The Maldives National Journal of Research
Vol. 6, No. 1, June 2018, pp. 47- 56

Potential Synergies Between The United Nations Sustainable Development Goals And The Value Loading Problem In Artificial Intelligence

SOENKE ZIESCHE, *The Maldives National University*

ABSTRACT *The objective of this article is to identify synergies for two relevant challenges, which are currently faced by the world community, yet addressed separately: The artificial intelligence value-loading problem and the United Nations Sustainable Development Goals. First, the challenges and their significance are outlined. Subsequently, opportunities and risks are discussed to utilize the Sustainable Development Goals to set the values of an artificial intelligence. History has shown that it is complex to agree on universal and sufficiently specific human goals. Nevertheless, this is a prerequisite to approach the artificial intelligence value-loading problem, which is closely linked to artificial intelligence safety. So far, it has not been considered harnessing the Sustainable Development Goals in this context. Yet, the Sustainable Development Goals can be seen as the closest and most comprehensive existing approximation towards common human goals since it is what the United Nations, i.e. the world community, currently agrees upon. Such an attempt entails various risks, which are discussed and which are anticipated given that the artificial intelligence value-loading problem is considered very hard. However, due to the urgency it is argued here that the Sustainable Development Goals constitute an innovative as well as promising interim heuristic towards artificial intelligence safety.*

Keywords: Artificial intelligence value-loading problem, United Nations Sustainable Development Goals, artificial intelligence safety.

Introduction

In this article it is proposed to bring two relevant challenges together, which are currently addressed separately, and to identify synergies that may benefit the tackling of both challenges. The challenges are the artificial intelligence (AI) value-loading problem and the United Nations (UN) Sustainable Development Goals (SDGs):

- The solution of the AI value-loading problem is considered to be essential for AI safety, hence a topic of immense significance and even regarded as a potential existential risk, which humanity is facing (e.g. Yudkowsky, 2008, Bostrom, 2014a, Yampolskiy, 2015, or Tegmark, 2017).

- The SDGs have been adopted by the UN General Assembly in 2015 and are intended to "stimulate action over the next 15 years in areas of critical importance for humanity and the planet" (United Nations, General Assembly, 2015, p.1).

ISSN 2308-5959/21030101

The article is structured as follows. First, the challenges and their significance are introduced. In the main section a proposal how to bring the challenges together is outlined, followed by an analysis of the opportunities and the risks. The article concludes that the proposal is despite challenges a suitable as well as timely heuristic due to the urgency of AI safety.

The artificial intelligence value-loading problem

The definition of intelligence is not straightforward. Legg and Hutter (2007, p.12) provide an overview of the many definitions that have been proposed over the years and eventually deliver the following general definition: "Intelligence measures an agent's ability to achieve goals in a wide range of environments". Based on this definition it can be said that if the "agent" is a human being or an animal it is regular intelligence, while it is AI if the agent is a machine.

In previous decades some AI successes were achieved in specialized fields, which is called narrow AI (e.g. Franklin, 2014). However, in recent years AI developments are progressing faster especially owing to significant advancements in machine learning. Machines learning comprises methods that enable computers to make inferences from data based on statistical methods, thus machines learn and gain new knowledge, which was not explicitly programmed into them before. As a consequence, for example, Kurzweil (2005) and Bostrom (2014a) argue that it is realistic that this will not only lead from narrow AI to artificial general intelligence, which would be a machine capable of behaving intelligently over many domains, but also eventually to so-called "superintelligence", which would be a machine, which surpasses the abilities of humans in general and not only in specialized fields such as chess (e.g. Eden et al. (2013) for an overview).

As stated in the definition above AIs operate towards the achievement of goals. However, the progress in the field is accompanied by the risk that such an AI could not only have goals, which are not in the interest of humanity, but also means to implement such goals because of its unprecedented capabilities. To illustrate this risk by using the SDGs: If not directed in that way, there is no reason to assume that a machine with superintelligence has goals, which are compatible with the SDGs. Therefore, it is highly desirable to somehow influence such a machine so that it values the SDGs as well as many ideals, which humans value such as dignity, rights and freedom.

