Artificial Intelligence for societal improvement: a commentary on targets and measures

Spyridon Samothrakis*, **Maria Fasli, Alejandro Quiroz Flores, Ana Matran-Fernandez, Haider Raza**

Institute for Analytics and Data Science, University of Essex, Colchester, UK

ssamot, mfasli, aquiro, amatra, h.raza@essex.ac.uk

Abstract

Artificial Intelligence methods focus on optimising specific targets (e.g. maximising reward, solving a logic puzzle, minimising a loss function). Once a problem can be translated into the language of optimisation, a whole panoply of methods can be used to tackle it. It is well known, however, that aggressively optimising for a target and using the same target as a measure quite often results in considerable bias and counter-intuitive results — this is known as Goodhart's or Campbell's law. Furthermore, unwanted side-effects (i.e., externalities) can act detrimentally to any solution resulting from a clear-cut optimisation process. We discuss the two problems of surrogate measures and (negative) externalities in more detail within the context of deploying artificial intelligence in societal improvement and propose a set of conceptual solutions.

1 Introduction

Measures of system performance rarely coincide with what a policy maker would like to actually optimise. This is mainly because measuring the actual qualities of a system is hard — direct measurements are often not available at all. A common example from Artificial Intelligence (AI) research is the often-used BLEU [Papineni *et al.*, 2002] metric for machine translation. This provides some heuristic rules to measure the quality of a translation, but it is just an approximation to human intuition — one can easily imagine very good BLEU translations that would seem awkward to a human.

In Economics, using proxy measures is extremely prevalent: Gross Domestic Product (GDP) or unemployment are not direct measures of societal happiness and wellbeing — rather, they are meant approximations of income or economic activity (e.g. [Coyle, 2015]), whose improvement should generally correspond with happier societies. If we are to deploy Artificial Intelligence to support and improve societal outcomes, we should at least be prepared to acknowledge that the measures we have inherited from social and economic sciences are not optimal and try to account for this.

*Contact Author

Using surrogate measures as targets for an optimisation process is problematic. For example, on the one hand, one can obviously force extremely low unemployment by establishing some kind of forced-to-work-for-nothing dystopia, but few would defend this as an optimal solution. This dubious relationship between targets and measures has been documented under various premises and is widely understood both in social sciences [Chrystal *et al.*, 2003] and economics [Sidorkin, 2016]. On the other hand, it is hard to see how the direct optimisation of surrogate measures can be avoided completely: these measures can not be avoided if we want to quantitatively assess progress, which is a vital concept in all modern societies [Michéa, 2009].

This paper is a commentary on surrogate measurements, ideal solutions and externalities and the strong domain-specific optimisation processes that underpin AI — how should we deploy AI systems?

The rest of this short paper is organised as follows: we discuss the notion of human progress in Section 2; we overview measures of progress in Section 3 and Externalities in Section 4. We propose two solutions on Section 5 - and we finally conclude in Section 6.

2 Progress and Growth

The idea of progress is historically relatively new and ties in with the inception and propagation of Liberalism [Michéa, 2009]. Generally speaking, the idea can be summarised as follows: life improvement is overall a good thing and human lives this year must be better than human lives the year before. This belief shaped almost all of 20^{th} century politics and is tightly coupled with the concept of (economic) growth. Whereas progress is a more abstract concept, growth is far more tangible and directly linked to the increase of GDP, after the effects of inflation have been taken into consideration. The fact that GDP is an imperfect surrogate measure and should therefore not be used as a form of societal welfare measure goes back to the person that introduced it [Kuznets, 1934]. More recently we have seen calls for the reversal of the (economic) growth trend as an effort to increase human happiness [Schneider *et al.*, 2010], which at very least brings some question to the universality of the idea. In terms of growth, we have an established surrogate measurement (GDP differentials), which is being dutifully followed and reported, but it is at least dubious whether it achieves its purpose.

3 Issues with Surrogate Measurements

The most obvious issue when dealing with surrogate measurements is the bias they introduce. Solutions that can be measured as acceptable can be horrible. For example, the tricks used to help chatbots pass Turing test-like competitions almost never correspond to anything deeper than cheats and do not approximate intelligence. The same applies to any societal measure as well, with the risk of even deeper issues: political campaigns turn these measures into societal targets, that can be widely cheated — creating widespread cynicism that makes any further efforts of improvement futile. Bureaucracies quite often revel in such measures, their primal reason of existence being the transformation of measures to targets.

