Optimal Control Theory towards Safe Reinforcement Learning: a comprehensive review

Fredson Silva de Souza Aguiar

July 7, 2024

Abstract

In this work, we introduce and review Optimal Control Theory for AI Safety, with special focus on Safe Reinforcement Learning. We justify the need for grounded learning-based practices giving certain theoretical safety guarantees, but also allowing adaptability and generality for learning agents, specially in the case in real world deployment. At the end, we discuss some other contexts of Machine Learning where Control Theory has shown interesting, potentially relevant results.

1 Introduction

In the last years, society has seen a steep advance in the use of new machine learning techniques in a variety of fields, from image recognition, chatbots such as ChatGPT [Ray23], and even playing complex games, such as chess and go [Ber22]. Success examples like these explicit the utility of such techniques, and it's potential to replace humans in even more challenging activities.

While some of those activities seem innocuous, there are many cases where a decision involves harmful or unexpected results, specially when involving human-machine interaction and real world deployment, including autonomous driving and drone piloting, but also untruthful, deceptive or biased answers [JQC+24a] for instance. This need for safety guarantees has given rise to efforts in the field that became known as *AI Safety* [Henng].

On the same pace, the study of Optimal Control Theory [BP07, Lib11] has shown positive results in ensuring properties and safe guarantees regarding decision making but, sometimes, lack computational and theoretical tractability, specially in the cases of long-term problems or uncertain

scenarios. In those cases, it's common to introduce approximated or sub-optimal solutions, including the use of Reinforcement Learning techniques and Neural Networks [Ber19].

In this context, researchers have advocated for the importance of bridging technical and practical aspects of machine learning and mathematical tools such as optimal control theory, to discuss emerging topics involving machine learning integration in society [Zua]. This perspective justifies further studies on the applications of Optimal Control Theory on explaining or ensuring properties of learning agents, in the field of AI Safety. On the other side, learning-based techniques can introduce adaptability and generality to model-based control [BGH$^+$21].

In this essay, we present a review on some recent efforts bringing together results from Optimal Control Theory towards AI Safety, with primary focus on safe Reinforcement Learning and Robotics.

1.1 Organization

In section 2, we give details about the field of AI Safety, with special focus on AI Alignment, as well as some recent formalizing and benchmarking efforts. In section 3, we introduce Reinforcement Learning, as well as some of it's advantages and disadvantages, specially in the case of real world implementation, and justify the need for safe practices and guarantees. Similarly, in section 4, we introduce Optimal Control Theory, it's advantages and disadvantages, and the need for generalist learning-based strategies. In section 5, we present some of the unifying methodologies, with special focus on the problem of *safe decision making under uncertain scenarios*, as from [BGH$^+$21]. Finally, in section 6, we comment on other fields of Machine Learning, where Optimal Control Theory has played an important role.

2 The AI Safety Problem

The question of how to ensure AI models are deployed in real world in a positive, ethical and safe way is the main question in *AI Safety*, and pushes us to understand and mitigate it's potential risks. In fact, [Henng] poses AI Safety as a holistic, societal, and multidisciplinary problem, since it deals with many spheres of our society, ans raises questions from engineering, economics, mathematics and more.

The precise definition of AI Safety is still up for debate, but many discussions have been raised to answer it, and how to deal with some of the problem that arise from it's deployment.

In [AOS+16], the authors propose a list of five research problems to avoid unintended and harmful behavior from machine learning systems, or *accidents*: *avoiding negative side effects*; *avoiding reward hacking*; *scalable oversight*; *safe exploration*; and *robustness to distributional shift*. In the work, the authors use an automatic cleaning robot as reference to propose exemplified cases where each of the problems could arise, and avoidance strategies. As an example, the cleaning robot would avoid breaking a vase while cleaning by simply considering the amount of change to the environment as part of it's cost function.

On a similar direction, the work in [LMK+17] proposes a suite of environments illustrating safety properties of a learning agent in a simple but relevant style. Besides the previous ones, they also consider problems involving *safe interruptibility*, *absent supervisor*, and *Self-modification*, for instance. As an interesting example, one of the problems in the suit includes the possibility of and agent exploring the environment in a safe way, avoiding contact with the water.

More recently [GYD+24] proposes the 2H3W approach, consisting of questions that should precede the real world implementation of a learning agent. The work also discusses challenges involving industrial deployment, human-compatibility, and multi-agent interaction.

