Scalar Reward is Not Enough

JAAMAS Track

Peter Vamplew
Federation University Australia
p.vamplew@federation.edu.au

Benjamin J. Smith
University of Oregon
benjsmith@gmail.com

Johan Källström
Linköping University
johan.kallstrom@liu.se

Gabriel Ramos
Universidade do Vale do Rio dos Sinos
gdoramos@unisinos.br

Roxana Rădulescu
Vrije Universiteit Brussel
roxana.radulescu@vub.be

Diederik M. Roijers
Vrije Universiteit Brussel
diederik.roijers@vub.be

Conor F. Hayes
University of Galway
c.hayes13@universityofgalway.ie

Fredrik Heintz
Linköping University
fredrik.heintz@liu.se

Patrick Mannion
University of Galway
patrickmannion@universityofgalway.ie

Pieter J.K. Libin
Vrije Universiteit Brussel
pieter.libin@vub.be

Richard Dazeley
Deakin University
richard.dazeley@deakin.edu.au

Cameron Foale
Federation University Australia
c.foale@federation.edu.au

ABSTRACT

Silver et al. [14] posit that scalar reward maximisation is sufficient to underpin all intelligence and provides a suitable basis for artificial general intelligence (AGI). This extended abstract summarises the counter-argument from our JAAMAS paper[19].

KEYWORDS

Scalar rewards; Vector rewards; AGI; Reinforcement learning

ACM Reference Format:
Peter Vamplew, Benjamin J. Smith, Johan Källström, Gabriel Ramos, Roxana Rădulescu, Diederik M. Roijers, Conor F. Hayes, Fredrik Heintz, Patrick Mannion, Pieter J.K. Libin, Richard Dazeley, and Cameron Foale. 2023. Scalar Reward is Not Enough: JAAMAS Track. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

Silver et al. [14] present the *reward-is-enough* hypothesis that "Intelligence, and its associated abilities, can be understood as subserving the maximisation of reward by an agent acting in its environment", and argue for reward maximisation as a means for creating AGI. We assert that the ability to consider multiple conflicting objectives is a critical aspect of intelligence, and is inadequately addressed by maximising a scalar reward. Even if scalar rewards are sufficient to create AGI, this approach greatly increases the likelihood of adverse outcomes. Therefore, we advocate explicitly multi-objective AI methods based on vector rewards.

2 THE LIMITATIONS OF SCALAR REWARDS

The relative merits of scalar and vector rewards have been extensively studied [8, 12, 13]. For many tasks an intelligent decision-maker must trade-off between multiple conflicting objectives. For example a biological agent must satisfy drives such as reproduction, hunger, thirst, avoidance of pain, following social norms, and so on. An agent based on scalar rewards must either be maximising only one of these objectives, or some scalarised combination of them.

Silver et al. acknowledge that multiple objectives exist, but argue "a scalar reward signal can represent weighted combinations of objectives". However it is well known that this places limitations on the solutions which can be found [5, 20], and so may not allow an agent to maximise its true utility [13]. In contrast, intelligence based on vector rewards and approaches that are explicitly multi-objective can directly optimise any desired measure of utility [8].

Vector rewards also support adaptation to changes in utility. A scalar reward encodes a single, fixed weighting of objectives, while vector rewards allow an agent to pursue its current goal, while simultaneously learning with regard to other possible future goals. Silver et al. state that "Intelligence may be understood as a flexible ability to achieve goals", but scalar rewards do not allow the degree of flexibility supported by multi-policy multi-objective methods.

Silver et al. also state "a solution to a specialised problem does not usually generalise; in contrast a solution to the general problem will also provide a solution for any special cases". We disagree with the implied assumption that maximising scalar reward is the general case. Scalar rewards (where the number of rewards $n = 1$) are a subset of vector rewards (where the number of rewards $n \geq 1$). Agents developed for vector rewards are also applicable to scalar rewards, as the scalar can be treated as a one-dimensional vector. The inverse is not true – mapping a vector reward to a scalar inevitably limits some capabilities of the agent. Therefore methods for scalar rewards are in fact the special case.

3 MULTI-OBJECTIVE REINFORCEMENT LEARNING IN NATURAL INTELLIGENCES

If our arguments in favour of multi-objective representations of reward are correct, then it would be expected that naturally evolved intelligences such as those in humans and animals would exhibit

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

evidence of vector-valued rewards. In fact, evolution has developed organisms that delegate learning not just into multiple objectives but even into multiple learning systems that are embedded within an organism. There are multiple objectives at a basic biological regulatory level, and these are matched with multiple objectives at every level of analysis of the organism.

4 INTERNALLY-DERIVED REWARDS

One could argue that an agent maximising a scalar reward may still develop the capacity to carry out multi-objective decision-making. For example, agents based on evolutionary algorithms or reinforcement learning might construct their own internal reward signals to guide their learning and decision-making [7, 15, 17].

