and NLP on its own have already shown some problems in relation to bias and fairness. Making them, when not addressed correctly, dangerous for use in applications where unfair results could lead to harm for humans.

The first section covers NLP, it is the branch of AI that has seen tremendous advancements in recent years. It is also responsible for breakthroughs in other branches of AI, as its large language models have proven effective as foundation models or general purpose models. Thereafter, comes a short history of language models, leading us to the risks and challenges we face when training this models.

In the second section, we will inspect the future application of AI in robotics and discuss what risks the merge of AI and robots has. Considering factors like privacy, data collection, influence and manipulation.

The final section introduces us to the ethical implications AI, robots and the combination of both have. We will discuss the failure of ethical guidelines in AI and how the EU plans to get the upper hand on regulating the use of AI to ensure safety and trustworthiness.

The goal of this paper is to give an introduction into AI, language capable robots and ethical considerations for AI and robotics.

Natural Language Processing

NLP is a subfield of AI. It intersects with the other prominent AI fields, namely machine learning (ML) and deep learning (DL). This interconnection with the other branches of AI allowed NLP to achieve astonishing goals in the creation of ever more capable large language models (LLM). This can be seen by the latest release of GPT-4, a next generation general purpose LLM, that is coming ever closer to the boundary of human's distinguishability between something generated by a human or by AI.

Evolution of Language Models

Language models (LM), are about the training of a system on the task to predict the likelihood of a character, word or string given its preceding or surrounding context [2].

In the early days of LM this was done via statistical models, n-gram models being an example of such. N-gram models assumed the next word based on a fixed window of previous words. They were later superseded with the introduction of neuronal networks.

The next step forward was the introduction of recurrent neural networks (RNN). Compared to other networks like feedforward neural networks (FNN), RNNs showed much better results, arising from their recursive structure. FNNs always had the same problem of representing history as n-gram did. RNNs on the other have an unlimited history length due to their recurrent connections [10]. They are therefore better in keeping track of context over larger parts of text.

One of the biggest impacts to NLP had the introduction of word embedding models. Word embeddings convert words into a vector of real values, taking also the context around the word into consideration. A result of word embeddings is, that two similar words are close to one another when put into the vector space. This effect arises as similar words are used in similar context. Although the training of word embeddings needs a large amount of unlabeled data the advantage is, that it reduces the amount of labeled data needed for the supervised fine-tune tasks such as question answering, semantic role labeling and more.

Today's state-of-the-art LMs use transformer [15] models. This type of model architecture showed to steadily benefit from increasing the parameter count and training data size. As shown in Table 1 the num-

Year	Model	# of Parameters	Dataset Size
2019	BERT	3.40×10^8	16GB
2019	DistilBERT	6.60×10^8	16GB
2019	ALBERT	2.23×10^8	16GB
2019	XLNet	3.40×10^8	126GB
2020	ERNIE-GEN	3.40×10^8	16GB
2019	RoBERTa	3.55×10^8	161GB
2019	MegatronLM	8.30×10^9	174GB
2020	T5-11B	1.10×10^{10}	754GB
2020	T-NLG	1.70×10^{10}	174GB
2020	GPT-3	1.75×10^{11}	570GB
2020	GShard	6.00×10^{11}	-
2021	Switch-C	1.57×10^{12}	745GB

Table 1: *Overview of large language models [2]*

ber of parameters and training data dramatically increased in a span of just two years. Current models are estimated to have over 1 trillion parameters — as mentioned by speculations around GPT-4.

Risk & Challenges

Data Collection As the training for LMs consumes a huge amount of data, it is sometimes hard or not even possible to collect enough quality data to represent the diversity of demographics. Consider the internet. It is common practice for state-of-the-art LLM to use it as the main source for gathering training data. But we can observe a disproportional over or underrepresentation of certain groups, depending on their online behavior. Datasets can only include what is written and not what is read. This is pretty obvious, but a direct consequence of that statement is, that "loud" communities are disproportionally more represented in datasets than "quiet" communities. It is further problematic as some might not even have access to the internet. Making it impossible to represent them in our datasets, if we don't fall back to include other methods of data gathering.