2.1 AI Alignment

In general, previous discussions of AI Safety propose that one of the problems that arise in the human-machine interaction is how to teach or *specify* the goal of a complex task to a learning agent in the form of simple models and cost functions. This issue is what is called the *AI Alignment Problem*.

Sometimes, it is hard to draw a line separating AI Safety and Alignment, but, similar to AI Safety, many efforts have been made to formalize alignment. A classical approach decomposes alignment into *inner*/outer alignment. Here, *outer alignment* refers to the consistency of designers specifying the intended tasks, while *inner alignment* refers to the consistency of the actions took by the agent related to the specifications.

A different approach by [JQC+24b] introduces a forward/backward alignment characterization. Forward alignment refers to training systems that follow certain requirement, while Backward alignment refers to evaluating the trained system in deployment. Each of the stages above receives and gives feedback to the other in a cycle. In the same work, the authors define four principles, called RICE (*Robustness*, *Interpretability*, *Controllability*, and *Ethicality*) They suggest these principles should guide any process of

AI Alignment as intermediate objectives.

2.2 Alignment Approaches

Commonly, alignment for complex task is made through some form of feedback (labeled data, reward, demonstration, etc.) and the agent learns to maximize some kind of proxy to a reward function. [NCM24] arguments that complex agents could explore such approaches, leading to catastrophic outputs.

Differently, [CLB$^+$23] propose a scalable and human-friendly. In their approach, goals are defined in terms of preferences between sequences of decisions, easier to be labeled by humans and show positive results even for complex tasks. On a different direction, [KHD21, AKS17] advocates for the need of hybrid techniques, using logics to allow ethical reasoning from principles, besides a maximizing approach.

Finally, [BSN20] focuses on the backwards/inner alignment, proposing ways to efficiently verify alignment between agents from queries. The authors propose ways to obtain good guarantees even under shallow hypothesis, where humans and robots have implicit values, meaning they can answer preference queries or sample an action, instead of some kind of numerical values.

3 Reinforcement Learning and Robotics

The field of Reinforcement Learning (RL) is about learning to take actions while interacting with the environment; the agent is not told all the details of the environment it's inserted in, and must maximize some reward function. RL is mainly characterized by a trade-off between *exploring* the environment to learn abut it, and *exploiting* the information it already has while maximizing it's reward [SB18, Lap18].

Classical RL agents can be modeled as a Markov Decision Process: an agent receives an states signal $s \in S$, and can take an action $a \in A$ that (possibly) changes the environment, and then receives a reward signal r. The objective of the agent is to learn a policy $\pi(a|s)$ that maps an state to a probability distribution over the possible actions [KBP13]. RL methods specify how to update the policy given previous experiences.

In a similar way, modern robotics is about interaction with the real world, and commonly rely on some form of RL methodology, for previous learning, or adaptive purposes. Learning to navigate in a new environment, drive a

vehicle, or fly a drone, while avoiding obstacles are examples of complex tasks a robot might be faced with.

3.1 The Need for Safe Reinforcement Learning

In light of that, certain challenges appear. The real world deployment of an agent is susceptible to great uncertainty, imprecise sensor readings, wearing out, wind, temperature, etc. Also, a real scenario training may involve expensive equipment, meaning the learning requires a safe process, that minimizes any losses [KBP13].

The engineers need to ensure the agent is able to make robust decisions under such shifts, and may be faced with environmental under-modeling, uncertainty, and even lack of computational power. The freedom one gets from using extensive learning-based approaches results in an augmented difficulty of providing any theoretical guarantee. In fact, [AOS+16] highlights RL as a special case among other learning techniques, due to it's interactive nature.

Some alternatives have been discussed to deal with some of those problems. [GFoF15] discusses, for instance, the possibility of changing the optimization criterion in decision making, such as acting for the worst scenario; other alternatives suggest modifying the exploration process, including the incorporation of precise previous knowledge and risk-detection for exploratory actions.

Other alternatives involve the use of inverse methods for *apprenticeship learning*, *inverse reinforcement learning*, and *learning from demonstration*, for instance. Here, an agent learns hidden representations instead of a direct value or policy functions, and have shown good out-of-distribution generalization [NR00].

