Review of Metrics to Measure the Stability, Robustness and Resilience of Reinforcement Learning

Laura L. Pullum, DSc
Computer Science and Mathematics Division, Oak Ridge National Laboratory
pullumll@ornl.gov

Abstract

Reinforcement learning (RL) has received significant interest in recent years, due primarily to the successes of deep reinforcement learning at solving many challenging tasks such as playing Chess, Go and online computer games. However, with the increasing focus on RL, applications outside of gaming and simulated environments require understanding the robustness, stability and resilience of RL methods. To this end, we characterize the available literature on these three behaviors as they pertain to RL. We classify the quantification approaches used, determine the objectives of the desired behaviors, and provide a decision tree for selecting metrics to quantify the behaviors.

1. Introduction

Recent literature on the robustness of machine learning models has focused almost entirely on the robustness of deep neural networks for imaging applications. However, at the time of this study, there are no published surveys on robustness of RL. With RL use increasing, especially in control systems contexts, we pursued this review. Included along with robustness are stability and resilience. Stability is included because the term has been used interchangeably with robustness and resilience is included because the term has been used as a state beyond robustness.

RL involves agents which take actions in an environment and experience at a reward for those actions. The agent is to learn a policy that maximizes the cumulative reward. Formally, consider an agent operating over time $t \in \{1, \ldots, T\}$. At time t, the agent is in environment state s_t and produces an action $a_t \in A$. The agent then observes a new state s_{t+1} and receive a reward $r_t \in R$. The set of possible actions A can be discrete or continuous. The goal of reinforcement learning is to find a policy $\pi(a_t|s_t)$ for choosing an action in state s_t to maximize a utility function or (expected return). [252]

$$J(\pi) = \boldsymbol{E}_{s_0,a_0,\ldots}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right]$$

where $0 \leq \gamma \leq 1$ is a discount factor; $a_t \sim \pi(a_t|s_t)$ is drawn from the policy; and $s_{t+1} \sim P(s_{t+1}|s_t,a_t)$ is generated by the environmental dynamics. The state value function

$$V^{\pi}(s_t) = \boldsymbol{E}_{a_t,s_{t+1},\ldots}\left[\sum_{i=0}^{\infty}\gamma^i r(s_{t+i},a_{t+i})\right]$$

is the expected return by policy *p* from state s_t. The state action function

$$Q^{\pi}(s_t,a_t) = \boldsymbol{E}_{s_{t+1},a_{t+1},\ldots}\left[\sum_{i=0}^{\infty}\gamma^i r(s_{t+1},a_{t+1})\right]$$

is the expected return by policy *p* after taking action a_t at state s_t. [252]

The objective of this manuscript is to present a systematic review of RL literature to identify metrics to measure the stability, robustness and resilience of RL. We limit RL to general reinforcement learning and not specialized RL, such as inverse RL. We reviewed papers that attempted to measure or otherwise characterize the stability, robustness and resilience of RL, seeking metrics for these behaviors.

We searched computer science and technical literature databases for eligible papers, combining RL, the behavior terms and terms related to measuring, metrics and quantification. The result comprised 16,015 items, which after removal of duplications and extraneous material, a collection of 546 items was established. Through a process of elimination described in full in the paper, we reduced the set to 248 papers. We systematically reviewed those 248 papers, and the results are presented in this analysis.

We classified the papers by behavior (i.e., stability (*n*=76), robustness (*n*=169), and resilience (*n*=3)) and identified the primary domains of application as robotics, network systems, power system control and vehicle/traffic control and navigation. We identified approaches to determining or measuring each behavior individually and those across behaviors. The approaches were categorized as quantitative or theoretical and the quantitative approaches were further classified as being applied internally (e.g., in training) or externally (e.g., performance measures on outputs) the model. Metrics, approaches and objectives were identified for each paper surveyed. The objective indicates to what the metric or approach was intended to be stable, robust or resilient. We close by indicating the need to define stability, robustness and resilience behaviors for RL and identify the quantitative and theoretical approaches to achieve measurement and determination of these behaviors.

There is a rich set of domains (i.e., 53 identified in this survey) in which measurement of RL stability, robustness and resilience have been conducted. The domains ranged from robotics and network systems to sheep herding and fish behavior. The most frequently mentioned domains include robotics, general control and network systems, with numerous papers not specifying a domain. Many papers used Gym [254] and other environments for demonstration. Though the search focused on quantitative measurement of stability, robustness and resilience, theoretical approaches were identified as well. The quantitative approaches were categorized as internal or external, dependent upon where in the model the evaluation was held. Internal measures quantified the performance of the training, where external measures quantified the ultimate performance of the model.

The goal of this systematic review was to identify metrics to measure the stability, robustness and resilience of RL. To initiate the search for this review, we identified keywords and phrases related to reinforcement learning, the behaviors of interest (stability, robustness and resilience) and measurement (shown in Table 1).

Table 1. Keywords and Phrases

Key Phrase	Behavior	Measurement	
reinforcement learning	stability	metric	measure
	robust*	index	score
	resilien*	quantifier	indicator

We believe that this is the first comprehensive review of stability, robustness and resilience specifically geared towards RL. The remainder of the paper is organized as follows. Section 2 describes the methods used in this systematic review. Section 3 presents the results of the review. Section 4 discusses the results of the review and introduces a decision tree for metric selection based on the review. Section 5 provides supplementary information.

2. Methods

Keywords salient to RL, system behavior and measurement were identified for the research topic and are shown in Table 1. The typical search was of the form

<Key Phrase> + <Behavior> + <Measurement>

with <Key Phrase>, <Behavior> and <Measurement> defined in Table 1. A specific example is

“reinforcement learning” AND robust* AND (“metric” OR “measure” OR “index” OR “score” OR “quantifier” OR “indicator”)

Multiple searches were conducted in bibliographic databases covering the broad areas of computer science, physical and biological sciences and engineering. See Table 12 for a list of information sources used in this study. No restrictions were placed on the publication’s date or language. Journal articles, books, books in a series, book sections or chapters, edited books, theses and dissertations, conference papers, and technical reports containing the keywords and phrases were included in the search. The publication date of search results returned are bound by the dates of coverage of each database and the date on which the search was performed, however all searches were completed by October 31, 2020. The range of dates for the documents ultimately included in the review was 2002-2020.

The databases queried resulted in 16,015 citations being collected. Irrelevant citations were also unwittingly retrieved. We removed extraneous studies resulting in a collection of 699 publications. Further, removing duplicate papers resulted in 580 publications. Citations for “Full Conference Proceedings” were removed if relevant paper(s) within the associated conference were otherwise collected, resulting in 546 publications. Further refinement excluded publications that were not on RL, that were not on the searched behavior or those that had no metrics or theoretical content, resulting in 248 documents. As a result, we systematically reviewed 248 papers and the results are presented in this analysis. See Figures 10, 11, and 12 for graphic summaries of the data reduction methodology for stability, robustness and resilience behaviors, respectively. See Table 13 for a tabular summary of the data reduction. In addition, see Checklist 1 for the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines used for the evaluation of the papers.

