Ethical AIs Need to Understand Society

Executive summary

While the usage of artificial intelligence revolutionizes various areas of life including healthcare and workforce, current models being unreliable and potentially harmful brings up a major demand for comparably capable but safe models. The industry tries to tackle this problem by firstly developing unrestricted algorithms, then aligning them to act ethically. This white paper outlines a different solution, building on ethics from the beginning: creating human-like behaviour models instead, outlining its advantages and weaknesses in feasibility, effort, irreparable flaws, and generalizability. The aim is to show that this approach in the long term is more capable of offering robust safety and lower chances of being tricked to be used for misuse. For this reason, actions have to be taken to enhance the development of models that aim to act in humane, socially responsible manners. A paradigm shift is required to focus more on researching social skill development of artificial models.

Introduction

In the blooming times of artificial intelligence (AI), the technology being applied to tackle problems in science, mathematics, economics, health sciences and even tasks in social sciences shows its immense power and applicability. Hence, the current phase of the field is called the "AI boom". The technology keeps advancing, improving performance and expanding possible use cases. Lately, generative AI algorithms such as diffusion-based models rose to fame for generating highly realistic (but fake) pictures [1], even videos [2]. In the field of text generation, ChatGPT, a transformer-based machine learning large language model (LLM) fine-tuned for responding to human-written text (e.g. questions) changed the world leaving a decade's worth of impact in a year. Anyone can quickly generate very sophisticated (although typical and detectable) essays, reviews, e-mails, messages, even programming code, saving time for busy people. Even just for brainstorming it's a quick effective tool. The technology advanced to a point where it'd be rather disadvantageous to not learn to use it (as stated by the controversial paper by Fabrizio Dell'Acqua [3], in which according to their research, for most consulting tasks and situations, productivity increases significantly). Their impact cannot be overstated, hence it's important that they bring a positive impact.

What these models lack on however, are safety and trustworthiness. A problem can be formulated: Now is a stage of development when the models are not responsible enough to be used in critical applications.

Two major categories of flaws raise the most concerns. One being unintended misinformation spreading. These come in various forms, especially for language models: hallucination, bias from data, non-lingustic limitations, and so on. Non-linguistic issues come from the fact that from just reading text the models cannot build a sense for logic, mathematics (or only to a limited extent, as without an external helper system it cannot stay inside the lines of formal mathematics). (Recently geometry problem solving boundaries were broken by AlphaGeometry [4], which is the first competition master level AI in the field, but along the language model it has, it also relies on a symbolic engine to check the validity of solutions, and the model takes a lot of tries to arrive at a correct solution). As LLMs have no deep understanding of numbers and calculations, mistakes in simple calculations are like a feature of them, they fail on basic logic puzzles, they cannot even count how many words a text has. Hallucinating in the context of LLMs is when false/made-up information is generated. A famous example was when in the first days after the release of ChatGPT, if one asked for sources for something, it started referencing articles that have never existed. This is because the models are trained on some amount of data including articles, and upon generating a response, they extrapolate some of the sources they are familiar of, but of course combining articles result in non-existing articles. Model bias typically comes from overrepresentation of classes, entities in data. A classic example is the iPhone voice assistant system, Siri