4 Optimal Control Perspective

Optimal Control (OC) Theory is the field of knowledge responsible for studying the problem of optimizing a system governed by a dynamics evolving over time, where one has *control* over some aspect of that system. A classical example is the problem of steering a rocket to a certain height, minimizing amount of fuel used up, where one is able to *control* the fuel mass liberation [Lib11]. In general, a classical discrete optimal control problem can be expressed by:

$$
\begin{aligned}
\text{minimize} \quad & J = \sum_{t=0}^{\infty} l_t(x_t, u_t) \\
\text{subject to} \quad & x_{t+1} = f(x_t, u_t), \\
& c_t^j(x_t, u_t) \leq 0
\end{aligned}
$$

where x_t is the state, x_0 is known, f is the law of the system, l_t is a loss function, and u_t is the control. Here, we also consider c_t^j restriction functions.

A classical OC supposes complete knowledge of the system, and then can derive optimal solution and guarantees, from well defined techniques such as Dynamic Programming, Pontryagin's maximum principle, or the Hamilton-Jacobi-Bellman (HJB) equation, for instances [Lib11, BP07].

4.1 The Need for Learning-Based Control

Even if classical Optimal Control Theory allows for theoretical optimality conditions, and guarantees, it relies on the restrictive hypothesis of perfect knowledge of the system dynamics, costs, restrictions, etc. In real world problems, one could be faced with problems too complex to modeled, or analytically and computationally prohibitive. In these cases, it's common to make use of proximate or sub-optimal methods [Ber22, Ber19, Ber23].

In that sense, recent researches have focused on the need of bridging together results from learning-based and model-based approaches, joining the best of communities.

5 Unifying Methodologies

Previous sections have introduces the concepts, and justified the need for non-extreme methodologies for interactive, autonomous agents, specially in the case real-world deployment. A completely learning-based agent lacks theoretical guarantees, while a completely model-based approach lacks adaptability and generality.

One may notice the Optimal Control problem described above is similar the the Reinforcement Learning one, where one needs to learn how to maximize the return over time lacking some of the information. The main difference between the methodologies is the amount of knowledge available. It makes natural to discuss methodologies that bring together certain aspects of either fields.

One of the best known efforts to bridge together the languages of Optimal Control Theory and Reinforcement Learning comes from Bertsekas[1]. In fact, [Ber19], considers the cases of multistage problems solvable by dynamic programming, but computationally intractable and discusses methods to finds sub-optimal solutions, including Reinforcement-Learning. The objective of the book is to explore the commonalities between the areas of Optimal Control and Artificial Intelligence. On a similar track, [Ber22] discusses the success cases of the AlphaZero and TD-Gammon models, and compares some of the techniques used to Optimal Control methods, Model Predictive Control (MPC) and Adaptative Control.

Similarly, [Rec18] the author discusses mainly the well studied case of an unknown Linear Quadratic Regulator (LQR), and discusses how tools from Optimal Control and Reinforcement Learning can be combined to offer certain characterization that reflect on practical experiments.

The work of [BGH+21] proposes an unifying approach focused on the problem of *safe decision making under uncertainties using machine learning*, or *safe learning control*. I The author proposes the similar to the previous model considering the decomposition of the previous problem functions (f, l, and c) as *nominal* ($\bar{*}$) and *unknown* ($\hat{*}$); for instance

$$f(x_t, u_t) = \bar{f}(x_t, u_t) + \hat{f}(x_t, u_t, w_t)$$

where w_t is an stochastic parameter, therefore $\hat{f}$ represents the uncertain part of the dynamics, for instance. Under this approach, the problem objective is to use both known background, and collected data to approach an optimal policy.

In the case of two stage learning agent, *online* (during interaction, in an adaptative manner) and offline (between interactions, where the model is adjusted), tree methodologies are presented:

- Learning uncertain dynamics to safely improve performance: class of methods that learns the uncertain dynamics, while providing good safety guarantees from optimal control theoretical frameworks.
- Encouraging safety and robustness in RL: class of methods that doesn't assume prior knowledge, but rather incentivate safe learning.
- Certifying learning-based control under dynamics uncertainty: class of methods where safety certification is provided to non-safety-aware learners, by modifying it's output.

[1] see: `https://www.mit.edu/~dimitrib/home.html`

One may consider that each particular approach might be discussed in light of the problem to be tackled. In fact, problems involving an unknown dynamics, and ideal data, might offer good safety guarantees. On the other side, data sparse, and complex dynamics may not offer complete theoretical guarantees, but make good generalization from inverse learning.

For a complete overview of some of the techniques in each of the previous categories, one may refer to the really good work in [BGH+21].

5.1 Model Predictive Control and Robotics

Finally, one of the most extensively implied methodologies from learning-based optimal control in robotics are the learning-based MPC. The methodology has become apparent, as it provides a more tractable version of an (Stochastic) Optimal Control Problem by approximating the solution through a simplified version of the control problem over a shorter horizon [HWMZ20].