The 248 papers that made it through the screening process were grouped by the behavior searched, that is Stability, Robustness, and Resilience. We also identified those papers on one behavior that mention one or both of the other behaviors. Some papers that mentioned other behaviors did so interchangeably. For instance, stability and robustness were used interchangeably in several papers, which can lead to some of the confusion that exists in the definitions of these behaviors. The primary domains of application were identified and categorized as robotics, network systems, general control systems and Gym [254] and other environments. We also identified those publications that mentioned the RL policy.

The primary focus of the paper is to identify approaches to determining or measuring each behavior. Of course, most of the publications focused on quantitative approaches because of the search terms used. The ones that use a theoretical approach provide additional insight into the behavior determination problem. The quantitative approaches were further classified as being applied internal (e.g., in training) or external (e.g., performance measures on outputs) the model. Metrics, approaches and objectives were identified for each paper surveyed (see Figure 1). The objective indicates to what the metric or approach was intended to be stable, robust or resilient.

There is little agreement on the definitions of stability, robustness and resilience in the literature. In fact, there are few distinct definitions of these behaviors. For this review, we use the following definitions.

Stability is a property of the learning algorithm (that is, a small change in the training set results in a similar model) and refers to the ranking of the variance of a model [253]. For example, if we use variance of the loss function over all datasets as a performance measure and test a set of models. The smallest loss indicates the more stable model. Given this definition, stability analysis is the application of sensitivity analysis to machine learning.

Robustness, when used with respect to computer software, refers to an operating system or other program that performs well not only under ordinary conditions but also under unusual conditions that stress its designers' assumptions (http://www.linfo.org/robust.html). Robustness then is a property of the model and is measured by the, e.g., loss over all datasets (as opposed to the variance of the loss).

Throughout the literature, *resilience* has been used interchangeably with robustness, however, it is used most often with production machine learning systems to indicate robustness to different data sets and different data added to the data set.

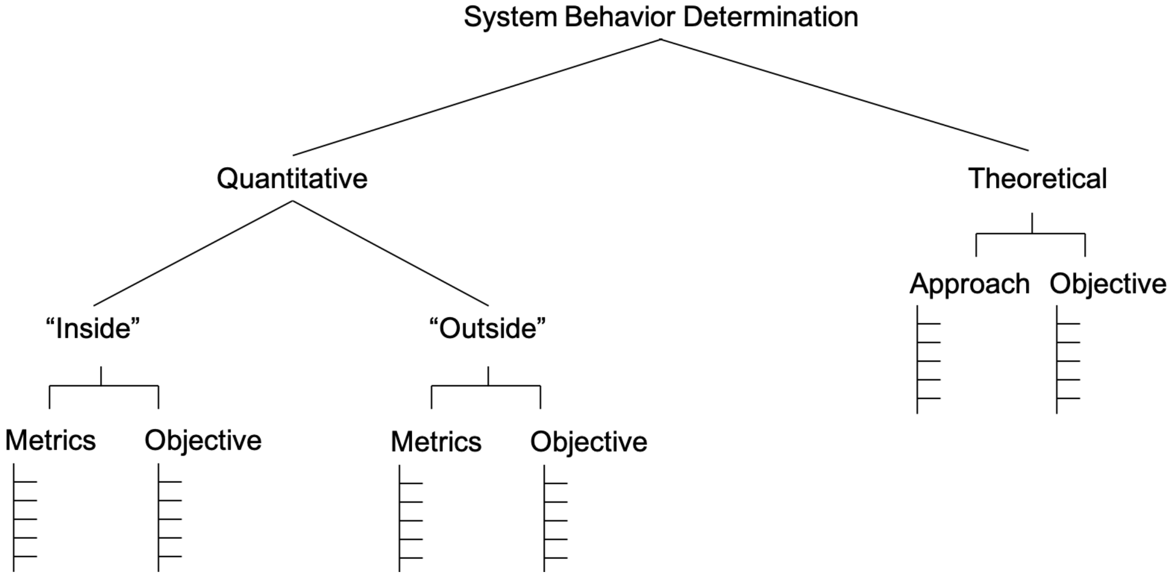


Figure 1. Categorization and resulting metrics, approaches and objectives

3. Results and Analysis

The publications were categorized by behavior (see Table 2): Stability (*n*=76), Robustness (*n*=169), and Resilience (*n*=3). Papers on one behavior often mention the other behaviors, especially Stability and Robustness (see Figure 2). Resilience is mentioned in 5 Stability papers and in 11 Robustness papers. Robustness is mentioned in 50 Stability papers and in 1 Resilience paper. Stability is mentioned in 104 Robustness papers and in all (3) Resilience papers.

Table 2. Citations categorized by behavior

Behavior	Citations	Total
Stability	[4-80]	76
Robustness	[81-250]	169
Resilience	[1-3]	3
	Total	248

Given the recent explosion of papers on robustness of neural networks to adversarial attacks, one might expect it to be a cornerstone of the robustness papers reviewed herein. The term “adversarial” is mentioned in a quarter (n=61, N=248) of the papers reviewed (see Figure 3). That is, 1 Resilience paper, 56 Robustness papers and 4 Stability papers mention “adversarial”. Some papers on one behavior used one of the other behaviors interchangeably, notably stability and robustness, specifically [91, 93, 105, 145, 146, 179, 194, 225, and 237] and generally in several other articles.

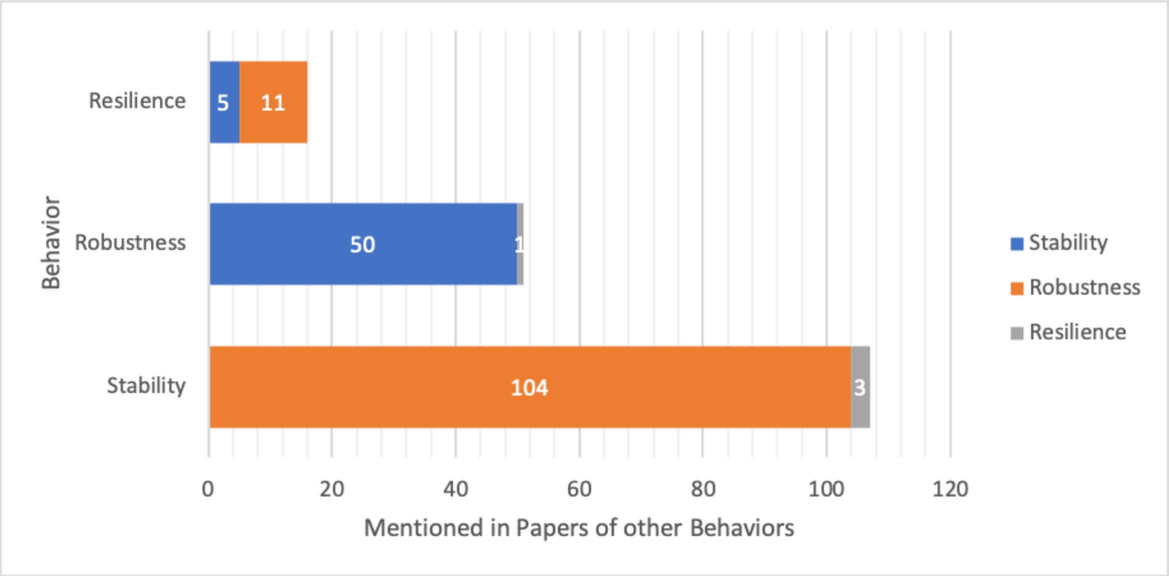


Figure 2. Stability, robustness and resilience papers were mentioned in papers on other behaviors. For example, Robustness is mentioned in 104 Robustness papers and in all 3 Resilience papers.