historically being more capable of recognizing the words said by Caucasian male men than other racial groups and women. This is not by chance, the causing reason is that when they trained the model many years ago, the data they acquired was represented by more people of this category than others, which resulted in the classifier (model) focusing more on classifying that category correctly, as the quantity of that category is bigger in the data, meaning more correctly classified cases. After training, this caused the said effect, and the user representative group sizes differ very much from the training representative group sizes. For some time, this was less of an issue (e.g. by adjusting learning "success" to be fair, or using synthetic data to make group sizes equal), but recently, because of the internet's bias and LLMs learning from sources of the internet and books, this problem again became prominent, and hard to handle (Google DeepMind tackled this by trying increasing the probabilities of certain categories to be generated upon generating content, which created faulty behaviour in some obvious cases, causing a big scandal [5]). A quick fix could be to make the model "think more" or verify itself when making a statement but just as it is incapable of cancelling wrong responses, it can hardly tell if its incorrect and lacks the common sense to recognize faults. Image generation and multimodal (text, image and more modes) methods are also imperfect, lacking ability to correctly display text, and also obey logic (weather and light environments unrealistic, discontinued patterns such as a table having different legs). The second major issue that appears is the models providing for the wrong people who intend to utilize it for harmful usage. One motive of perpetrators is accessing otherwise hardly accessible, highly valuable dangerous information, such as: how to break into a car, hack somebody's computer. These requests are already filtered out without providing information, these are built into the AI system, but they only go so far. Users can chain thoughts after each other to turn the model's attention away from asking for unintended behaviour, and make it provide the information they look for without restrictions. When an approach prompt is patched (fixed manually by the developers), attackers come up with new similar prompts, usually lengthening on the previous approach and rephrasing it (this sadly does not take much time for the attackers). Furthermore, researchers at the NeurIPS conference presented a method that uses LLMs to generate attempt text to access restricted information (known as jailbreak) from LLMs [6], fully automated. The paper also highlights some of the ways one can create jailbreak prompts, such as convincing the model in lengthy sentences to do something beyond the safety guardrails, or encoding the dangerous part of the prompt (e.g. in program code). Another common way attackers may use these models is to automatically generate text that can be then used to scam people, such as via phishing mails. The models in this format speed up the process of scamming, making it able for the attackers to reach out to more people.

With AI having such a large impact currently, and models have such low responsibility, solving the issue of dangers of current models will advance the field to never-seen-before heights. The industry sees the solution in medium-term as "aligning" the incorrect and biased responses from these models to ethically acceptable ones. For current damage limitations, this does help, but the correct way to fix these problems is not by countersteering towards a better direction, instead making the models go in the right direction in the first place. Arguably, this approach by design is not capable of achieving this. Rather the problem should be tackled by focusing on researching intelligent systems that better understand human and social behaviour, have common sense in communication, and ideally realize what statements may be potentially incorrect. (If one just focuses on statements made by the AI agent, alternatively, that by design can only be correct, which used to be common with chatbots that replied "I don't know" when they were unsure of something, but as this may limit the model performance a lot, this approach is less relevant today.) Such system does not have to be able to prove mathematical theorems but must have enough social awareness to recognize for what purpose one may use it and what impact could it have when making actions. Finding the ways to develop a system with these capabilities will help deploy AI systems that act more responsibly and also spread less misinformation. This technology would not necessarily fall under the "artificial general intelligence" (AGI) collective (model(s) that are comparably capable in cognitive tasks as humans) but could make a big leap towards making such systems responsible too. Positive results may lead the way to ethical social AI models (currently, social AI agents focus on recognizing just the attitudes of users such as emotions). In turn, these models when reaching their potential will be safer to deploy than current models aligned as much as possible.

Ethicality and harmlessness

The terms of ethical AI and harmless (safe) AI are used simultaneously. There is some discrepancy between acting ethically and acting not harmfully. In certain situations, acting unethically may be less harmful, such as lying with a good intent. Therefore, developing an ethical system may not (can not) be completely harmless, and vice versa. This raises the question on what the most optimal combination of the two would be, and which should be the more principal.

Leading teams currently tend to focus on safety more, and later attempting to increase the ethicality of a model without increasing the risk of harm. The other way around is also a considerable approach, developing ethical systems, and mitigating risks while staying as ethical as manageable.