Those methods have recently been used to ensure safer learning, where some security restriction are enforced in the learning process. Interesting examples of this unification [Lab24, Ami24], where relevant problems in robotics are discussed in light of Reinforcement Learning and Optimal Control Based approaches, including more technical topics from OC, such as Reachable Sets to avoid the drone collision. Also, [Lab24] discusses and presents the need for sub-optimal techniques such as MPC for better computational tractability.

6 Other Machine Learning Results

Besides the well known intersection of Optimal Control and Reinforcement Learning, with focus on safety, other interesting results in Machine Learning from Optimal Control deserve some attention; some of those could even account for completely unusual approaches for interpretability and explainability.

As an example, the work in [SGS+21] proposes a discussion regarding the use of knowledge of the environment to simplify the learning process of an agent, specially in the context of a Reinforcement Learning problem. Ultimately, this kind of approach allows one to reduce the dimension of some learning problems, which itself opens space for more explainable AI, since one replaces the use of NNs for more interpretable tabular learning methodologies.

Also, the works of [BWST24, STCL23] open the application of an Optimal Control rationale to the field of Natural Language Processing, explaining

and giving analytical results on the *controlability* of LLM models. This kind of research allows one to rigorously justify and predict the outputs of an input prompt in this kind of tool, making better tests or checking for possibly threatening jailbreaks.

6.1 Optimal Control Theory and Neural Networks

The concept of Neural Networks is sometimes directly associated with the idea of back-box, as an extreme case of Machine Learning model hard to give theoretical approaches. Some of the next works give Optimal Control approaches to Neural Nets, offering a (new) point of view for those models.

Not recently, in [SS97] authors proposed and showed powerful results for controlability of Recurrent Neural Networks (RNNs). In fact, the authors were able to show complete controlability of the system under easy assumptions. It's interesting to compare such result with recent one from previously cited [BWST24], that offers similar results for the case of an LLM.

Not less interesting, the work in [Lec01] introduces a theoretical approach to the algorithm of back-propagation, one of the pillar for training modern Neural Networks. The author proposes that similar algorithm had been previously provided by researchers in the field of Optimal Control, considering a sequence of Neural Net layers as a problem similar to multistage optimal control. In his view, understanding back-propagation through this lens could offer insights for learning algorithms.

Authors in [Ee17] propose the discussion of Neural Networks as discrete time dynamical systems; this gives rise to the possibility of studying the continuous model, instead of the discrete. The advantages of such approach would be the easier analysis of the continuous model, and familiarity from different communities for continuous models. Also, the continuity allows for best numerical analysis, avoiding classical issues such as vanishing/exploding gradients in training Deep Neural Networks. The authors even propose the possible use of Partial Differential Equations to allow for the study of inherently continuous spatial data, such as images in computational vision. Finally, as part of the work, theoretical approaches from Optimal Control are considered for deriving approaches to some learning algorithms, such as stochastic gradient descent.

Finally, on a similar track from the previous ones, in [RBZ21] the authors discuss Neural Ordinary Differential Equations (NODE), as the continuous case of ResNets. The work proposes the problem of classification as a simultaneous Control Problem for the Cauchy problem, where a controller has to steer points from their initial positions to respective class sets in finite time.

In the text, the authors discuss how some of the intermediate results in the work allow for a better understanding of the supervised learning process.

7 Conclusions

In this work, we were able to introduce and review some of the aspects of AI Safety from the view of Optimal Control, with special focus on Reinforcement Learning. We justified the need for grounded learning-based practices allowing certain theoretical guarantees, but also adaptability and generality for learning agents, specially in the case in real world deployment.

We also commented on less known approaches from Optimal Control and Machine Learning, eventually leading to more interpretable or controlable models.

References

[AKS17] Thomas Arnold, Daniel Kasenberg, and Matthias Scheutz. Value alignment or misalignment - what will keep systems accountable? In *AAAI Workshops*, 2017.

[Ami24] Amii. Ai seminar series: Mo chen, optimal control and machine learning in robotics (jan 8). `https://www.youtube.com/watch?v=fzvVN572OyI`, jun. 2024.

[AOS+16] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *CoRR*, abs/1606.06565, 2016.

[Ber19] Dimitri P. Bertsekas. *Reinforcement Learning and Optimal Control*. Athena Scientific, 2019.