4 Externalities

Apart from bias, a second problem that arises is the risk of (negative) externalities [Omohundro, 2008]. Any action taken upon a system (e.g., the creation of a factory) may have unwanted side-effect (e.g., pollution, longer working hours). The conclusion of Omohundro (2008) on how to deal with this is best portrayed on his often-quoted segment: "Social structures which cause individuals to bear the cost of their negative externalities would go a long way toward ensuring a stable and positive future". Solving for externalities might be imposing serious constraints on established social order.

5 Mixing Artificial Intelligence and Progress

In light of the above problems, what should be some rough guideline for the deployment of AI in order to improve societal outcomes? We propose some sample solutions based on the issues discussed above:

Improving surrogate quality The most obvious solution is improving the quality of the surrogate measures. Why use growth when it is well known that it is not properly mapped to societal welfare? And what should we measure instead? A substantial number of papers have been published on the topic (e.g. see [Lawn, 2003]), but the work is almost always ignored, as a well-known target is much easier to work on. We should, in fact, be both reporting and aiming on improving simultaneously on multiple fronts. Furthermore, one should examine which are the possible biases of current deployed solutions.

Improve on the ideals The concept of growth in itself is not neutral — it comes with specific philosophical baggage. A separate study with the ideals we work under should be commenced as soon as any work is to start in order for biases in the ideals to surface. Ideals are too often either confused with surrogate measurements or not thought out properly and quite often expressed as "Wishes". Given that it might be difficult to tackle worldwide problems, it might be easier to form ideals as short wishes where every other variable remains the same, bar the one we are interested in tackling.

Explicitly measuring externalities Externalities are a major source of issues - even if a deployed AI system measures and achieves what the designers actually have in mind, the secondary goals of measuring externalities and incorporating them within the optimisation process should allow for some fairness. Right now, externalities can be easily hidden and pushed on weaker elements of society — for example byproducts of uranium are mostly disposed away from the source of consumption.

Interpretability Even when all the above issues have been addressed, we are still faced with the situation where machines do not have direct access to what one would term "human common sense" [Davis and Marcus, 2015]. We are thus forced to have models and methods that we can provide to current experts to help us check the model sanity. Indeed, we have lately seen a serious amount of effort towards this direction (e.g., [Ribeiro *et al.*, 2016], [Selvaraju *et al.*, 2017]).

6 Discussions and Conclusion

We have provided a conceptual framework for the deployment of AI solutions to societal problems, focusing on not how to achieve certain results (we expect this to come from the AI process) but *what* are good results.

References

[Chrystal *et al.*, 2003] K Alec Chrystal, Paul D Mizen, and PD Mizen. Goodhart's law: its origins, meaning and implications for monetary policy. *Central banking, monetary theory and practice: Essays in honour of Charles Goodhart*, 1:221–243, 2003.

[Coyle, 2015] Diane Coyle. *GDP: A Brief but Affectionate History-Revised and expanded Edition*. Princeton University Press, 2015.

[Davis and Marcus, 2015] Ernest Davis and Gary Marcus. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 58(9):92–103, 2015.

[Kuznets, 1934] Simon Kuznets. National income, 1929-1932. In *National Income, 1929-1932*, pages 1–12. NBER, 1934.

[Lawn, 2003] Philip A Lawn. A theoretical foundation to support the index of sustainable economic welfare (isew), genuine progress indicator (gpi), and other related indexes. *Ecological Economics*, 44(1):105–118, 2003.

[Michéa, 2009] Jean-Claude Michéa. *Realm of Lesser Evil*. Polity, 2009.

[Omohundro, 2008] Stephen M Omohundro. The basic ai drives. In *AGI*, volume 171, pages 483–492, 2008.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM, 2016.

[Schneider *et al.*, 2010] François Schneider, Giorgos Kallis, and Joan Martinez-Alier. Crisis or opportunity? economic degrowth for social equity and ecological sustainability. introduction to this special issue. *Journal of cleaner production*, 18(6):511–518, 2010.

[Selvaraju *et al.*, 2017] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[Sidorkin, 2016] Alexander M Sidorkin. Campbell's law and the ethics of immensurability. *Studies in Philosophy and Education*, 35(4):321–332, 2016.