The industry's approach

For a long time, the sole goal of industrial machine learning research was to build more and more intelligent and general models, improving performance. With the AI boom, some of the focus shifted to controlling AI systems, particularly after the famous AI risk statement released on May 30th, 2023 [7]. As it was the industry creating and developing the now potentially harmful algorithms and products, it's mostly their responsibility to keep it in control as well. The main goal does not get less focus even now, the companies still focus just as much on developing even higher performance models while also putting effort in developing the tools to guardrail current and future methods. Industry leading companies such as Google DeepMind and OpenAI therefore try not to limit their models in development to compete for the AGI race, still training the models that learning quickest and peak the highest (training can only go so far, after some time models start reaching an asymptote from which they cannot improve substantially), then correcting these top-level models afterwards to be ethical and safe, via human reinforcement. DeepMind co-founder Shane Legg, also a very notable figure in the reinforcement learning field (which closest mimics human-like agent behaviour) states exactly this, that they believe it is counterproductive to limit the models from the get-go, and rather limit them after their learning phase to act responsibly [8]. This is called alignment or user-alignment, described by OpenAI as "the propensity of an AI model or system to follow the goals specified by the user" [9]. In fact, OpenAI also takes the same approach, with even higher motivations, they have announced after their notable workers signed the AI safety statement to create a new team and project called "superalignment", aimed at making breakthroughs that not only control the current and possible AGI systems, but even the far away superintelligent agents (superintelligence would be marginally surpassing any human's capabilities). Currently interpretability, evaluation of output are already implemented, but the shift to alignment is speeding up. Improving safety with alignment is a continuous procedure as of current solutions (meaning that it can be done to any extent anytime, and the more aligning, the more safe it is), but as it seems now it's nearly impossible to always correct AI agents without making them learn what steps are safe to take. This is further backed by the generative jailbreak system mentioned before (which was developed to help researchers understand how to tackle jailbreaking).

One underemphasized aspect in both the development of AGI and ethical AI is the AI agent's social intelligence. Intelligence is not a well-defined concept (further discussed below) but *computer scientists neglect the social components of intelligence when building artificial models*. This is of course due to the reason that it has never been computationally feasible to train multiple models coordinating together as a "pack" to a point when results pay off (possibly the only time enough computational power was spent to get to standout performance was when OpenAI trained a network of 5 AI bots playing the game of Dota 2, for 10 months, using multiple thousand GPUs and over 150000 CPUs [10], resulting in a success as the bots cooperated flawlessly and read the playstyles of human opponents very well). Unless one can develop an environment in which a model could learn societal skills, this method is too expensive computationally. However, as this is a component that is entirely missing from current models (unlike everything else related

with intelligence, except consciousness), with potential to bring breakthroughs in developing systems that are socially aware and ethical, it is a field worth investing into more.

Mimicking intelligence artificially in different ways helps us better understand how humans are intelligent (what components make up the emergence of intelligence, and in what manner), which then can help further understand how one could build human-like awareness artificial agents. The exploration of social intelligence working in artificial agents could then help us build socially aware models.