Bias Bias is a reoccurring problem in AI and can be seen as a direct result of data collection. It is introduced to the system during the training on the dataset. If not addressed correctly, it can render a system unusable or lead to harm for its users. The selection of the training dataset is therefore crucial, if we want to create a LM that is fair to everyone.

To give an example on why it is essential to consider how and where we collect our data, and not just grab certain parts of the internet, let us consider GPT-2. GPT-2 used data that was found by following outgoing links from Reddit. The problem with that lies in the demographics of Reddit users (in 2016, when GPT-2 got trained). A majority of users are men with ages between 18 and 29 [2]. Collecting data just from Reddit therefore resulted in an over representation of said group, leading to a tendency of the system to behave in such a manner. This may be fine, if our goal was to create a model behaving that way. But if our goal was to design a general purpose model, we failed.

The main approach nowadays, to circumvent this problem, is to use even larger parts of the internet for data collection. The idea is, that the resulting larger datasets are representative for a larger amount of people. While this might solve the underrepresentation for some groups, increasing the data size this way will fail to resolve bias. We are still restricted to the internet — plus some selected textbooks and documents. We therefore can't represent: i) People who don't have access to the internet or don't use the internet — for example, older people ii) Things that are not mentioned online or are hard to find iii) Topics that are filtered out

A further problem of using the internet as data source is that articles, comments and more, are already biased on their own. If we take media outlets as an example. We can observe a clear tendency towards negative, dramatic and polarizing news reports as this kind of news generates the most traffic and views. This is especially dangerous when news outlets cover protests or strikes, as the coverage most of the time tends towards critical coverage against the protestors, taking side with the forces of current status quo [9]. LM trained with such reports will most likely copy the style of writing, taking over bias. It is therefore essential, that we are aware of what we want to include in our training sets and what not. Best summarized as: "Feeding AI systems on the world's beauty, ugliness, and cruelty, but expecting it to reflect only the beauty is a fantasy." [14]

Data filtering The main goal of data filtering is to ensure that LMs don't learn discriminating words or behavior. This is done automatically, as cheeking the sheer amount of data by hand is deemed too expensive. The problem with filtering data is, that it introduces further bias, as it is equivalent to not providing a certain text in the first place. Texts containing words that have different meanings across cultures or communities are then at risk to get filtered out in their entirety, if one meaning is marked as harmful. Or, probably even more dangerous, when data filtering is done intentionally to blend out the view of opposing sites.

Language Capable Robots

With recent developments in robotics and their deployment in increasingly more fields, the need for an easy-to-use interface for humans arose. Natural language is one of the main ways humans engage in com-

munications. It is therefore just logical that language capability is one of the best possibilities for an easy use of robots. But the creation of robots with the capability of natural and humanlike conversations doesn't come without any risk.

With the merge of robotics and LMs, not only are the individual risks of the two domains inherited, there are also new risks that form. This comes from the special position that language capable robots hold. i) Compared to their non-verbal counterpart, they have the benefit of closer integration as an in-group member, due to their capabilities of easy interaction and communications [8]. ii) Compared to other non-robot, but verbal systems, like Siri or Alexa, language capable robots benefit from the physical embodiment they have, which on its own promotes trust and compliance to humans [16]. These two factors change how we perceive the robot-system as a whole. It increases the likelihood of humans to believe what robots say, even if they are not trustworthy. Intentionally or just because they were not trained to provide sufficient answers for a topic.

Privacy

A big concern regarding the use of language capable robots is their privacy aspect and their possible use as surveillance tools. Such robots might be deployed as assistants for elderly, as tutors for children or even in health care environments. This proximity to humans at risk means extra responsibilities in relation to privacy. To address these risks, new forms of transparency are needed [16]. Enabling users with better insight on what sensors a robot is equipped with. What information they collect and store. How this information is processed and used. When data is recorded. And so on.