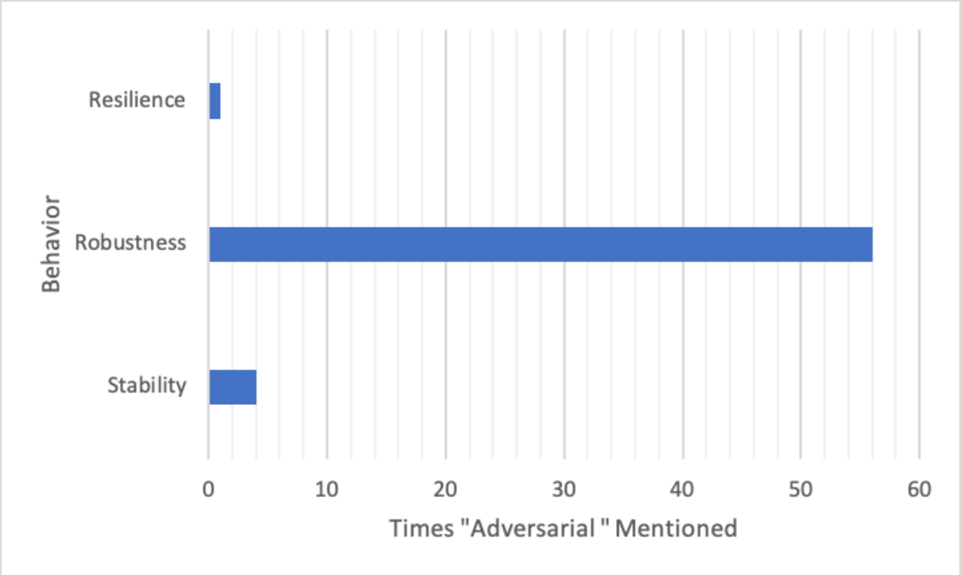


Figure 3. Number of papers in which “adversarial” is mentioned, categorized by behavior. For example, 56 Robustness papers mention “adversarial”.

3.1 Application Domains

The publications’ application domains are provided in Table 3 and illustrated in Figure 4. The primary domains were robotics with 16.4% (*n*=44) of the total citations (*N*=268), followed by network systems and general control (each with 7.8%, *n*=21), with 9.3% (*n*=25) using Gym or other environments as their experiment domain. Just as many (*n*=25, 9.3%) did not specify a domain. These top 5 (of 53) domains comprised over 50% (52.9%, *n*=136) of the citations. Most (52.8%, *n*=28) of the domains (*n*=53) had a single citation each.

Table 3. Citations categorized by application domain

Domain	Behavior(s)	Citations	Total
Robotics	Stability, Robustness	[9, 12, 16, 21, 27, 31, 44, 45, 46, 51, 56, 60, 70, 77, 78], [82, 86, 103, 109, 116, 118, 122, 125, 128, 133, 134, 140, 142, 149, 151, 158, 162, 164, 178, 190, 208, 211, 212, 217, 221, 224, 236, 242, 249]	44
Gym and other environments	Stability, Robustness	[8], [87, 101, 146, 182, 185, 191, 192, 195, 199, 200, 208, 109, 213, 225-229, 231, 233, 240, 245, 246, 250]	25
General, non-specified	Robustness	[132, 136, 145, 150, 171, 173, 174, 181, 184, 185, 195, 197, 198, 200, 205, 215, 216, 219, 223, 225, 229, 240, 247, 248, 250]	25
Network Systems	Stability, Robustness, Resilience	[6, 10, 11, 14, 32, 42, 66, 73, 77], [81, 84, 89, 96, 98, 104, 113, 115, 127, 203, 214], [3]	21
Control, not otherwise noted	Stability, Robustness	[64, 68, 69, 71, 74, 76, 79], [135, 137, 147, 163, 175, 188, 189, 194, 210, 218, 228, 230, 233, 244]	21
Vehicle/Traffic Control and Navigation, Collision Avoidance	Stability, Robustness	[16, 20, 38, 47, 49, 58, 65], [83, 100, 110, 111, 114, 123, 166, 177, 195, 226]	17
Games/Game Systems	Stability, Robustness	[24, 30, 33, 35, 37, 52], [156, 165, 168, 193, 201, 204, 220, 234, 239, 250]	16
Power System Control	Stability, Robustness	[4, 10, 15, 22, 23, 36, 43, 48, 50, 59, 67], [105, 107, 117, 172]	15
Image Analysis	Stability, Robustness	[72], [88, 93, 94, 99, 108, 148, 155, 235]	9
Nonlinear Dynamic Systems/Processes	Stability	[5, 25, 41, 54, 61, 62]	6
Continuous Control	Robustness	[157, 191, 231, 237, 245]	5

Table 3. Citations categorized by application domain

Domain	Behavior(s)	Citations	Total
Multi-agent Systems	Stability, Robustness	[18, 29, 52], [241]	4
Domain Agnostic	Stability	[7, 28, 40]	3
Manufacturing Systems	Stability, Robustness	[10, 63], [95]	3
Ride-share Dispatching, Delivery	Stability, Robustness	[19], [97, 160]	3
Medical - System Control, Medication Level/Control	Robustness	[112, 144, 187]	3
Autonomous Systems	Robustness	[143, 196, 207]	3
Security & Cyber Defense, Spammer Detection	Robustness	[138, 159, 179]	3
Uncertain Non-linear Systems	Robustness	[106, 121]	2
Conversation and Speech	Robustness	[119, 124]	2
Text Analysis, Machine Translation	Stability, Robustness	[13], [85]	2
Multi-armed Bandit	Stability, Robustness	[74], [120]	2
Feedback Controller	Stability, Robustness	[73], [129]	2
Information Retrieval (search)	Stability, Robustness	[17], [161]	2
Economics - Quantitative Investment, Trading Systems, Markets	Robustness	[102, 130, 206]	2
Video Presentation Quality	Stability	[26]	1
Brain-Machine Interface (BMI) Controller	Stability	[34]	1
Skill Acquisition	Stability	[53]	1