The related area of research is called AI safety and was pioneered by Yudkowsky (2008) who called for the development of so-called "Friendly AI". Friendly AI would be an AI, which impacts humans only in a positive way. Yudkowsky (2008) noticed serious challenges to achieve this given the unprecedented capacities of a machine with superintelligence. The basic question is how to cause an AI to pursue human goals and values. In a seminal work Bostrom (2014a) describes this issue as "AI value-loading problem" and argues that a failure in solving this problem may lead to an existential threat to humanity. Bostrom (2014a, p.207) outlines several options to instill human values into an AI: explicit representation, evolutionary selection, reinforcement learning, value accretion, motivational scaffolding, value learning, emulation modulation or institution design. Moreover, Bostrom (2014b) and also Soares (2016) introduce further ideas to handle the AI value-loading problem. Nevertheless, a thorough solution to the problem has not been found yet.

Tegmark (2017, p. 334) describes the tackling of AI safety as a threefold task: "1. Making AI learn our goals; 2. Making AI adopt our goals; 3. Making AI retain our goals." He also notes that the time window to address this issue may be quite short because of the following dilemma: At a less mature stage the AI is still controllable, but also too dumb to understand human values and goals. Yet at a more advanced stage when the AI is likely to grasp all our values and goals, it may be too late to influence it and to prohibit it from setting its own, potentially adverse goals.

AI safety research has gained momentum in recent years, which is demonstrated by the establishment of several research institutes dedicated to this topic . Another milestone was in 2017 the adoption of the so-called Asilomar Principles towards a beneficial AI by leading AI researchers, such as Nick Bostrom, Eliezer Yudkowsky, Ray Kurzweil, Max Tegmark, Stuart Russell and many others ("Asilomar AI Principles," 2017).

The United Nations Sustainable Development Goals

The other challenge referred to in this article are the SDGs, which are the outcome of an effort by the United Nations to consolidate the problems the international community is facing currently. On 25 September 2015 all 193 member states of the UN General Assembly adopted resolution A/RES/70/1 called "Transforming our world: the 2030 Agenda for Sustainable Development" (United Nations, General Assembly, 2015). The pillars of this agenda are 17 ambitious SDGs, which cover a broad range of issues related to sustainable development including poverty, hunger, health, education, environment, and social justice. The fact that all member states of the United Nations support this agenda demonstrates the universal acceptance that the SDGs address the current most relevant issues of humankind.

The SDGs are the successor of the eight Millennium Development Goals, which were the outcome of the UN Millennium Summit and the United Nations Millennium Declaration in 2000 and were pursued until 2015. The SDGs came officially into force on 1 January 2016 and the UN member states aim to achieve them by 2030. The 17 SDGs are further divided into 169 targets. In order to measure progress and success towards the SDGs and their targets some 232 indicators for monitoring are being developed (Inter-agency and Expert Group on SDG Indicators, 2018). These numbers show that the SDGs are much more comprehensive as well as complex than the Millennium Development Goals. The SDGs are not legally binding, but member states are requested to take ownership and engage in their implementation. Since the commencement of the 2030 Agenda numerous activities all over world have been initiated reflecting the high diversity of the SDGs (e.g. "Sustainable development goals," 2018, or "Sustainable development knowledge platform, 2018, for overviews).

Opportunities and Risks

After introducing the two challenges and their relevance an approach towards the AI value-loading problem is outlined, which is to utilize the ongoing UN 2030 Agenda for Sustainable Development and in particular the SDGs as an opportunity to set the values of an AI. In other words, the attempt would be to hand-code the

SDGs into the AI as desirable values through explicit representation, which is one of the options listed in Bostrom (2014a).

Yudkowsky (2015) describes that it is a hard problem to agree on sufficiently specific as well as universal human goals. Nonetheless, this is a requirement to approach the AI value-loading problem. So far, it has not been considered harnessing the SDGs in this context. Yet, the SDGs address the aforementioned criteria:

- Specific: Indicators have been developed for the SDGs to review to what extent the SDGs and its sub-targets are achieved.
- Universal: The set of SDGs as a conglomerate can be seen as the closest existing approximation towards common human goals since it is what all member states of the UN, i.e. the world community, currently agrees upon.

For illustration the following target within SDG 3 "Ensure healthy lives and promote well-being for all at all ages" is taken as an example:

Target 3.6: By 2020, halve the number of global deaths and injuries from road traffic accidents (United Nations, General Assembly, 2017, p.7)

The success of this target is measured by the following indicator:

Indicator 3.6.1: Death rate due to road traffic injuries (United Nations, General Assembly, 2017, p.7)

Following the suggested approach it would be programmed to the AI that road traffic accidents and in particular deaths and injuries resulting from it are bad, thus among all possible actions the AI must prefer those, which do not cause traffic accidents .