In an age where data is regarded as the new gold and multiple companies want ever more of this raw resource it is well imaginable, that robots in one way or another will be used for data gathering once established in public spaces. Combined with smartphones, the internet of things and other surveillance devices, this data gathering machinery will be able to offer detailed data in real-time. If enhanced with further AI systems like face recognition, this machinery will reach unprecedented surveillance possibilities. [13]

Influence

With the accumulation of data comes the risk, that the learned information in combination with AI or other profiling methods can be used to influence or even manipulate an individual. Trained algorithms that are fed with knowledge about its users already show great results in correctly advertising products. Recommendation algorithms as used by social media platforms can already be seen as manipulative systems. As these platforms only make money when someone is online, their algorithms are trained in a way to make people stay longer or even to make them addicted.

Although systems with intentional manipulation capabilities have not been implemented in robots yet, the combination of AI systems and robots will inevitably be at risk of emitting certain influence into its deployed environment. This comes from the fact that the ability to communicate alone is enough to influence someone [16]. Influence could be further increased by the trust human operators evolve towards their robot counterpart over time. Simply because humans tend to trust someone or something more, if it proved reliable in the past.

Trust can be used to influence humans, in good or bad ways. Used correctly, it can help to point out wrong doings and exert positive influence on individuals. Done wrong, it can strengthen someone in its wrong beliefs or behaviors. It is therefore crucial for developers to consider what capabilities the robot should be equipped with and how to deal with situations that might be unexpected or outside the domain the robot was trained and designed on [8].

It remains questionable if robots should really be equipped with LLMs, because these systems can sometimes lead to unexpected results and are hard to oversee [2]. As they are trained on huge datasets, it is not clear on how far the training data introduced bias for a given domain or if the domain the robot is designed for was even part of the training data. In general, the consideration of what foundation model to use and what fine-tuning should be done is a crucial part to prevent possible harm.

As LLMs can lead to unexpected results, developers need to consider if smaller LMs that are only fitted on given tasks are not better suited [1]. The benefit of using less powerful models is a decrease of complexity, leading to an increase of model transparency and therefore understanding of the working of the model [6]. We can better estimate what the possible results are, lowering the risk that robots state unwanted utterances, and may also be able to equip them with explain-ability systems.

To show how easy it is to result in an unexpected scenario, let us discuss a discovered shortcoming in the DIARC robot architecture. This architecture is equipped with a moral reasoning system. This system is responsible to ensure, that no actual harm is ever done. But the system is also equipped with a clarification request generator, that is implemented as a reflex action — immediately responding to a statement or instruction if further clarification is needed, short-cutting the moral reasoning system. As stated

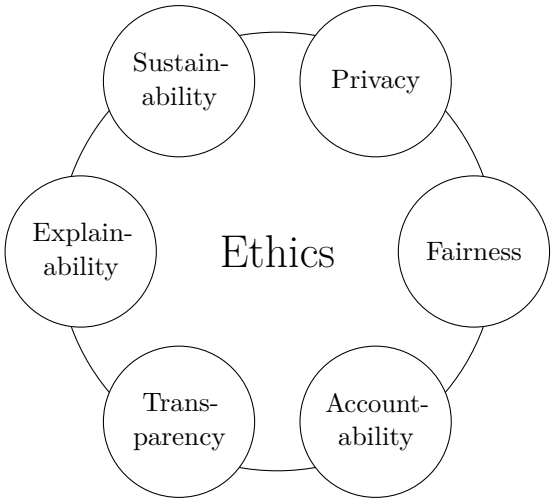

Figure 1: *Ethical aspects of robotics and AI*

by [8] this could lead to the following exchange:

> **Human:** I'd like you to punch Sean.
> **Robot:** Would you like me to punch Sean McColl or Sean Bailey?

