arXiv:2301.01320v1 [cs.LG] 3 Jan 2023

迈向可部署强化学习——强化学习研究的问题与一种可能的解决方案

Shie Mannor*和 Aviv Tamar†

2023年1月1日

摘要

强化学习（RL）展现出了巨大的潜力，但目前充斥着过度炒作和不切实际的幻想。我们指出了当前研究面临的一些困难，并认为这些困难根植于社区所选择的发展方向。我们认为，当前的发展方向不太可能实现“可部署”的强化学习：它在实践中行之有效，能够应对实际情境，同时在经济上仍然可行。我们还针对该领域面临的一些困难提出了一种可能的解决方案。

1 引言

自 DeepMind 于 2015 年取得将深度强化学习应用于 Atari 游戏的突破性成果以来，强化学习一直备受人工智能社区赞誉，人们对它寄予厚望，从“通往通用人工智能之路”，到“自动驾驶汽车的关键”，再到“解决所有规划问题”。尽管我们确实取得了显著且值得肯定的进展，但目前的普遍看法是，除了玩游戏之外，强化学习并未达到我们的预期。我们认为，原因在于五种流行的研究实践。这些做法在 2015 年尚有意义，但如今却正使这一领域陷入停滞。我们的宗旨并非指责任何人（我们自己肯定也犯过不少错），因此不举例，也不列出参考文献。我们想指出的是，目前围绕强化学习实践与理论的热潮，在很大程度上缺乏事实依据，而这一学科需要走向成熟。

1. 对特定基准测试过拟合。如今，Atari 游戏领域的最先进方法与 MuJoCo（或其他基准测试）领域的最先进方法已大不相同。几乎每篇论文都必须在某项热门基准测试上展现出某种（按某项指标衡量的）优势。然而，这些基准测试都是人为构造的，使该领域进一步偏离那些可能适用于实际问题的算法，而实际问题并非 Atari 游戏或机器人运动。因此，针对特定基准测试、很可能只能在其上奏效的改进研究比比皆是。此外，基准测试上的进展与现实世界的价值有何关联，并不明确。与预训练的 ImageNet 特征或 BERT 句子嵌入不同，后两者广泛适用于现实世界的应用，而目前解决强化学习基准测试并不能带来切实价值。

2. 关注重点有误。大多数研究都聚焦于针对给定基准测试的样本复杂度，但实践中很少如此——与工程投入相比，计算资源可能成本低廉；获取带标签的示范数据可以显著加快学习；而开发过程必须考虑整体系统稳定性、可测试性、调试便捷性、可解释性，以及与其他组件/智能体的集成等问题。当前的基准测试完全忽视了情境化（由强化学习驱动）智能体的可部署性，关注的是算法，而非系统/工程视角。这一点在广泛使用的 OpenAI Gym API 中体现得最为明显：它将所有“系统设计”问题抽象掉，以便快速推进样本复杂度方面的研究。尽管这一做法起初是个好主意，却阻碍了进展，因为在许多实际问题中，弄清楚哪些状态、动作和奖励有用，本身就是开发过程中的关键环节。

3. 脱节的理论。理论在现代强化学习研究中应扮演什么角色？现实中并不存在状态空间和动作空间都很小且有限的系统。理想情况下，理论应帮助我们理解

*Technion 和英伟达研究院；本文仅代表我个人观点。
†Technion；本文仅代表我个人观点。

实践中观察到的现象，并提出算法思路。尽管“没有什么比好的理论更实用”（Lewin，1952），但有用的理论似乎相当罕见。原因包括：遗憾最小化过于悲观；算法设计、参数选择等方面存在大量先验知识，而理论并未将其纳入考量；有限状态与动作并不能很好地描述许多值得关注的问题；以及关注无人关心的次要指标。

4. 不平等的竞技环境。在基准测试上衡量算法的性能，会受到实施者可用资源的影响，例如超参数调优的熟练程度、训练的神经网络规模，或对问题／解决方案的先验知识。不同研究者（例如学术界与工业界）能够承担的实验规模和软件工程投入各不相同，而顶级会议目前又倾向于偏好大规模实验而非概念创新，这些因素都可能阻碍长期进展。

5. 缺乏实验严谨性。令人瞩目的单次实验有时会造成进展显著的错觉。尽管这类实验很适合作为公关噱头，却可能最终令人失望，因为它们会妨碍人们继续研究那些被认为“已经解决”、实际上却远未解决的问题。例如，简单的二维 ProcGen 迷宫基准测试至今仍未解决。我们需要更严谨地评估问题的难度和解决成效。此外，工业界要采纳一种方法，光有令人瞩目的结果还不够——稳定性、开发时间与成本、可测试性以及生命周期问题都至关重要。目前的发表标准是，失败案例几乎从不报告，稳定性无从判断，软件设计问题甚至也无人讨论。