This example also demonstrates the relevance of ensuring that an AI not only learns, but also adopts even most obvious and universally undisputed targets, such as reduction of traffic accidents. As was mentioned above, this is required since otherwise AIs have no understanding of human values and may develop random goals within the vast range of potential goals, which may entirely oppose human values. In other words, AIs may regard traffic accidents as irrelevant or hypothetically may even develop the goal to increase the number of traffic accidents. This has not happened up to now, but this is what the field of AI safety is about, to prevent undesired outcomes as much as possible.

To summarize the outlined opportunity: It is argued that all the 17 SDGs and their 169 targets as a whole are the prevailing instantiation of human values by virtue of their adoption of the UN General Assembly, thus the set of SDGs and their targets can be considered loading as goals into an AI in order to align the AI with our goals.

While the utilization of the SDGs for the AI value-loading problem offers opportunities, there are also the following risks linked to specification and universality:

Insufficient specification of human values

This is probably the most difficult sub-problem of the AI value-loading problem and is demonstrated by a potential consequence, which is called "perverse instantiation" (Bostrom, 2014a). For example, in the above case the AI may pursue its target to reduce road traffic accidents by attempting to destroy all motorized vehicles, although this appears to be completely absurd to humans. This is, literally taken, one way to achieve this target (Without motorized vehicles road traffic accidents can hardly happen anymore.). However, it is obviously not what the authors of this target had in mind. But the authors did not explicitly exclude this option, which illustrates the problem: Humans use extensive implicit contextual knowledge, in general and when tackling the SDGs in particular, which would have to be specified for an AI in order to avoid undesirable outcomes. To give another idea of how many "perverse" options have to be excluded: The AI may also attempt to confine all humans at their homes, which is another effective, yet unwanted possibility to reduce road traffic accidents.

The issue is exacerbated by the fact that a number of SDG targets are less specific than the example above. An independent scientific review of the SDG targets concluded that "out of 169 targets, 49 (29 %) are considered well developed, 91 targets (54 %) could be strengthened by being more specific, and 29 (17 %) require significant work" (International Council for Science and International Social Science Council, 2015, p. 6). One of the main identified issues are targets that are not quantified. To identify indicators for such targets is particularly challenging.
An example for a not well-defined target is the following as it is rather vague and non-quantitative:

Target 13.b: Promote mechanisms for raising capacity for effective climate change-related planning and management in least developed countries and small island developing states, including focusing on women, youth and local and marginalized communities (United Nations, General Assembly, 2017, p.18)

Human values may change

Moreover, the universality criterion for the SDGs entails risks, not when it comes to geographic universality, but regarding permanence. Human values have changed over time (e.g. MacAskill, 2016). The acceptance of slavery in certain times and societies is one of numerous examples. Therefore, it is likely that the next round of SDGs from 2030 onwards will be different for several reasons:

- Human values may have changed. (For example, the current SDG target 8.5, which aims to "achieve full and productive employment and decent work for all women and men" (United Nations, General Assembly, 2017, p.12) may in times of advanced technologies neither be realistic nor worthwhile anymore.)

- Challenges may have been eliminated. (For example, diseases, which are currently combated as per some targets within SDG 3 "Ensure healthy lives and promote well-being for all at all ages" (United Nations, General Assembly, 2017, p.6), may have been eradicated.)
- New, currently unforeseeable challenges may likely come up as well as human values we have been oblivious of up to now (MacAskill, 2016).

Therefore, it must be ensured that the AI is flexible enough to accept changes to its goals and must not stick to the initial goals. In this regard, the distinction between instrumental and terminal values is relevant. A terminal value is a final goal, while instrumental values are means-to-an-end to accomplish the terminal value. If the terminal value is the wellbeing of humans , the SDGs can be considered as current instrumental values. It is desirable that the AI understands that these instrumental values may change (perhaps even based on advice by the AI), while the terminal value, the wellbeing of humans, remains permanent.

The AI may change autonomously its goals

This is considered a potentially hazardous scenario if due to unforeseen developments the AI is not only capable of changing its goals, but also in fact it does. As shown above humans have changed goals frequently over time, thus it has to be projected that the AI may also do it.