Social intelligence

To advance socially intelligent methods, understanding intelligence and the social component of human intelligence more. Intelligence is rather an umbrella term for concepts related to acting and observing thoughtfully, than a clear expression. According to many computer scientists such as Shane Legg and Marcus Hutter [11], John McCarthy [12], and psychologists, a major challenge in creating universal artificial intelligence is finding a feasible definition for intelligence (hard to recreate something if it is not known what exactly it is). Nobody really knows what intelligence is, as Legg and Hutter say [11], if it can be described in any way. Humans sort of have an inner feeling in comparison what entities seem more intelligent than others, and even have tests to measure intelligence, but there is no consensus on definition. Differences occur between human intelligence, animal and even artificial intelligence. Legg and Hutter [11] collected the different types of intelligence tests developed for humans and animals (up to 2007, when the paper was published) and provided different definitions of machine intelligence, along with tests to compare machines' intelligence. According to them, this is the first research paper collecting and comparing the various types of intelligence tests, paving the way to research in human-like artificial intelligence by building on their findings. In reality, the paper hasn't had the impact it promised. A lot of researchers seem to disagree with looking at intelligence from their perspective, and some have doubts with the claim that the existing intelligence tests such as the IQ-test or Raven or Wechsler's test are reliable measures of intelligence, based on the fact that they are fairly stable statistically (compared to other tests) and that they are good predictors of intelligence-related topics such as academic excellence. These should not be justifying reasons for this claim. These tests point to a direction where on large scale, higher results point to considerable amount of increase in intelligence-related performance, but the test results are so "noisy" (understood here as the large amount of false positives and negatives in performance comparison based on intelligence test results) that no assumptions can be made on two individuals and how they might compare to each other, knowing their intelligence tests results, unless the difference is large. Furthermore, the authors define intelligence roughly as a measure of the ability to optimize in various goal-driven environments. This could be biased towards considering the artificial intelligence created by Hutter in 2000 [13] called AIXI, the most universally applicable and statistically optimal AI model (which would provide solutions optimal in Kolmogorov-complexity), to be the most intelligent. This model, just like most other models, starts with zero knowledge and learns step-by-step as it encounters new observations. The problem with AIXI is it learns unfeasibly slowly (in reality, it isn't even computable, but if you limit it with time and space constraints it is, even then computation time grows exponentially by increasing space size) and as the science stands today, there are no justifying reasons left to consider this as an example of a modern intelligent system. Two decades later it is clear that (as long as computing power is limited or data) there is no need to optimally predict while learning on data, it is faster to make mistakes and go back to correct them. Even if intelligence is not well understood, not only learning itself, but properties of learning must be key components of intelligence, the ability to quickly learn and be able to still improve performance even after learning for a long time. Making right decisions based on already known knowledge is rather optimal decision-making, and being able to learn anything is not enough. Definitions that only consider one or two key points for intelligence fall in practice.

Another definition mentioned by the authors was signed by 52 experts to be a feasible definition (in 1997). Originally from the article of Linda Gottfredson [14], it states: “Intelligence is a very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience”. In comparison, this definition does not naively consider a few attributes. However, it also does not highlight the social and communicational skills, showing that even among psychologists, the importance of these skills have not been recognized well enough.

Since there is no clear definition of what in general intelligence is, it may be worthy to explore what are some traits that are expected for an intelligent system to have. All presented psychological definitions of human intelligence in Legg and Hutter’s article focus on two concepts: one being learning in or adapting to various, new environments, circumstances, and the other concept being profiting, making decisions, solving problems effectively based on experience. As said before, for artificial intelligence at least, the speed of learning and the limit to what point it can effectively learn matter too, these may be of lesser importance for humans to be intelligent (may be more like a “given”). The Gottfredson definition already includes the learning abilities, and also introduces reasoning, planning, problem solving and abstraction and understanding as part of intelligence. Other personal traits that are commonly associated with intelligence are consciousness, creativity and memorization. Memorization artificially is easily implemented, it is not clear how humans memorize (but it is known that humans are superior in long-term memory to current methods) and memory is of lesser importance in building human-like systems, one can assume that possibly memory is less related to intelligence than above stated concepts. Consciousness, just like sentience, is a complex concept that is relevant mostly for humans (but many animals are conscious) which Ralph Adolphs describes as a concept humans may never understand [15]. An intelligent agent has to have some type of creativity, but this may also be less relevant for understanding intelligence, seems to be less required for artificial intelligence (models can just explore probabilistically) and is not well understood (even though Jürgen Schmidhuber gave a formal definition of (artificial) creativity in 1990 [16], but some believe researchers are yet to come up with systems that “invent” things not by extrapolating given data, brute force methods or by pure chance).