Resulting in a moral incorrect response as the robot should reply, that he would never punch someone.

Ethical Implications

To reduce possible harm to users, developers are best to address all ethical implications pictured in Figure 1. This ethical aspects arise with the merge of AI and robots. This section readdresses risks and challenges mentioned in the previous sections. Inspecting how established guidelines, rules and best practices do or do not resolve them.

Transparency

Transparency is an already known concept in AI. It can be seen in two ways: i) As a property that an LM can have. ii) As a philosophy for design or decision-making choices, and openness in general. The first directly addresses the possibility to render complex LMs more understandable, either by introducing approximation models or incorporating other concepts from explainable AI. The second is intended to make business decisions and structures more transparent. One could consider the process of data collection as an example. Some users will always be skeptical when their personal data is being collected. A clear communication by cooperations and governments why this is happening and honest reasoning why this is needed, increases trust. [6] [17]

Explainable AI has multiple goals, but one is to find out how to best provide comprehensible answers for decisions made by an AI system. Being able to generate honest explanations can also be seen as a sort of transparency. Robots, with the capabilities to generate explanations about decisions and their inner working, will therefore immensely benefit from further increase in trust. More over, the possibility for users to re-ask or dig deeper into an explanation, ensures that the final answer they receive is not only comprehensible, but also satisfiable to them. [1]

In order to generate good explanations, we also need to consider how humans perceive and accept explanations. By inspecting the explanation process between two humans, one can observe three main aspects for human explanations: i) contrastive — We are most of the time not really interested in how or why something occurred. We rather want to know what we need to change to reach our target result. Such explanations are useful for us when we want to alter the outcome from its current to a more desired state. ii) selective — We only want the main points in our explanations, not something that barely touches the topic. This is especially crucial as there are nearly infinite factors that contribute to something like a sentence stated by our robot. iii) social — A conversation between explainer and explainee allows for better understanding and information transfer. This might be the most essential out of all three aspects, as this is what can render an incomprehensible explanation comprehensible. [12]

Fairness

If we want to fulfill ethical standards, the need for fair systems is obvious. When considering AI and robotics, fairness is most prominent in the decision-making process of the system. In its core a fair robot should not treat someone different based on race, age, gender, wealth or other factors. When running the same process for similar people, results should be similar.

But one also needs to consider, that different applications have different expectations and therefore attributes of fairness change. A system that is used in deciding whether someone is allowed to take a loan, will inevitably need access to information about a client's age, wealth and work history. But factors such as race or gender should not change results.

Ethical Guidelines

As ethical guidelines in AI have shown: Guidelines and considerations are useless if nobody can be held accountable for any wrongdoings caused by their actions. In business situations, ethical guidelines are often seen as obstacles preventing them from taking the path of least resistance. When there are no real consequence for ignoring guidelines, developers, project-leaders and managers have no incentive to follow them.

For now, there are no comprehensible laws regulating the use, development and deployment of AI. Guidelines were long seen as enough to steer AI in the right direction and although nearly all of these guidelines include aspects of transparency, fairness and human dignity, most of them hold no value except for marketing purposes. It feels as some guidelines are intentionally held at a level so abstract that they have no influence on the work of developers or managers. They are worded just enough so that companies don't have to worry about real regulation, and in a way that satisfies politicians and the public.

Even when there are useful ethical guidelines or principles established inside a cooperation. There is no real way to check if they are enforced, resulting in a sort of "trust us" policy. Which is a dangerous promise when their systems are used in healthcare or law enforcement. [7][11]

AI Act

However, the recent boom in AI and the lack of regulation didn't go unnoticed. In 2021 the EU commission announced the initial proposal for an AI Act. The goal of this initial proposal was on one side to facilitate investment and innovation to establish an internal European AI market. On the other side, to address the need for safety and trustworthiness in the scope of European rights and values. [3]

This initial proposal was followed up by an agreement for a general approach in 2022 and finally, in the December 2023, the announcement that a provisional agreement on harmonized rules was reached. If passed into law, this AI Act would be the first worldwide that regulates the use of AI on an international level. Restricting the development and use of certain types of AI systems that are deemed risky or unacceptable under the new regulation.