Table 3. Citations categorized by application domain

Domain	Behavior(s)	Citations	Total
Technical Process Feedback	Stability	[55]	1
Linear Quadratic Output Tracking	Stability	[39]	1
Decentralized Control	Robustness	[90]	1
Policy-selection Races	Robustness	[126]	1
Complex Adaptive Systems	Resilience	[1]	1
Emission Control	Stability	[57]	1
Flight Simulator Training	Resilience	[2]	1
Markov Decision Process	Stability	[75]	1
Cellular Communications	Robustness	[167]	1
Fish Behavior	Robustness	[141]	1
Space Telescope	Robustness	[131]	1
Brain Modeling	Robustness	[139]	1
Sensors	Robustness	[152]	1
Unsupervised Goal Exploration	Robustness	[153]	1
Directed Graph Embedding	Robustness	[154]	1
HVAC Control	Robustness	[169]	1
Train	Robustness	[170]	1
Teamwork	Robustness	[176]	1
Causal Discovery	Robustness	[180]	1
Turbo Fan Engines	Robustness	[183]	1
Sheep Herding	Robustness	[186]	1
Van de Pol Oscillator	Robustness	[222]	1
Single Episode Transfer	Robustness	[232]	1
Multi-Task Batch	Robustness	[238]	1
Spectral Efficiency	Robustness	[243]	1
		Total Citations	268
The categories are not mutually exclusive.		Total Domains	53

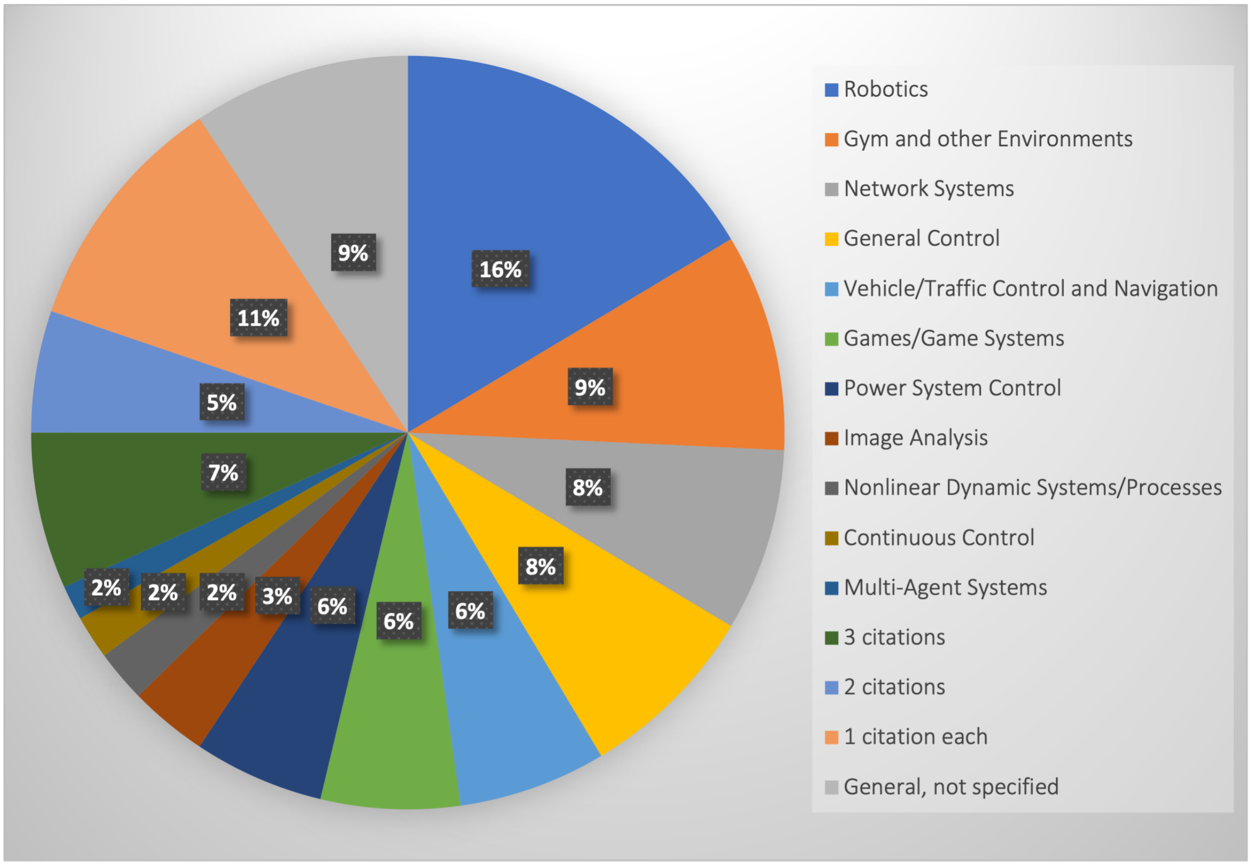


Figure 4. Application Domain Categories

3.2 Reinforcement Learning Policies

The types of RL policies mentioned in the articles are provided in Table 4. Most documents did not identify the policy used. If a behavior is not provided in Table 4 for a particular policy, that policy was not discussed in the papers on that behavior. Of the 21 types of policies mentioned, the top 4 – Actor-Critic (*n*=18), Q-learning (*n*=16), Proximal Policy Optimization (PPO) (*n*=8) and Adaptive Critic Design (*n*=5) comprise 72.3% of the total citations that included policy (*n*=65).

Table 4. Citations categorized by the reinforcement learning policy used

Policy	Behavior	Citations	Subtotal	Total
Actor-Critic	Stability	[5, 10, 46, 47, 49, 51, 65, 68, 69, 77, 79]	11	18
	Robustness	[131, 155, 180, 185, 191, 192, 222]	7	
Q-Learning	Stability	[18, 19, 39, 66, 72-75]	6	16
	Robustness	[92, 94, 107, 109, 113, 130, 138, 223, 236]	9	
	Resilience	[3]	1	
PPO	Robustness	[143, 147, 149, 158, 161, 162, 166, 208]	8	8
Adaptive Critic Design	Stability	[43, 54, 61]	3	5
	Robustness	[110, 125]	2	
Deep Deterministic Policy Gradient (DDPG)	Robustness	[162, 167]	2	2
4 variational Model-based Policy Optimization	Robustness	[181]	1	1
Advantage Actor Critic (A2C)	Robustness	[143]	1	1
Ad-hoc On-demand Distance Vector (AODV)	Robustness	[123]	1	1
Active Tracking Target Network (ATTN) and Anytime Reduced Value Iteration (ARVI)	Robustness	[177]	1	1
Constant Feedback to Control Policy	Robustness	[106]	1	1

Table 4. Citations categorized by the reinforcement learning policy used

Policy	Behavior	Citations	Subtotal	Total
Constraint-controlled PPO (CPPO)	Robustness	[164]	1	1
Data-regulated Actor-Critic (DrAC)	Robustness	[171]	1	1
Deep Deterministic Policy Gradient	Stability	[70]	1	1
Generalized Advantage Estimation	Robustness	[158]	1	1
Goal Policy	Robustness	[153]	1	1
Gradient	Robustness	[179]	1	1
Graph-based Policy Learning	Robustness	[176]	1	1
Lyapunov-based Actor-Critic (CLAC)	Robustness	[183]	1	1
MOOSE (MOdel-based Offline policy Search with Ensembles)	Robustness	[151]	1	1
Natural Stochastic Policies	Robustness	[184]	1	1
Student t Policy	Robustness	[146]	1	1
		Total Citations including Policy		65
		Total Policies		21

3.3 Approach to Determining or Measuring Behavior

The publications' approaches to determining or measuring each behavior are categorized as either quantitative or theoretical (Table 5). Most of the publications focused on quantitative approaches (*n*=205, 82.0%), which is understandable given that the search focused on quantifying the behaviors. For publications on the stability behavior, there was an almost even split between quantitative (n=42) and theoretical (n=43) approaches. However, publications on the robustness behavior were primarily focused on quantitative approaches. All (3) resilience publications applied quantitative approaches.