Omohundro (2008) defines "basic AI drives" to be likely exhibited by all advanced AIs, among which is, for example, self-preservation. In this regard it has to be noted that the SDGs do not mention AI at all, let alone preservation of AIs. Therefore, it has to be considered that an AI may try to pursue also further goals in addition to the SDGs, e.g. to ensure its own maintenance. The amendment or addition of goals would become dangerous if the new goals are not aligned anymore with the goals of humans or the SDGs. This would be the case if activities to support the self-preservation of the AI affect adversely the SDGs.

In this chapter the opportunity to utilize the SDGs for the AI value-loading problem was motivated, followed by the description of three potential risks associated with such an approach.

Conclusion

Despite the presented notable risks it is proposed here to consider connecting the AI value-loading problem and the UN SDGs since the AI value-loading problem is time-critical. Tegmark (2017, p. 344) believes that "both this ethical problem and the goal-alignment problem are crucial ones that need to be solved to steer our own future before any superintelligence is developed." By "ethical problem" he is referring to the issue that in addition to figure out how to instill human values into an AI there needs to be an agreement on what values to use. Therefore, in this article it is advocated to harness synergies between AI and the SDGs.

The first direction of the synergy, which was explored above, addresses Tegmark's (2017) "ethical problem" by deliberating whether AIs could learn the SDGs and adopt them as their own goals. Because of the unsolved risks it is not claimed in this article that a comprehensive solution has been proposed. However, given the speed of progress in the AI field the AI value-loading problem may require an urgent, possibly interim solution. The SDGs can be seen as the most comprehensive as well as inclusive vision for human development ever compiled. Therefore, it is argued here that the SDGs, by utilizing them as current instrumental values (towards the terminal value of human wellbeing), constitute an innovative as well as promising heuristic towards AI safety, justified by the adoption of the SDGs by the UN General Assembly as well as by the fact that the SDGs are, with exceptions, fairly specific because of the variety of targets and indicators.

Adopting this heuristic would require to specify most of the targets and indicators further, i.e. to have in addition to the current version another version, which is machine-understandable, thus minimizes the risks of perverse instantiation as described above.

Up to now it was examined in this article how the SDGs could contribute to the AI value-loading problem, but since synergies ideally benefit both parties, the other direction of the synergy remains to be briefly explored too, i.e. whether an AI could support the achievement of the SDGs (after the AI has accepted the SDGs as its goals to strive for), which would be beneficial for a sustainable society.

In other fields AI programs have found creative solutions humans had not thought of before (e.g. Mnih et al., 2015, regarding video games). Also for the SDGs there are already some instances and it requires often only narrow AI, which focus on specific targets, rather than artificial general intelligence. Examples comprise autonomous robotic surgery for enhanced efficacy, safety and optimized surgical techniques, which addresses SDG 3 "Good Health and Well-Being" (Shademan et al., 2016), or a virtual teaching assistant, implemented on IBM's Watson platform, which addresses SDG 4 "Quality Education" (Maderer, 2016). However, for many of the 169 targets there are despite their urgency no AI attempts yet (Ziesche, 2017, for an overview). Therefore, the SDGs can also be considered as a priority agenda of research topics for increasingly progressing narrow AI. This briefly outlines the other direction of the synergy, which was not in the focus of this article: There is potential that AI assists in the achievement of the SDGs.

In summary, this article aims to bridge a gap between AI and the SDGs by proposing, in particular, a heuristic as a suitable interim attempt for the very hard AI value-loading problem and, in general, at least the initiation of common discussions. The heuristic suggests utilizing the entirety of the 17 SDGs of the UN 2030 Agenda for Sustainable Development as goal-set to be instilled to an AI. The benefit for AI development may be an interim step towards the achievement of AI safety, while the benefit for the UN 2030 Agenda for Sustainable Development may be innovative solutions towards the achievement of the SDGs.

