Reward (Mis)design for Autonomous Driving (Abstract Reprint)

W. Bradley Knox[1,2], **Alessandro Allievi**[1,2], **Holger Banzhaf**[3], **Felix Schmitt**[4], **Peter Stone**[2,5]

[1]Robert Bosch LLC , United States of America
[2]The University of Texas at Austin, United States of America
[3]Robert Bosch GmbH, Germany
[4]Bosch Center for Artificial Intelligence, Germany
[5]Sony AI, United States of America

Abstract Reprint. This is an abstract reprint of a journal article by Knox, Allievi, Banzhaf, Schmitt, and Stone (2023).

Abstract

This article considers the problem of diagnosing certain common errors in reward design. Its insights are also applicable to the design of cost functions and performance metrics more generally. To diagnose common errors, we develop 8 simple sanity checks for identifying flaws in reward functions. We survey research that is published in top-tier venues and focuses on reinforcement learning (RL) for autonomous driving (AD). Specifically, we closely examine the reported reward function in each publication and present these reward functions in a complete and standardized format in the appendix. Wherever we have sufficient information, we apply the 8 sanity checks to each surveyed reward function, revealing near-universal flaws in reward design for AD that might also exist pervasively across reward design for other tasks. Lastly, we explore promising directions that may aid the design of reward functions for AD in subsequent research, following a process of inquiry that can be adapted to other domains.

References

Knox, W. B.; Allievi, A.; Banzhaf, H.; Schmitt, F.; and Stone, P. 2023. Reward (Mis)design for autonomous driving. *Artificial Intelligence*, 316: 103829.