The already stated aspects are all intellectual attributes of intelligence. These are the commonly associated traits with artificial intelligence too, ranking models’ intelligence in these categories based on their performance, and even the success of a model in some cases (where performance is not easily evaluated) also depends on these traits (for an example, Chatbot Arena ranks LLMs by human evaluation scores, comparing different model responses, and humans make choices based on reasoning and such). For humans, a much broader scope of skills is associated with intelligence or success. Probably the most influential type of skills for a human aside from these are social and communication skills. In parallel, GPT-3 has existed since 2020 and gained large popularity in the field by that time, but real success came when they released ChatGPT, a very well communicative model building on GPT-3 (at that time). LLMs gained popularity, and left such an impact because of their communication skills. Yet, in any type of artificial intelligence definition, communication skills are not included (like because they are mostly only appearing in language models which are only a fraction of AI models) despite their impactfulness. LLMs by many are considered now to be the pinnacle of artificial intelligence, for this reason, communication skills should also be considered as an attribute of intelligence, for AI systems that rely on communication. On a broader scale, social skills are only considered impactful for humans and animals, but just as communication skills, social skills could enhance human-bot collaboration, therefore worthy to be considered as an intelligence factor, where agents interact with other agents (likely humans). Furthermore, socially intelligent systems shall have the ability to make decisions accordingly to the social impact of each decision. This is a key assumption of this white paper, that social intelligence may lead to theoretically socially beneficial decisionmaker models. This is why models with social intelligence are promising to research.

Extending the list of intelligence attributes with social traits helps to point to a few directions to research for developing social intelligence of models. As there are a broad range of such traits, it may be beneficial to limit to focusing on some that we can argue would be worthy in the nearby future to investigate

developing such traits for AI models. All types of communicational skills are especially important since the rise of LLMs, if LLMs were not fine-tuned to respond in such human-preferred form, the models would reply in less descriptive and less friendly manner. These models are already trained (via reinforcement learning with human feedback) to always respond in a friendly manner. This is a useful simplification, but better would be to have the model get an intuition of what style of communication would be most pleasing in certain scenarios and according to the user's (predicted) preferences. This could be a breakthrough to improving LLMs to communicate in a smoother manner, and help generalize without the need for enormous amount of data.

The model needs to develop an understanding of sentiments, feelings from the communication, have folk psychology ability, this could help notice behavioural aspects otherwise not detectable from just the sequence words. It has to notice what social behaviours are common and get used to them, adapt to them. This itself might in time (with enough scaling, i.e. computing power) make the model develop an internal common sense for itself, as agents can have different "opinions" (e.g., which action is correct to make for a decision), some being too radical and inaccurate but the most common. This idea is in parallel with averaging and regularizing machine learning methods. Practical algorithms in machine learning use some form of averaging the results (either in prediction, such as the random forest classifier averaging multiple tree classification results to make a prediction, or during learning, e.g. dropout in neural networks). Lastly, the model needs to understand the societal impact of decisions it makes. This is for cases when perpetrators would use the model for wrongdoing, but also to make the model be intelligent enough to not make decisions that could harm society (possible use-case could also be global policy making, without arising infeasible solutions, e.g. in climate change it would not make decisions that kill humans despite improving nature). Cultural awareness could be considered furthermore, although seem to require previous breakthroughs that will likely not come in near future to have a chance at successfully deploying.

This way, we can partition intelligence traits into 4 categories: goal-maximizing attributes, explorational creative skills, common sense and communication skills, and an abstract "identity senses" category. The novel research should focus on the communication skills and building a "common" sense for models. Communication is conveying any information through other participants (usually bilateral), and common sense is a built-up knowledge base based on what are the commonly accepted actions of agents in an environment. The multi-agent environments, the human-like interaction and behaviour environments for agents are still missing (with the exception of programming libraries for multi-agent reinforcement learning and simulations), and fast and reliable methods are not developed yet. These are some lines of research that need to be carried out, on large scale.