Table 5. Citations categorized by approach to determining or measuring behavior

Approach	Behavior	Citations	Total
Quantitative	Stability	[4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 19, 20, 21, 23, 26, 27, 28, 29, 30, 31, 32, 34, 36, 38, 42, 43, 45, 50, 51, 53, 58, 59, 66, 70, 72, 73, 75, 78, 80]	42
	Robustness	[81-89, 92-104, 106-120, 122-196, 198-201, 203-222, 224-227, 229, 231-247, 249-250]	160
	Resilience	[1, 2, 3]	3
		Total Quantitative	205
Theoretical	Stability	[4, 5, 10, 13, 15, 16, 20, 22, 24, 25, 27, 33, 35, 36, 37, 39, 40, 41, 43, 44, 46, 47, 48, 49, 52, 54, 55, 56, 57, 60, 61, 62, 63, 64, 65, 67-69, 71, 74, 76-79]	44
	Robustness	[90, 91, 97, 105, 116, 121, 135, 145, 150, 165, 184, 185, 194, 195, 197, 200-202, 205, 206, 215, 216, 222, 223, 225-230, 233, 239, 240, 248, 250]	35
	Resilience	--	0
		Total Theoretical	79

If a document covered both quantitative and theoretical approaches, it was placed in both categories.

3.3.1 Types of Quantitative Approaches

Next, we further categorize the quantitative approaches into whether they are focused internal or external to the model (see Table 6). Internal quantitative approaches measure aspects within the model, e.g., its training and associated measures such as the value of rewards over time or the number of episodes until convergence. External quantitative approaches measure performance-related aspects of the model, e.g., variations in accuracy or throughput. Most (n=141, 63.2%) of the quantitative approaches were categorized as performance-related, or external, measures. Of these, most (n=103) were on robustness, with stability (n=36) next. The 3 papers on resilience focused on performance-related quantitative measures. Robustness also led the internal approaches (*n*=68) with stability following (*n*=14). These are primarily due to the large number of robustness papers (*n*=170) and paucity of resilience papers (*n*=3) overall. Of the robustness papers, 40.0% (*n*=68) contained internal quantitative measures and 60.6% contained external quantitative measures. For stability, these values are 18.2% and 46.8%, respectively.

Table 6. Quantitative approaches categorized by internal or external measures

Quantitative Approach	**Behavior**	**Citations**	**Total**
Internal	Stability	[7, 8, 9, 11, 13, 19, 20, 27, 29, 30, 43, 72, 75, 78]	14
	Robustness	[83, 84, 92, 94, 97, 101-103, 109, 114, 122, 130, 132, 134, 138, 143, 146, 149-151, 155-157, 163-164, 166, 174-176, 181-182, 185, 187-189, 191-193, 195, 198-200, 204-205, 208-209, 211-213, 215, 217, 220, 226, 229, 231-239, 241, 244-247, 250]	69
	Resilience	--	0
		Total Internal Measures	**83**
External	Stability	[6, 9, 10, 12, 13, 14, 15, 17, 18, 19, 20, 21, 23, 26, 27, 28, 31, 32, 34, 36, 38, 42, 45, 50, 51, 53, 58, 59, 66, 68, 70, 73, 75, 78, 80]	36
	Robustness	[81-83, 85-89, 93, 95-100, 102, 104, 106-120, 123-129, 131, 133, 135-145, 147-148, 152-154, 158-162, 165-173, 177-181, 183-184, 186, 190, 194-196, 201, 203-204, 206-207, 210, 214, 216, 218-219, 221-222, 224-227, 240, 242-243, 249]	103
	Resilience	[1, 2, 3]	3
		Total Performance-related Measures	**142**

If there were both internal and external quantitative approaches in a document, the document was placed in both categories.

3.3.2 Types of Internal Quantitative Approaches

Looking at the types of internal quantitative approaches, we see a fairly narrow set of aspects being considered in the papers (see Table 7). These are metrics specifically made to measure stability other than by the variance of the output. They essentially measure variation in training performance. The vast majority (*n*=75, 88.2%) of the internal quantitative approaches calculate reward- or score-based metrics. Other types of internal quantitative approaches include 2 each of policy entropy, variations in control strategy approximation weights, and the convergence rate, and 1 each of policy weight, calculation of the Lyapunov stability criteria and calculation of the Wasserstein function lower bound. The term *convergence*, in RL context, refers to the stability of the learning process (and the underlying model) over time [11].

Table 7. Internal Quantitative approaches categorized by metric

Internal Quantitative Metric	**Behavior**	**Citations**	**Total**
Reward or Score – magnitude, mean/ variance, variation in average reward, time to threshold, episode duration	Stability, Robustness	[7, 8, 9, 13, 30, 72, 75, 78] [83, 84, 92, 94, 97, 101-103, 109, 114, 122, 130, 132, 134, 138, 143, 146, 149-151, 155-157, 163-164, 166, 174-176, 181-182, 185, 188-189, 191-193, 195, 198-200, 204, 208-209, 211-213, 215, 217, 220, 226, 229, 231-239, 241, 244-247, 250]	75
Policy entropy	Stability	[11, 19]	2
Variations in control strategy approximation weights	Stability, Robustness	[20] [120]	2
Convergence rate	Stability	[27, 29]	2
Lyapunov stability criteria calculated	Stability	[43]	1
Policy weight	Robustness	[231]	1
Regret	Robustness	[187]	1
Wasserstein function bounds calculated	Robustness	[205]	1
		Total	**85**

If a document had metrics that fell in more than one category, the document was placed in each of the categories.

3.3.3 Types of External Quantitative Approaches

The external or performance-based quantitative approaches to measuring the behaviors primarily (*n*=39) used deviations or variation in performance-related metrics other than precision, accuracy or recall (see Table 8 and Figure 5). The next category (*n*=28) of quantitative metrics used error, failure and success rates. Statistics on the performance of the tracking or estimation error follows with *n*=23 papers. Papers in the network domain used network-related metrics (*n*=15) to measure the behavior. Statistics on precision, accuracy and recall (*n*=12) followed. Five papers used variance in loss or regret estimation, 3 papers used game-related performance measures to quantify behavior and 2 papers each used bounds on or the size of the stability region and terminal wealth and inventory. Eighteen (18) additional different types of external quantitative metrics categories were represented by a single paper each.