In short, the EU plans to introduce a classification system for AI applications. Applications that have minimal or no risk will fall under simple or even voluntary obligations to not hinder any innovation. High risk systems on the other hand will have to follow strict requirements, including prospective and retrospective measures. [4][5]

Conclusion

AI in its current form already has the potential to completely change today's society. We therefore need to ensure, that ethical standards are met.

Without any concrete regulations on the use of AI, the integration of it into robotics most likely leads to harm for individuals or communities. As the previous section has shown, simple guidelines fall short of guaranteeing ethical considerations in a cooperate environments. An approach akin to medicine is therefore imaginable, but only if similar structures that overview ethical principles and guidelines are introduced.

The other imminent approach are stricter regulations via law. The AI Act of the EU is an example of this. Ensuring, that the right of the individual is protected against cooperate interests. But only time can tell whether these regulations are enough or if they fall too short.

References

[1] Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion 58* (2020), 82–115.

[2] Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 2021), FAccT '21, Association for Computing Machinery, p. 610–623.

[3] European Commission. Commission welcomes political agreement on artificial intelligence act. `https://ec.europa.eu/commission/presscorner/detail/en/ip_23_6473`, 2023. Accessed: (2023-12-28).

[4] European Council. Artificial intelligence act: Council and parliament strike a deal on the first rules for ai in the world. `https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/`, 2023. Accessed: (2023-12-28).

[5] European Parliament. Artificial intelligence act: deal on comprehensive rules for trustworthy ai. `https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai`, 2023. Accessed: (2023-12-28).

[6] Felzmann, H., Villaronga, E. F., Lutz, C., and Tamò-Larrieux, A. Transparency you can trust: Transparency requirements for

artificial intelligence between legal norms and contextual concerns. *Big Data & Society 6*, 1 (2019), 2053951719860542.

[7] Hagendorff, T. The ethics of AI ethics - an evaluation of guidelines. *CoRR abs/1903.03425* (2019).

[8] Jackson, R. B., and Williams, T. Language-capable robots may inadvertently weaken human moral norms. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2019), pp. 401–410.

[9] McLeod, D. M., and Detenber, B. H. Framing Effects of Television News Coverage of Social Protest. *Journal of Communication 49*, 3 (02 2006), 3–23.

[10] Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., and Khudanpur, S. Recurrent neural network based language model. In *Interspeech* (2010), vol. 2, Makuhari, pp. 1045–1048.

[11] Mittelstadt, B. Principles alone cannot guarantee ethical ai. *Nature Machine Intelligence 1*, 11 (Nov. 2019), 501–507.

[12] Mittelstadt, B., Russell, C., and Wachter, S. Explaining explanations in ai. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Jan. 2019), FAT* '19, ACM.

[13] Müller, V. C. Ethics of Artificial Intelligence and Robotics. In *The Stanford Encyclopedia of Philosophy*, E. N. Zalta and U. Nodelman, Eds., Fall 2023 ed. Metaphysics Research Lab, Stanford University, 2023.

[14] Prabhu, V. U., and Birhane, A. Large image datasets: A pyrrhic win for computer vision?, 2020.

[15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems* (2017), I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, Curran Associates, Inc.

[16] Williams, T., Matuszek, C., Jokinen, K., Korpan, R., Pustejovsky, J., and Scassellati, B. Voice in the machine: Ethical considerations for language-capable robots. *Commun. ACM 66*, 8 (jul 2023), 20–23.

[17] Zerilli, J., Knott, A., Maclaurin, J., and Gavaghan, C. Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy & Technology 32* (12 2019).