Regardless of whether vector rewards are derived externally or internally, the agent must make decisions based on those vector values. Silver et al. argue that an agent maximising a scalar reward could theoretically develop multi-objective capabilities. However we believe it is more practical to construct multi-objective agents via explicitly multi-objective algorithms. Similarly, we argue that it makes sense to design multi-objective reward structures for computational agents rather than relying on them to identify such structures themselves. In fact, we contend that it typically will be easier to specify multi-objective rewards directly than to design a scalar reward which captures all of the various factors of interest.

5 REWARD MAXIMISATION AND AGI

One of the main arguments of Silver et al. is that maximising of a simple scalar reward in the context of a suitably complex environment may suffice for the emergence of general intelligence. They illustrate this via the the scenario of an agent given a reward of +1 for collecting a round pebble, arguing this could lead it to develop tools, form an understanding of the natural processes which form pebbles, persuade people to collect pebbles, and so on.

While the development of open-ended, far-reaching intelligence from such a simple reward is presented positively by Silver et al., this scenario is strikingly similar to the infamous *paperclip maximiser* thought experiment from the AI safety literature [4]. While unrestricted maximisation of a scalar reward may indeed result in the development of complex, intelligent behaviour, it is also inherently dangerous [11]. For this reason, AI safety researchers have argued in favour of approaches based on satisficing rather than unbounded maximisation [16], or on multi-objective measures of utility which account for factors such as safety or ethics [18].

Therefore we argue that even if scalar rewards are enough for the development of general intelligence, they are not sufficient for the far more important task of creating human-aligned AGI. While safety and ethics are not the focus of Silver et al.'s paper, it is concerning that these issues are not acknowledged in a paper which is actively calling for the development of AGI.

Reward specification is difficult even in trivial systems, and reward misspecification or reward hacking often lead to surprising, unintended, and undesirable behaviour [3]. In more complex systems with more general agents, the potential for reward misspecification significantly increases [6]. We argue that the use of scalar rewards leads to significant risks of unpredictable and undesirable behaviour. Given the limitations of their human designers, scalar rewards will most likely not be enough for the development of AGI with guaranteed behavioural properties, and predictable reward design is better achieved using multi-objective methods.

One possible implementation of a multi-objective approach to safe and ethical AGI would be a review-and-adjust cycle [8]. A multi-objective AGI plans or learns a set of optimal policies for all possible utility functions. A policy is then selected to be executed, possibly with direct or indirect user feedback. The outcome can then be reviewed by an overseer (either a human, the AGI itself, or another AGI), along with the AGI's explanation of its policy selection. The MOMDP, utility function or set of solutions can then be updated based on this review. We note that such reviews can not only be triggered by incidents, but also by regular inspection.

We see such a cycle as essential for future AI systems. As AI researchers we have to enable responsible deployment. It is our opinion that the above-mentioned benefits are not merely desirable, but that it is a moral imperative for AI developers to obtain them, in order to create systems that more likely benefit society.

6 CONCLUSION

Silver et al. argue that maximisation of a scalar reward suffices to explain all observed properties of natural intelligence, and to support the construction of artificial general intelligence. However, this requires representing all of the different objectives of an intelligence as a single scalar value. which places restrictions on the behaviour which can emerge. Therefore, we contend that the *reward-is-enough* hypothesis does not provide a sufficient basis for understanding all aspects of naturally occurring intelligence, nor for the creation of computational agents with broad capabilities.

In the context of AGI, a focus on maximising scalar rewards creates an unacceptable exposure to risks of unsafe or unethical behaviour by the AGI agents. This is particularly concerning given that Silver et al. are highly influential researchers and employed at DeepMind, one of the organisations best equipped to expand the frontiers of AGI. While Silver et al. "hope that other researchers will join us on our quest", we instead hope that the creation of AGI based on reward maximisation is tempered by other researchers with an understanding of the issues of AI safety [9, 10] and an appreciation of the benefits of multi-objective agents [1, 2].

7 ACKNOWLEDGEMENTS

This research was supported by funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" program, and by the National Cancer Institute of the U.S. National Institutes of Health under Award Number 1R01CA240452-01A1. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or other funders. Pieter J.K. Libin and Roxana Rădulescu acknowledge support from the Research Foundation Flanders (FWO, fwo.be) (postdoctoral fellowships 1242021N and 1286223N). Johan Källström and Fredrik Heintz were partially supported by the Swedish Governmental Agency for Innovation Systems (grant NFFP7/2017-04885), and the Wallenberg Artificial Intelligence, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. Conor F. Hayes was funded by the University of Galway Hardiman Scholarship. Gabriel Ramos was partially supported by FAPERGS (grant 19/2551-0001277-2) and FAPESP (grant 2020/05165-1).