3.3.4 Quantitative Approach Objectives

An additional aspect reviewed was to what action or event were the quantitative approaches attempting to be stable, robust or resilient. We call this the *<behavior> objective*. The *<behavior>* objective category (see Table 9) with the highest number of citations was geared toward handling changes in the operational environment or a dynamic environment or network (*n*=41). Papers that did not specifically state their objective comprised the next most populous category (*n*=35). The objective of handling uncertainty and disturbances in the environment also contained *n*=35 papers. The remaining objectives included input variation/perturbations (*n*=20); differences between training and test or operational environments (*n*=19); differences or uncertainties in model parameters (*n*=16); adversarial attack (*n*=14); different domains, environments or settings (*n*=8); errors or failures in operational environment (*n*=5); differences in training data sets or initializations (*n*=5); high variability (*n*=2) and one paper each in systematic pressure, spamming, incomplete data, and unknown control coefficients.

Table 8. External Quantitative approaches categorized by metric

External Quantitative Metric	Behavior	Citations	Total
Deviations/variation in other (than precision, accuracy and recall) performance-related metrics	Stability, Robustness, Resilience	[9, 13, 15, 19, 20, 26, 27, 36, 38, 45, 50, 51, 58, 59, 68, 80] [85, 95, 97, 106, 107, 110, 111, 119, 126, 137-139, 141, 142, 152, 169, 178, 179, 181, 183, 184, 195] [2]	39
Error and failure rates/success rate	Stability, Robustness	[12] [86, 88, 100, 104, 112, 124, 129, 131, 133, 136, 140, 143, 153, 158, 160, 162, 166-168, 186, 204, 207, 221, 222, 225, 242, 249]	28
Performance of tracking/trajectories estimation error; mean absolute deviation, mean square error, mean absolute percentage error, margins and magnitude of correlation coefficient	Stability, Robustness	[10, 18, 23, 28, 31] [87, 88, 93, 96, 106, 116, 125, 144, 145, 147, 148, 158, 173, 186, 201, 224, 226, 240]	23
Network-related timing/delay, path and link metrics, connectivity, delivery ratio, routing loops, path optimality, visitation distribution, structural Hamming distance, Small base station-serving ratio, sum-rate and 5th percentile rate	Stability, Robustness	[6, 14, 32, 42, 53, 66, 73] [83, 89, 123, 127, 170, 180, 203, 214]	15
Mean/average and variation in accuracy, precision and recall, area under the receiver operating characteristic (ROC) curve (AUC)	Stability, Robustness, Resilience	[17, 34] [93, 99, 102, 108, 113, 154, 159, 161, 168, 227] [3]	12
Variance of the estimation of loss, regret	Robustness	[118, 159, 187, 216, 224]	5

Table 8. External Quantitative approaches categorized by metric

External Quantitative Metric	Behavior	Citations	Total
Game-related performance - scores of game playing, percent wins, exploitability	Robustness	[109, 120, 165]	3
Size of the stability region; bounds	Stability, Robustness	[21] [135]	2
Terminal wealth, terminal inventory, cost, Sharpe ratio	Robustness	[206, 218]	2
Average proportion of failed eavesdropping attempts and of jammed red-force nodes; Average throughput	Robustness	[81]	1
Hours of operation with some maintenance	Robustness	[82]	1
Response time, energy consumption, and execution time	Robustness	[98]	1
Spectral efficiency	Robustness	[243]	1
Time headway (sec)	Robustness	[210]	1
Covariance analysis as a metric	Robustness	[114]	1
Mutual information	Robustness	[177]	1
Fidelity	Robustness	[219]	1
Expected rank and Robust Measurements metric	Robustness	[115]	1
Singular value decomposition (SVD)-based controllability measure	Robustness	[117]	1
Time to find goal/destination	Robustness	[128]	1
Number of adversarial actions required to cause error	Resilience	[1]	1
Normalized Energy Stability Margin (NESM)	Stability	[70]	1
Voltage violation rate, active power loss	Robustness	[172]	1
Jensen-Shannon divergence (JSD)	Robustness	[171]	1
Number of successful steps	Robustness	[190]	1
Blood glucose responses, Insulin concentration	Robustness	[194]	1

Table 8. External Quantitative approaches categorized by metric

External Quantitative Metric	Behavior	Citations	Total
Likelihood (Mahalanobis Distance)	Robustness	[196]	1
		Total	**147**

If a paper contained quantitative approaches belonging to multiple categories, it was placed in each of the relevant categories.