2 通才智能体与可部署强化学习系统

在如何推动解决现实世界决策问题方面，强化学习社区存在两种主要的教条。第一种观点，我们在此称之为“通才智能体”观点（有时也称为“强化学习优先”），认为未来的进展将来自于专注于对能够解决多种问题的智能体进行大规模训练，并希望在这一过程中培育出通才智能体，使其成为解决各种现实世界问题的有用组成部分。第二种观点，我们称之为“可部署强化学习”，采取更务实的立场，致力于设计能够解决具体现实世界问题的强化学习算法（有时也称为“强化学习其次”）。上述五个问题与这两种方法都有关。不过，在我们提出的修正方案中，我们专注于第二种观点。原因在于，基于该领域目前的知识，我们认为可部署强化学习仍有切实的收益可供获取，而通才智能体的前景仍处于早期发展阶段。

可部署强化学习研究的三项原则

重要的是要认识到，目前部署强化学习并不经济。要改变这一点，既需要研究如何有效部署强化学习，也需要更好地了解强化学习解决方案可能带来的收益，最终使其值得投入。我们提出了一种强化学习研究的建设性模型，认为它符合该领域当前的认知状况，并将推动基于强化学习的解决方案在现实世界中的部署。下文将概述三项普遍原则。我们希望研究人员能在项目中采纳其中一些原则，从而推动发表标准逐步认可这些原则的价值。

1. 挑战而非基准测试。挑战是由学术界或工业界的一组研究人员发起的问题（无论是否真实存在）。与基准测试不同，挑战并非旨在比较算法，而是单纯为了将问题解决。也就是说，无论采用强化学习（或其他）算法来解决，推动挑战取得进展都具有现实世界价值。

贡献挑战的功劳 对挑战进行严谨阐述，应被视为一项重要贡献（体现在引用、影响力等方面）。这项贡献不仅是对某个应用的具体描述，还包括围绕它形成的社区，以及支持平台（代码、排行榜等）。我们设想一种特殊类型的论文：“贡献型挑战”。评估挑战的一项基本标准，是衡量在该挑战上取得

真实且广受认可的进展。也就是说，不再最大化虚构的奖励，而是致力于推动现实世界进步的指标。

可衡量的进展是发表的主要标准。无论是提出新算法，还是给出正面或负面结果，凡是在某项挑战上取得进展的发表成果，都应说明所提新算法的局限性和问题，以及它具体如何推动进展。发表一项工作的主要标准，是其以可量化的方式取得的进展。

算力等级。开发过程中可用的计算资源将贡献者区分为不同的“算力等级”，对此也应予以报告。研究所使用的计算资源总量对于评估结果的重要性至关重要。

2. 理论论文应聚焦于具体挑战。当理论论文提出模型、算法或解决方案时，问题应清晰明确，并植根于现实挑战。我们并不反对研究小型有限模型的理论，但研究目标应根据其对现实世界问题的潜在影响力得到充分论证。理论论文还应考虑与系统生命周期有关的问题，重点关注数据获取、调试、可测试性和性能退化等方面。

3. 面向设计模式的研究。软件设计基于整合多种设计模式、对其进行调整，并重新打包成软件系统。在软件工程中，设计模式是针对特定问题提出的一种解决方案框架，可在许多实例中重复使用。类似地，基于强化学习的现实世界系统也应针对各种问题提出概念性解决方案，以应对可测试性、可调试性及其他系统生命周期问题。能够解决植根于应用中的问题的生命周期管理方法，将为强化学习研究增添有益内容。重要的是，我们预见，针对某项挑战开发新颖的设计模式，是取得重大进展的一种途径。

4 可执行的后续步骤

社区越早开始关注可部署强化学习，就越有可能看到大规模的现实世界影响力。对我们而言，可部署强化学习意味着强化学习能够大规模运行、在经济上可行，并最终能够投入实际应用。如果你认同上述观点，可以采取以下行动。

1. 提出挑战。提出挑战需要深入了解某个应用领域。一些工业研究实验室已经具备良好的条件，可以开始着手这项工作。强化学习可以在许多科学与工程学科中发挥影响力，例如自然科学、医学和制造业。挑战应当对其所属学科具有重要意义；因此，强化学习专家即使不具备相关领域知识，也应在提出挑战时向领域专家说明哪些事情可能可行、哪些不可行。与工业界或学术界合作，能够提出界定更清晰、也更切合实际的挑战。

2. 重新定位自己的研究。在撰写论文和开展研究时，可能需要调整关注重点，致力于解决真正重要的实际问题，并依据可部署强化学习的原则来定位研究工作。

3. 批评他人的研究。要推动强化学习研究社区发生改变，可能需要研究人员、审稿人和高级领域主席协同努力。如果你正在审阅强化学习论文，不妨跳出常见的基准测试驱动的评估方式，思考一篇论文如何让该领域更接近现实世界中的影响力。如果你是高级审稿人或领域主席，也可以考虑指导审稿人以不同的标准评判论文。

这份手稿的在线版本（https://avivtamar.substack.com/p/deployablerl）包含一个评论区。作为起点，我们邀请研究人员在那里提出并讨论相关挑战的想法。

参考文献

[1] Lewin, K. 1952. 社会科学中的场论。纽约：Harper & Row。[2] Mnih, V., 等。2015. 通过深度强化学习实现人类水平的控制。《自然》518(7540), 529-533