REFERENCES

[1] Abdolmaleki, A., Huang, S., Hasenclever, L., Neunert, M., Song, F., Zambelli, M., Martins, M., Heess, N., Hadsell, R., Riedmiller, M.: A distributional view on multi-objective policy optimization. In: International Conference on Machine Learning, pp. 11–22. PMLR (2020)

[2] Abdolmaleki, A., Huang, S.H., Vezzani, G., Shahriari, B., Springenberg, J.T., Mishra, S., TB, D., Byravan, A., Bousmalis, K., Gyorgy, A., et al.: On multi-objective policy optimization as a tool for reinforcement learning. arXiv preprint arXiv:2106.08199 (2021)

[3] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D.: Concrete problems in ai safety. arXiv preprint arXiv:1606.06565 (2016). URL https://arxiv.org/pdf/1606.06565.pdf

[4] Bostrom, N.: Ethical issues in advanced artificial intelligence. Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence pp. 12–17 (2003)

[5] Das, I., Dennis, J.E.: A closer look at drawbacks of minimizing weighted sums of objectives for pareto set generation in multicriteria optimization problems. Structural optimization **14**(1), 63–69 (1997)

[6] Dewey, D.: Reinforcement learning and the reward engineering principle. In: 2014 AAAI Spring Symposium Series (2014)

[7] Elfwing, S., Uchibe, E., Doya, K., Christensen, H.I.: Co-evolution of shaping rewards and meta-parameters in reinforcement learning. Adaptive Behavior **16**(6), 400–412 (2008)

[8] Hayes, C.F., Rădulescu, R., Bargiacchi, E., Källström, J., Macfarlane, M., Reymond, M., Verstraeten, T., Zintgraf, L.M., Dazeley, R., Heintz, F., Howley, E., Irissappane, A.A., Mannion, P., Nowé, A., Ramos, G., Restelli, M., Vamplew, P., Roijers, D.M.: A practical guide to multi-objective reinforcement learning and planning. Autonomous Agents and Multi-Agent Systems **36** (2022)

[9] Krakovna, V., Orseau, L., Ngo, R., Martic, M., Legg, S.: Avoiding side effects by considering future tasks. arXiv preprint arXiv:2010.07877 (2020)

[10] Leike, J., Martic, M., Krakovna, V., Ortega, P.A., Everitt, T., Lefrancq, A., Orseau, L., Legg, S.: AI safety gridworlds. arXiv preprint arXiv:1711.09883 (2017)

[11] Omohundro, S.M.: The basic AI drives. In: AGI, vol. 171, pp. 483–492 (2008)

[12] Rădulescu, R., Mannion, P., Roijers, D.M., Nowé, A.: Multi-objective multi-agent decision making: a utility-based analysis and survey. Autonomous Agents and Multi-Agent Systems **34**(1), 1–52 (2020)

[13] Roijers, D.M., Vamplew, P., Whiteson, S., Dazeley, R.: A survey of multi-objective sequential decision-making. Journal of Artificial Intelligence Research **48**, 67–113 (2013)

[14] Silver, D., Singh, S., Precup, D., Sutton, R.S.: Reward is enough. Artificial Intelligence p. 103535 (2021)

[15] Singh, S., Lewis, R.L., Barto, A.G., Sorg, J.: Intrinsically motivated reinforcement learning: An evolutionary perspective. IEEE Transactions on Autonomous Mental Development **2**(2), 70–82 (2010)

[16] Taylor, J.: Quantilizers: A safer alternative to maximizers for limited optimization. In: AAAI Workshop: AI, Ethics, and Society (2016)

[17] Uchibe, E., Doya, K.: Finding intrinsic rewards by embodied evolution and constrained reinforcement learning. Neural Networks **21**(10), 1447–1455 (2008)

[18] Vamplew, P., Dazeley, R., Foale, C., Firmin, S., Mummery, J.: Human-aligned artificial intelligence is a multiobjective problem. Ethics and Information Technology **20**(1), 27–40 (2018)

[19] Vamplew, P., Smith, B.J., Källström, J., Ramos, G., Rădulescu, R., Roijers, D.M., Hayes, C.F., Heintz, F., Mannion, P., Libin, P.J., et al.: Scalar reward is not enough: A response to silver, singh, precup and sutton (2021). Autonomous Agents and Multi-Agent Systems **36**(2), 1–19 (2022)

[20] Vamplew, P., Yearwood, J., Dazeley, R., Berry, A.: On the limitations of scalarisation for multi-objective reinforcement learning of pareto fronts. In: Australasian joint conference on artificial intelligence, pp. 372–378. Springer (2008)