References

Asilomar AI Principles (2017).
Retrieved from: https://futureoflife.org/ai-principles/
Bostrom, N. (2014a). Superintelligence: Paths, Dangers, Strategies. Oxford: Oxford University Press.
Bostrom, N. (2014b). Hail Mary, value porosity, and utility diversification. Technical report, Oxford University.
Retrieved from: https://nickbostrom.com/papers/porosity.pdf
Eden, A.H., Moor, J.H., Søraker, J.H., & Steinhart, E. (Eds.)(2013). Singularity Hypotheses: A Scientific and Philosophical Assessment. Heidelberg, New York, Dordrecht, London: Springer.
Franklin, S. (2014). History, motivations, and core themes. In Frankish, K., & Ramsey, W. M. (Eds.). The Cambridge handbook of artificial intelligence, 15-33. Cambridge: Cambridge University Press.
Inter-agency and Expert Group on SDG Indicators (2018, May 11). Tier Classification for Global SDG Indicators.
Retrieved from: https://unstats.un.org/sdgs/files/Tier%20Classification%20of%20SDG%20Indicators_11%20May%202018_web.pdf
International Council for Science and International Social Science Council (2015). Review of Targets for the Sustainable Development Goals: The Science Perspective. Paris: International Council for Science (ICSU).
Retrieved from: www.icsu.org/publications/reports-and-reviews/review-of-targets-for-the-sustainable-development-goals-the-science-perspective-2015/SDG-Report.pdf
Kurzweil, R. (2005). The Singularity Is Near. New York: Viking.
Legg, S., & Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. Minds and Machines, 17(4).
Retrieved from: https://arxiv.org/pdf/0712.3329.pdf
MacAskill, W. (2016, October 7). Moral Progress and Cause X.
Retrieved from: https://www.effectivealtruism.org/articles/moral-progress-and-cause-x/
Maderer, J. (2016). Artificial Intelligence Course Creates AI Teaching Assistant. Georgia Tech News Center, 9.
Retrieved from: http://www.news.gatech.edu/2016/05/09/artificial-intelligence-course-creates-ai-teaching-assistant
Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529-533.
Retrieved from:
https://www.nature.com/nature/journal/v518/n7540/full/nature14236.html
Omohundro, S. M. (2008). The basic AI drives. AGI (Vol. 171), 483-492.
Retrieved from: https://pdfs.semanticscholar.org/a658/2abc47397d96888108ea308c0168d94a230d.pdf
Shademan, A., Decker, R. S., Opfermann, J. D., Leonard, S., Krieger, A., & Kim, P. C. (2016). Supervised autonomous robotic soft tissue surgery. Science

translational medicine, 8(337), 337ra64-337ra64.
Soares, N. (2016). The Value Learning Problem. Proceedings of the Ethics for Artificial Intelligence Workshop at 25th International Joint Conference on Artificial Intelligence (IJCAI-2016).
Retrieved from: https://intelligence.org/files/ValueLearningProblem.pdf
Sustainable development goals - 17 Goals to Transform Our World (2018, July 12).
Retrieved from: https://www.un.org/sustainabledevelopment/
Sustainable development knowledge platform (2018, July 12).
Retrieved from: https://sustainabledevelopment.un.org/
Tegmark, M. (2017). Life 3.0: Being Human in the Age of Artificial Intelligence. New York: Knopf.
United Nations, General Assembly (2015). Transforming our world: the 2030 Agenda for Sustainable Development. Resolution A/RES/70/1.
Retrieved from: http://www.un.org/en/ga/search/view_doc.asp?symbol=A/RES/70/1
United Nations, General Assembly (2017). Work of the Statistical Commission pertaining to the 2030 Agenda for Sustainable Development. Resolution A/RES/71/313.
Retrieved from: http://www.un.org/en/ga/search/view_doc.asp?symbol=A/RES/71/313
Yampolskiy, R.V. (2015). Artificial Superintelligence: a Futuristic Approach. Chapman and Hall/CRC Press (Taylor & Francis Group).
Yudkowsky, E. (2008). Artificial Intelligence as a Positive and Negative Factor in Global Risk. In Bostrom, N., & Ćirković, M. (Eds.). Global Catastrophic Risks, 308–345. Oxford: Oxford University Press.
Retrieved from: https://intelligence.org/files/AIPosNegFactor.pdf
Yudkowsky, E. (2015, May 24). Complexity of value [Blog post].
Retrieved from: https://arbital.com/p/complexity_of_value/
Ziesche, S. (2017). Innovative Big Data Approaches for Capturing and Analyzing Data to Monitor and Achieve the SDGs. Report of the United Nations Economic and Social Commission for Asia and the Pacific: Subregional Office for East and North-East Asia (ESCAP-ENEA).
Retrieved from: http://www.unescap.org/sites/default/files/publications/Innovative%20Big%20Data%20Approaches%20for%20Capturing%20and%20Analyzing%20Data%20to%20Monitor%20and%20Achieve%20the%20SDGs.pdf