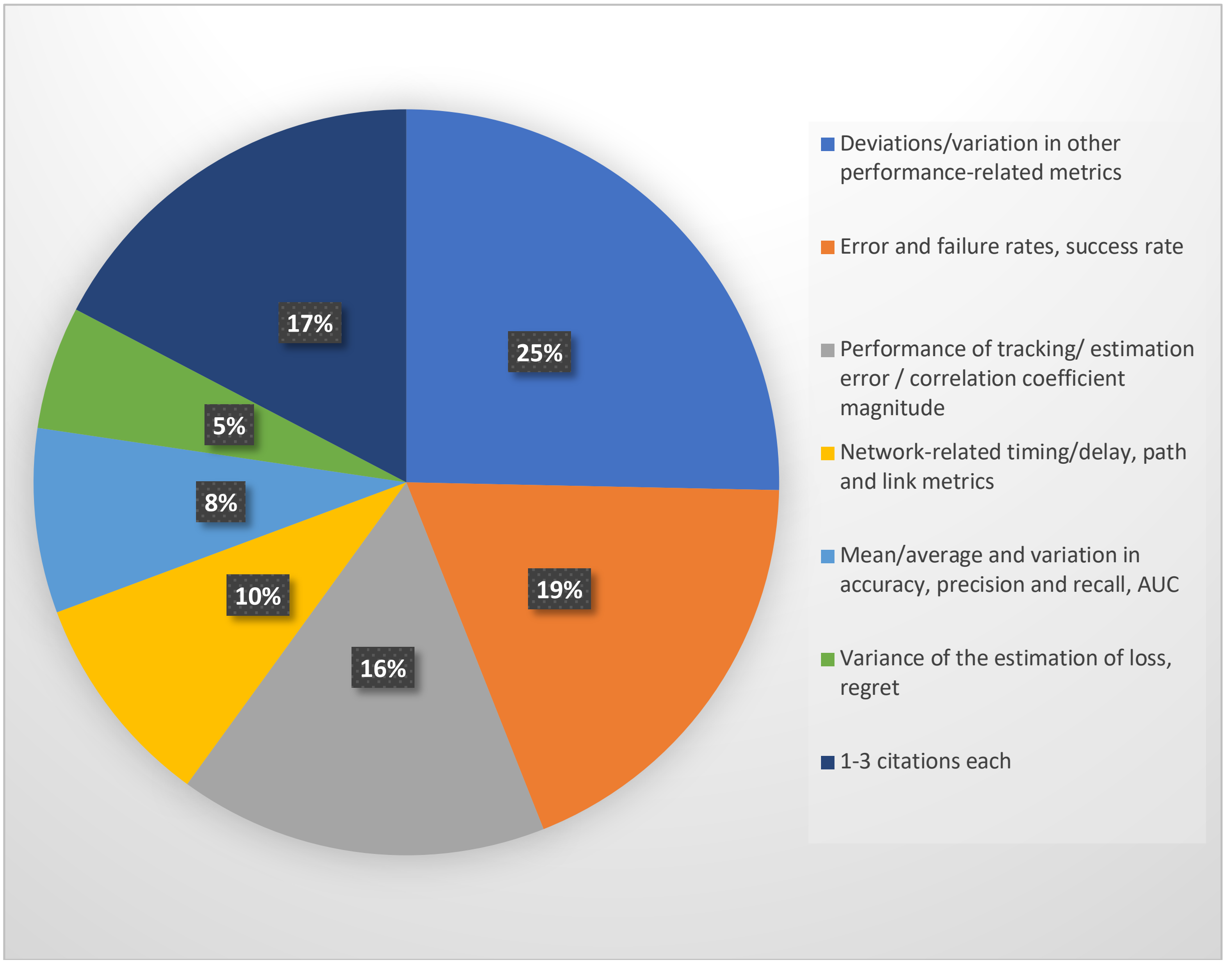


Figure 5. External quantitative metrics

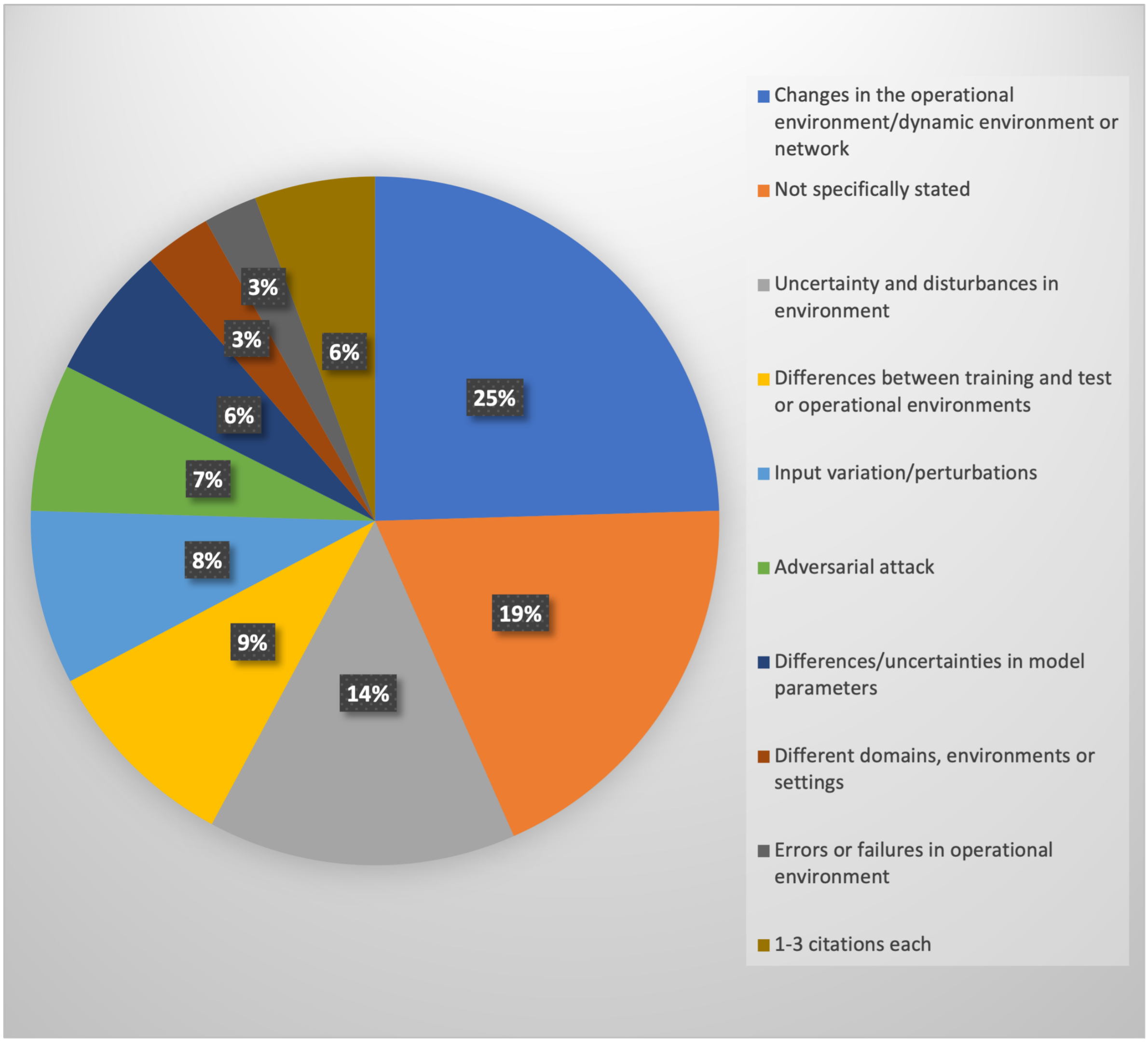


Figure 6. Quantitative <*behavior*> objectives

3.3.5 Types of Theoretical Approaches

A majority of the types of theoretical approaches in the papers reviewed were based on Lyapunov theory (n=50, 61.0%) (Table 10). The next highest types of theoretical approaches used are convergence to Nash equilibrium (n=10) and value-based guarantees such as error and output deviation bounds (n=8). Of the remainder, 3 papers used the Wasserstein distance to explore stability, 3 papers proved the methods were doubly robust, 2 papers proved the methods exhibited Lipschitz continuity, and stochastic stability theory to prove stability, stability guarantees, policy-based guarantees, regret bounds, minimization of the Jacobian on input, and per-episode Bellman-error regret guarantees/bounds were used by a single paper each to establish stability of the RL methods discussed.

Table 9. Quantitative <*behavior*> objectives

<***behavior***> **Objective**	**Behavior(s)**	**Citations**	**Total**
Changes in the operational environment/dynamic environment or network; distribution shift	Stability Robustness Resilience	[26, 27, 32, 45, 50, 66, 70, 73] [85, 86, 89, 93, 94, 96, 98, 99, 102, 107, 108, 111, 112, 114, 117, 118, 122, 123, 128, 131, 142, 154, 167, 171, 172, 177, 187, 194, 207, 214, 220, 243] [3]	41
Not specifically stated	Stability Robustness	[6, 8, 9, 10, 11, 13, 14, 15, 17, 18, 20, 21, 23, 28, 31, 38, 43, 51, 58, 59, 78] [113, 116, 125, 127, 130, 134, 165, 184, 198, 209, 216, 227, 240, 247]	35
Uncertainty and disturbances in environment, e.g., noisy sensor data, measurement noise, distractors, nuisances	Robustness	[133, 135, 138, 140, 141, 147, 148, 152, 153, 162, 169, 173, 175, 178, 186, 188, 190, 191, 193, 195, 196, 199, 208, 211, 217, 218, 219, 222, 224, 225, 234-236, 244, 249]	35
Input variation/perturbations, outliers	Stability Robustness	[7, 12, 19] [101, 103, 106, 132, 139, 143, 146, 149, 157, 158, 170, 187, 195, 208, 221, 226, 231]	20
Differences between training and test or operational environments	Stability Robustness	[31, 42] [82-84, 87, 95, 100, 109, 119, 120, 124, 129, 136, 149, 174, 191, 221, 245]	19

