Remarks on the Ethical Behavior of Artificial Intelligence

Zafeirakopoulos Dimitrios
NTUA
Athens, Greece
dimzaf252@yahoo.gr

Stefaneas Petros
NTUA
Athens, Greece
petrosstefaneas@gmail.com

KEYWORDS

Artificial intelligence, ethical framework, AI medical assistant, responsibility

1 BIOGRAPHICAL SCETCHES

Zafeirakopoulos Dimitrios is a PhD student at the National Technical University of Athens. He received his graduate Degree from the National Technical University of Athens with the diploma thesis "Development of Models for Ethical Behaviour of Artificial Intelligence through use of Deontic Logic" in 2018. He participated and was a speaker at the Rules: Logic and Applications 2018 (Athens, NTUA Campus, 19-20 December 2018) workshop, where he talked on "Deontic Logic and Robot Ethics". His research interests include artificial intelligence, logic, formalisation and various programming languages.

Prof. Petros Stefaneas co-ordinates the Logic and Formal Methods Group (λ-ForM) of the AALL Laboratory. He is an Assistant Professor at the Department of Mathematics of the School of Applied Mathematical and Physical Sciences at NTUA.

2 FULL PAPER

Artificial intelligence can inspire a great deal of hope as well as fear among researchers and the general population alike. As it is a new, rapidly growing area of technological innovation, we are not yet aware of the degree towards which it will help shape our culture and society at large, given that we have only just begun seeing its impact. Naturally, the areas that can be improved by such a unique and multi-faceted piece of technology are various and involve different fields. These can range from the way we do our searches online, through the world wide web, how data and pieces of information are presented to us, often in ways tailored to our individual preferences, to robotic assistants and workers in dangerous fields, such as doing manual labour in highly hazardous environments and conditions, in addition to fields which require constant presence and vigilance, without allowing for breaks and interruptions. Such a case could be a mechanical assistant supervising and offering help to patients in a medical hospital. The general population can have quite the number of concerns regarding artificial intelligence. An important reason for that is that how artificial intelligence agents function and what they can achieve is a very complex matter, very much unlike more everyday household appliances that are used commonly and thus, the way AI functions is not something that can be adequately explained to the average user. Furthermore, one must not neglect to acknowledge the impact that popular culture has had on this issue. Popular culture, usually in the form of books and films, has definitely had an impact on how the public perceives and at times fears artificial intelligence and though such fears might not be grounded in reality, as they stem from the realm of fiction, they must be nonetheless taken into consideration when considering how to address the public's fears. What we would like to achieve, is to find a way to counter the public's fears. After understanding and examining the causes that give rise to these concerns regarding artificial intelligence, we conclude that in order to counter such understandable fears, we would need two equally important pillars when deciding how to develop proper artificial intelligence agents. These two pillars are firstly, a clear understanding of what we would like each individual artificial intelligence agent to achieve, what goals it would satisfy specifically and secondly a method by which to communicate to the general population that any given artificial intelligence agent will have sound behaviour. Sound, correct behaviour applies to how each AI would act, both with regard to ethical as well as legal obligations we want it to respect. The general goal we have for each given artificial intelligence agent is actually not different from what we would want from any other piece of technology that gets developed and used, whether by the scientific community or the general public. Specifically, we want it to fulfil its function as it was designed, satisfying each and every goal we might have set for it, while also not crossing any boundaries and thus engaging in forms of behaviour that ought to be prohibited. Such behaviour might be types of illegal actions or in general any behaviour that could be considered harmful and end up violating the rights of a potential user, or, furthermore, any other affected party. This is a task that can be considered particularly difficult, due to the fact that we expect that artificial intelligence agents will grow and thus keep developing new approaches to tackling the goals we might set for them. That means that ensuring that they do not cross any boundaries regarding what is acceptable and what is unacceptable, both when it comes to ethical obligations and to legal obligations, will not be an issue that will be tackled once during the development process and then considered solved after the development of an AI agent is finalized, as is the case with more mundane pieces of technology, such as everyday electrical appliances. On the contrary, ensuring that an AI has solely correct behaviour, is going to be a continuous task that will remain of vital importance during the whole period that any given artificial intelligence agent continues to operate and develop new methods towards tackling its goals, as any new method might involve an ethical or legal risk that wasn't considered earlier. One method to help us better understand artificial intelligence agents and limits they have, as well as what expectations we can set for them is found in the works of James H. Moor. James H. Moor's approach to the ethics of robots introduced different AI categories in relation to ethics. As such, there is a range including four different types of artificial intelligence agents.

