Ideas for a usable pro-social data set

willmarechal87@gmail.com

This challenge is about curtailing AI actions in what is considered moral or ethical by us humans. As many, including the organizers of the challenge have put forth, AI is in its infancy, and it is up to us, the "adults", to teach it "good manners and good morals", with the aim to enable kinder machines. The challengers are asked about ideas on how to generate the best set of pro-social examples, on which to apply machine learning.

With AI reaching in every domain of human activity, and being tailored for the task at hand each time. The representatives of the AI will be plural in their actions, cognitive and processing capabilities. As Wallach and Allen [1] propose, they should be regarded whether they are more operational morality or functional morality, or as I would call them "workers" and "central processing units" or central AI, including a moral agent [2] [3] [4] [5]. The first follow a pre-design moral architecture and the second are capable of "moral reasoning", making them AMAs—artificial intelligent machines capable of moral reasoning [4]. The "workers" refer to relatively cheap AIs with minimal sensing & processing capabilities, that undertake most of the daily/routine tasks and the central AI for rarer and changing/evolving tasks and could have to oversee or control the workers, as discussed by Moor [2]. Furthermore, we have here two ends of a spectrum and they will also have to be subcategorized following their applications and data collection capability. For each type/families of robots we can define their general interactions and which responses we deem appropriate or not. This requires to know what the robot can do, and what type of information does he has access to.

Wallach and Allen [1] have given 3 methods or ways to implement this moral guiding system: the top-down or direct programming track, bottom-up or developmental approaches, and the hybrid of these two.

- Regarding the bottom-up method, robots can communicate between each other, they can be in communication with a central AI with a larger capability, like having security cameras, with which it can evaluate the behavior of the "workers" and inform them on their behavior. The data can then be shared over a larger network with other AIs in different locations, creating a shared prosocial dataset. The behavior evaluation could be based on image recognition of sentiments and/or facial expression of nearby humans, or it could be based on social network aggregation of news or reactive post relating to the actions of the "workers". An example of this social network aggregation is the categorization of internet news in Pro and Con news for a same event by the website www.knowherenews.com [6] using an AI, which then generate 3 views/stories of the event: Positive, Negative & Impartial. A similar job can be done to rate the behavior of an AI "worker".
- In the top-down method, data sets or "Asimov-like" laws will either have to be set by the legislative power or an official constructer policy. But, with new functionality being implemented at an alarming rate, these codes of conduct will either be very general or

only implemented after an incident or else they will be fought in the courts as not to slow access to market for new products by private interest. Unforeseen usages of the AI/robot can emerge from consumers, creating an ethical problem for the AI will not have at his disposal data sets, rules or laws in order to guide/evaluate its behavior.

- Law and policy are a lengthy process, and will lag behind AI innovation and the constant evolution of the AI usage. This calls for a layered behavior code, the hybrid method. We humans have several degrees dictating our behavior. On top we have the official law of the land, then we have our religious, spiritual law or morale, the rules of conduct of our social group and then we have the agreed upon which are in use for a short period. ALL of them evolve, but a different rate. Official laws can take several years to evolve, while the "agreed upon" change dynamically all the time. The AI could follow general ethical rules set by the legislative or constructer and tune their behavior with the shared data sets on the network and feedback from the central AI, which could enact local guide who is to be audited by human officials in order to upgrade the law.

The central AI or supervisor could compare with its database and hopefully find enough similar cases to decide on the rule and it could fail to do so. But if humans are present in the vicinity, try an approach and modify it in regards to the reaction of the humans by like facial expression recognition. Actually, this should always be a fail-safe for AI operations.

Now how to consider this "AI pro-social behavior" in operations? For the behaviors that violate the rules or laws, its simple, we just forbid them and the AI will know which type of actions leads to those forbidden results and not do them. But what about the "not yet ruled", emerging and "not thought of" behaviors? They fall under the dynamic, bottom-up scenarios where the AI has to guess what to do based on incomplete data. The simple answer is "don't do unless covered by law". But that will stop in its track many robots and limit the innovation industry who is better known for its "not explicitly forbidden, let's do it" approach. Several authors [7][8][9][10] have proposed a reward mechanism recording the ratting of past behavior of the AI/robot/agent and seeks to optimize the rating of the present action by choosing among the ones with the best record. This is a good strategy and could pull event records for the shared data set network.

But human societies are different from one another, and they evolve. A good behavior today isn't necessarily good tomorrow or yesterday. An agent trained with data from one part of the world isn't necessarily going to do well in another part of the world. Massive amount of data from each parts of the world would be needed. This is costly and not necessarily available. When we humans conduct ourselves in society, we module our behavior depending on the situation, on the context. We don't behave the same on a busy day, on a calm day us when we hear that there is social unrest or discontent. In order for the examples to be used across the maximum number of regions and contexts, the example must not only contain a description of the action of the AI and whether or not it is a "pro-social behavior", but it should also be classified by the "cultural context", and a "social stress level" of the current area. The "cultural context" could be determined by the GPS coordinate, the state/country and the identification of nearby buildings (church, bank, school, residence, stadium…). The "social stress level" can be determined by using a classifier on available news outlets, just like www.nowherenews.com [6] is doing for its PRO/CON articles.

In conclusion, regarding your challenge for ideas on how to build data sets in order to achieve a "pro-social behavior" for the AI. I suggest it contains: a description of the AI performing the action (sensing & mobility capability), a description of the action, the "cultural context", the "social stress level" and whether or not it is considered a pro-social behavior.

Bibliography:

[1] Wendell Wallach and Colin Allen, *Moral Machines: Teaching Robots Right from Wrong,* New York: Oxford University Press, 2009.

[2] James H. Moor, "The Nature, Importance, and Difficulty of Machine Ethics," in *Machine Ethics*, ed. by Michael Anderson and Susan Leigh Anderson, New York: Cambridge University Press, 2011, 13-20.

[3] Anderson, Michael and Susan Leigh Anderson, *Machine Ethics,* New York: Cambridge University Press, 2011.

[4] John Sullins, "Artificial Moral Agency in Technoethics," in *Handbook of Research on Technoethics*, ed. by Roccio Luppicini and Rebecca Adell, Hershey: IGI Global Information Science, 2009, 205-221.

[5] R.J. M. Boyles, A case for Machine Ethics in Modeling Human-level Intelligent Agents, KRITIKE, 2018, V.12, number 1, June, 182-200

[6] nowherenews, 2018, URL: http://www.nowherenews.com

[7] P. Abbeel, and A. Y. Ng, Inverse reinforcement learning. In Encyclopedia of machine learning, 2011, Springer US, 554-558.

[8] A. Y. Ng, and S. J. Russell, Algorithms for inverse reinforcement learning. In Icml, 2000, 663-670.

[9] N. C. Almeida, M. A. Fernandes, and A. D. Neto, "Beamforming and power control in sensor arrays using reinforcement learning". Sensors, 15(3), 2015, 6668-6687.

[10] U. Kose, A. Pavaloiu, Dealing with machine Ethics in Daily Life: A View with Examples, The 5th International Virtual Conference on Advanced Scientific Results, www.scieconf, 2017,