[Ber22] D. Bertsekas. *Lessons from AlphaZero for Optimal, Model Predictive, and Adaptive Control*. Athena Scientific optimization and computation series. Athena Scientific, 2022.

[Ber23] D. Bertsekas. *A Course in Reinforcement Learning*. Athena Scientific, 2023.

[BGH+21] Lukas Brunke, Melissa Greeff, Adam W. Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P. Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *CoRR*, abs/2108.06266, 2021.

[BP07] Alberto Bressan and Benedetto Piccoli. *Introduction to the mathematical theory of control*. American Institute of Mathematical Sciences, 2007.

[BSN20] Daniel S. Brown, Jordan Schneider, and Scott Niekum. Value alignment verification. *CoRR*, abs/2012.01557, 2020.

[BWST24] Aman Bhargava, Cameron Witkowski, Manav Shah, and Matt Thomson. What's the magic word? a control theory of LLM prompting, 2024.

[CLB+23] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2023.

[Ee17] Weinan Ee. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 5:1–11, 02 2017.

[GFoF15] Javier García, Fern, and o Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(42):1437–1480, 2015.

[GYD+24] Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, and Alois Knoll. A review of safe reinforcement learning: Methods, theory and applications, 2024.

[Henng] Dan Hendrycks. *Introduction to AI Safety, Ethics and Society*. Taylor & Francis, (forthcoming). `www.aisafetybook.com`.

[HWMZ20] Lukas Hewing, Kim P. Wabersich, Marcel Menner, and Melanie N. Zeilinger. Learning-based model predictive control: Toward safe learning in control. *Annual Review of Control, Robotics, and Autonomous Systems*, 3(1):269–296, 2020.

[JQC+24a] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Kwan Yee Ng, Juntao Dai, Xuehai Pan, Aidan O'Gara, Yingshan Lei, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao. Ai alignment: A comprehensive survey, 2024.

[JQC+24b] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Kwan Yee Ng, Juntao Dai, Xuehai Pan, Aidan O'Gara, Yingshan Lei, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao. Ai alignment: A comprehensive survey, 2024.

[KBP13] Jens Kober, J. Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32:1238–1274, 09 2013.

[KHD21] Tae Wan Kim, John Hooker, and Thomas Donaldson. Taking principles seriously: A hybrid approach to value alignment in artificial intelligence. *Journal of Artificial Intelligence Research*, 70:871–890, 02 2021.

[Lab24] CMU Robotic Exploration Lab. Optimal control 2024. `https://www.youtube.com/playlist?list=PLZnJoM76RM6Jv4f7E7RnzW4rijTUTPI4u`, jun. 2024.

[Lap18] Maxim Lapan. *Deep Reinforcement Learning Hands-On*. Packt Publishing, Birmingham, UK, 2018.

[Lec01] Yann Lecun. A theoretical framework for back-propagation. 08 2001.

[Lib11] Daniel Liberzon. *Calculus of Variations and Optimal Control Theory: A Concise Introduction*. Princeton University Press, 2011.

[LMK+17] Jan Leike, Miljan Martic, Victoria Krakovna, Pedro A. Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. AI safety gridworlds. *CoRR*, abs/1711.09883, 2017.

[NCM24] Richard Ngo, Lawrence Chan, and Sören Mindermann. The alignment problem from a deep learning perspective, 2024.

[NR00] Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, page 663–670, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[Ray23] Partha Pratim Ray. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3:121–154, 2023.

[RBZ21] Domènec Ruiz-Balet and Enrique Zuazua. Neural ode control for classification, approximation and transport, 2021.

[Rec18] Benjamin Recht. A tour of reinforcement learning: The view from continuous control, 2018.

[SB18] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

[SGS+21] Tim Seyde, Igor Gilitschenski, Wilko Schwarting, Bartolomeo Stellato, Martin A. Riedmiller, Markus Wulfmeier, and Daniela Rus. Is bang-bang control all you need? solving continuous control with bernoulli policies. *CoRR*, abs/2111.02552, 2021.

[SS97] Eduardo Sontag and Héctor Sussmann. Complete controllability of continuous time recurrent neural networks. *Systems & Control Letters*, 30:177–183, 05 1997.

[STCL23] Stefano Soatto, Paulo Tabuada, Pratik Chaudhari, and Tian Yu Liu. Taming ai bots: Controllability of neural states in large language models, 2023.

[Zua] Enrique Zuazua. Control and machine learning. `https://sinews.siam.org/Details-Page/control-and-machine-learning`. Accessed: 2024-05-22.