WAIEL2020, September 3, 2020, Athens, Greece

The simplest one includes agents that either intentionally or unintentionally have ethically correct impact, called "Ethical Impact Agents". An example of such an agent could be a watch that which helps a person be punctual in their appointments. Moving on, we have "Implicit Ethical Agents". These would be machines that are constrained to avoid unethical outcomes. The following category is the one that we deem more interesting when it comes to actually designing artificial intelligence agents while trying to ensure they behave properly in ethical and legal terms. That would be "Explicit Ethical Agents" [1]. What separates this category from the previous one is that Explicit Ethical Agents are not simply constrained to avoid unethical outcomes, but instead they explicitly strive towards ethical behaviour and furthermore provide proof of that fact with the algorithm that defines the way they function and operate to fulfil the goals we set for them while also being ethically sound. There is also a fourth, final category. That would cover artificial intelligence agents labelled "Full Ethical Agents" [2]. These are hypothetical machines that would be ethical in the same way that humans are. This would mean they would have consciousness, free will and intentionality. Such machines would of course also be able to pass the Turning test. [3] One question we definitely need to consider is, when a given artificial intelligence agent, much like any other, more or less complex, piece of technology might, inevitably malfunctions and causes minor or major harm, who should be held responsible for what happened? We will present an example to help explain the types of problems and questions that might arise from a malfunctioning AI better. The example we choose to present is that of an AI medical assistant in a hospital. Said artificial intelligence agent supports patients by giving them the medicine they need while also respecting patient autonomy. In the event where a patient refuses to take the necessary medication and the AI, programmed to also respect the patient's autonomy and respect their wishes, chooses to withhold the medication, as the patient requested, we might be led to the patient dying, something that obviously ought to be considered a major malfunction in the AI's programming. A question that immediately arises in such a situation is who would be legally and ethically responsible for the death of the patient in a scenario like the one described above? There are many potential answers to that question. One could be that the patient themselves are responsible, due to their choice to not take the medication they needed to stay live. Perhaps the hospital that utilizes such an AI for patient supervision should be held responsible for what happened. Another potential answer is the company that created the AI and decided how it would function when designing it is responsible. Could perhaps the AI itself be held responsible? Such a question might indeed arise as AI becomes more prominent in our society and of course, it is one legal science ought to tackle and address. One can't help but wonder where the inevitable development of more complex artificial intelligence agents might lead, with regard to ethics and legality. If artificial intelligence agents were to be held responsible for their actions, will we potentially need to grant them rights too? In situations like the one described above, we no doubt need to be able to figure out who is to blame for what happened, but it is also critically important to be able to verify exactly what caused the malfunction that resulted in an artificial intelligence agent to act in a way that crossed legal or ethical boundaries. In order to be able to achieve that efficiently we should be striving towards developing artificial intelligence agents that go beyond just fulfilling their obligations regarding ethical and legal matters. We should aim towards developing AI that succeed in the above, in a way that allows us to examine the process by which they make decisions and give us the opportunity to analyze what causes them to make choices in the way they do. That would mean giving us a clear view on each artificial intelligence agent's "thought" process. This would allow us to achieve two very important goals. Firstly, it would let us find out exactly what went wrong in the event of a malfunction and thus be able to improve the AI, in order to ensure such mistakes do not happen again. And secondly it would allow us to be able to accurately place blame for the mistake. Furthermore, artificial intelligence agents that operates in a way that communicates their inner workings and is thus more transparent in their function could be more easily accepted by the public, as the visibility of their "thought" process would allow us to provide explanations as to why we expect a given AI to stay within the desired parameters we have set for it. Let us revisit our previous example, that of a medical assistant AI in a hospital that engaged in behaviour that led to a patient under its supervision dying. If said artificial intelligence agent operates based on an algorithm which also provides an ethical framework for its behaviour, we would be able to more efficiently verify what caused the actions that led to an unethical or illegal outcome, like the death of a patient, and thus be able to accurately place blame. More specifically, if, due to the AI's design, letting the patient die was something the medical assistant considered ethically acceptable, then we can conclude that the company that designed it is at fault, due to the problematic ethical parameters that were set for the AI. However, if the AI considered the patient dying an ethically unacceptable outcome, but it also strived towards maintaining patient autonomy and respecting the patient's rights and thus the solution it found for the dilemma was to inform a doctor and ask for their help with the situation, we can conclude that the hospital is ethically and legally responsible as it failed to act upon the information received from the AI to save the patient. Of course, such an example is theoretical and doesn't cover all eventualities. In a real life situation, there would be many more variables that ought to be considered when analyzing what happened. Nevertheless, even such a small-scale theoretical example can show how an AI having an ethical framework, can help us pinpoint design or operational flaws more easily and thus place blame for violations of ethical or legal rights as well as improve the AI in question. Therefore, we propose that any algorithm that operates an AI must also include an ethical framework in order for its function to be able to be evaluated fully.

3 CONCLUSIONS

Due to the rapid growth of artificial intelligence and the impact it has on our lives, we need to be prepared to tackle any problems it might cause, both in ethical and legal matters. Towards that end, utilizing algorithms that include an ethical framework will indeed help us ensure correct AI behaviour and allow us to more easily pinpoint where malfunctions stem from and who is to blame for unwanted outcomes.

4 REFERENCES

[1]Zafeirakopoulos Dimitrios, “Development of Models for Ethical Behaviour of Artificial Intelligence through use of Deontic Logic”, Diploma Thesis
[2] James H. Moor, “Four Kinds of Ethical Robots”
[3] Alan Turing, “Computing Machinery and Intelligence”