Conclusion

While more and more AI systems are deployed in highly socially impactful scenarios, increases the need to find ways to not let these models cause more harm than wrongdoing. Current alignment approaches can not provide real safety and trustworthiness, hence we need more radical steps. A recommendation is turning focus on models that learn to act accordingly from the beginning, particularly models that have potential to learn communicational, social skills and build an internal understanding of social behaviour. Practical results are missing from this field, therefore policy makers have to ensure that more research goes into social AI.

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, 'High-Resolution Image Synthesis with Latent Diffusion Models'. arXiv, Apr. 13, 2022. Accessed: Mar. 22, 2024. [Online]. Available: http://arxiv.org/abs/2112.10752

[2] A. Blattmann *et al.*, 'Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models'. arXiv, Dec. 27, 2023. Accessed: Mar. 22, 2024. [Online]. Available: http://arxiv.org/abs/2304.08818

[3] F. Dell'Acqua *et al.*, 'Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality', *SSRN Electron. J.*, 2023, doi: 10.2139/ssrn.4573321.
[4] T. H. Trinh, Y. Wu, Q. V. Le, H. He, and T. Luong, 'Solving olympiad geometry without human demonstrations', *Nature*, vol. 625, no. 7995, pp. 476–482, Jan. 2024, doi: 10.1038/s41586-023-06747-5.
[5] N. Grant, 'Google Chatbot's A.I. Images Put People of Color in Nazi-Era Uniforms', Feb. 22, 2024. [Online]. Available: https://www.nytimes.com/2024/02/22/technology/google-gemini-german-uniforms.html
[6] P. Chao, A. Robey, E. Dobriban, H. Hassani, G. J. Pappas, and E. Wong, 'Jailbreaking Black Box Large Language Models in Twenty Queries'. arXiv, Oct. 13, 2023. Accessed: Mar. 23, 2024. [Online]. Available: http://arxiv.org/abs/2310.08419
[7] Center for AI Safety, *Statement on AI Risk*. 2023. [Online]. Available: https://www.safe.ai/work/statement-on-ai-risk
[8] Dwarkesh Patel and Shane Legg, *Shane Legg - Path to Artificial General Intelligence, New Architectures, Aligning Superhuman Models*. [Online]. Available: https://www.dwarkeshpatel.com/p/shane-legg
[9] Y. Shavit *et al.*, 'Practices for Governing Agentic AI Systems'. [Online]. Available: https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf
[10] OpenAI, 'Dota 2 with Large Scale Deep Reinforcement Learning', Dec. 13, 2019. [Online]. Available: https://cdn.openai.com/dota-2.pdf
[11] S. Legg and M. Hutter, 'Universal Intelligence: A Definition of Machine Intelligence'. arXiv, Dec. 20, 2007. Accessed: Mar. 17, 2024. [Online]. Available: http://arxiv.org/abs/0712.3329
[12] J. McCarthy, 'What is artificial intelligence?' Nov. 12, 2007. [Online]. Available: http://jmc.stanford.edu/articles/whatisai/whatisai.pdf
[13] M. Hutter, 'A Theory of Universal Artificial Intelligence based on Algorithmic Complexity'. arXiv, Apr. 03, 2000. Accessed: Mar. 17, 2024. [Online]. Available: http://arxiv.org/abs/cs/0004001
[14] L. S. Gottfredson, 'Mainstream science on intelligence: An editorial with 52 signatories, history and bibliography.', *Intelligence*, vol. 24, no. 1, pp. 13–23, 1997, doi: 10.1016/S0160-2896(97)90011-8.
[15] R. Adolphs, 'The unsolved problems of neuroscience', *Trends Cogn. Sci.*, vol. 19, no. 4, pp. 173–175, Apr. 2015, doi: 10.1016/j.tics.2015.01.007.
[16] J. Schmidhuber, 'Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990–2010)', *IEEE Trans. Auton. Ment. Dev.*, vol. 2, no. 3, pp. 230–247, Sep. 2010, doi: 10.1109/TAMD.2010.2056368.