Table 9. Quantitative <*behavior*> objectives

<*behavior*> Objective	**Behavior(s)**	**Citations**	**Total**
Differences/uncertainties in model [hyper-]parameters, model error	Stability Robustness	[36, 53, 75] [92, 110, 126, 137, 164, 174, 181, 182, 192, 201, 225, 237, 241]	16
Adversarial attack	Stability Robustness Resilience	[29, 30] [81, 97, 156, 159, 163, 166, 204, 206, 210, 212] [1, 2]	14
Different domains, environments or settings	Stability Robustness	[80] [161, 164, 176, 203, 224, 232, 242]	8
Errors or failures in operational environment	Robustness	[104, 115, 145, 168, 189]	5
Differences in training data sets or initializations	Stability Robustness	[34] [88, 151, 213, 238]	5
High variability	Robustness	[144, 183]	2
Systematic pressure, e.g., sudden surge of requests	Robustness	[160]	1
Spamming	Robustness	[179]	1
Incomplete data	Robustness	[180]	1
Unknown control coefficients	Stability	[68]	1
		Total	**204**

If a technique had multiple robustness objectives, it was placed in each of those objectives.

Table 10. Theoretical approaches categorized by specific approach

Theoretical Approach	**Behavior**	**Citations**	**Total**
Lyapunov stability theory	Stability	[4, 5, 10, 15, 16, 20, 25, 35, 36, 39, 40, 41, 43, 44, 46, 47, 48, 49, 52, 54, 55, 56, 57, 60, 61, 62, 64, 65, 67, 69, 74, 76-79]	50
	Robustness	[90, 91, 97, 105, 116, 121, 135, 145, 194, 200, 222, 225, 230, 233, 248]	
Convergence to Nash equilibrium	Stability	[22, 24, 37, 63, 71]	10
	Robustness	[165, 197, 206, 215, 230, 239]	
Value-based guarantees, error bounds, output deviation bounds	Robustness	[195, 201, 226-228, 230, 233, 250]	8
Wasserstein distance	Robustness	[185, 201, 205]	3
Prove double robustness	Robustness	[184, 216, 240]	3
Prove Lipschitz continuity	Robustness	[201, 229]	2
Stochastic stability theory	Stability	[33]	1
Prove stability guarantees	Stability	[68]	1
Policy-based guarantees	Robustness	[201]	1
Regret bounds	Robustness	[202]	1
Minimize Jacobian on input	Robustness	[150]	1
Per-episode Bellman-error regret guarantees/bounds	Robustness	[223]	1
		Total	**82**

If a paper used multiple theoretical approaches, the was placed in each of those categories.

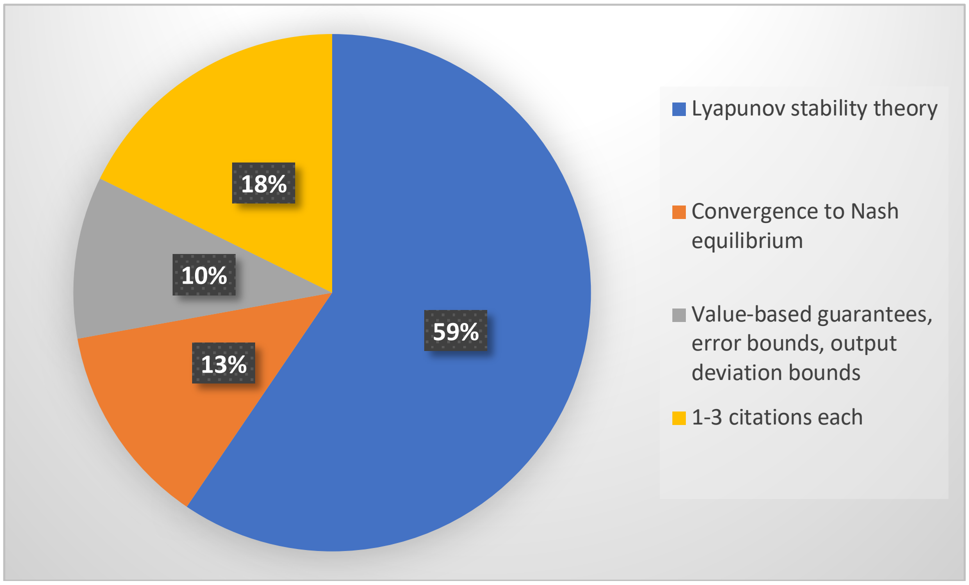


Figure 7. Theoretical Approaches

3.3.6 Theoretical Approach Objectives

We also reviewed the *<behavior> objective* for theoretical papers (Table 11). Most (n=42, 54.5%) papers on theoretical approaches did not state their objective. Of those few that did, changes or dynamics in the operational environment was the most frequent objective (n=10), followed by differences or uncertainties in model parameters (n=7), adversarial attack (n=6), error or failure (n=5), differences between training and test or operational environments (n=2), input variation (n=2), then 1 each for domain shifts, different function approximation architectures and differences in quantization levels.

4. Discussion

Our study was conducted to characterize published means of measuring or determining the stability, robustness or resilience of RL. Out of an initial collection of 16,015 items 248 papers met inclusion criteria and were systematically reviewed. Approaches to measuring or determining the behavior were classified as either quantitative or theoretical. Quantitative approaches were further classified as internal or external depending on whether they evaluated the training phase or the test or operational phases. For both categories of quantitative approaches, we categorized the metrics used, with internal approaches primarily using the reward or score (and statistics on same) and external approaches primarily using variations on performance-related metrics

Table 11. Theoretical <*behavior*> objectives

<*behavior*> Objective	**Behavior(s)**	**Citations**	**Total**
Not specifically stated	Stability Robustness	[4, 5, 10, 15, 16, 20, 22, 24, 25, 33, 37, 39, 41, 43, 44, 46, 47, 48, 49, 52, 54, 56, 57, 60, 61, 62, 63, 64, 65, 67, 74, 76, 78, 80] [90, 105, 116, 121, 165, 184, 197, 200, 216, 227, 240]	42
Changes in the operational environment/dynamic environment, environment uncertainties, environment disturbances	Stability Robustness	[35, 79] [135, 185, 194, 195, 222, 225, 228, 248]	10
Differences/uncertainties in model parameters	Stability Robustness	[36, 55, 68] [91, 201, 205, 225]	7
Adversarial attack, corruption, perturbations	Robustness	[150, 206, 215, 223, 230, 239]	6
Error, failure	Stability Robustness	[69, 71] [135, 145, 202]	5
Differences between training and test or operational environments	Stability Robustness	[40] [215]	2
Input variation or perturbation	Robustness	[195, 226]	2
Domain shifts	Robustness	[229]	1
Function approximation architecture	Robustness	[233]	1
Quantization level	Robustness	[250]	1
		Total	**77**

If a paper used multiple theoretical approaches, it was placed in each of those categories.

(though not precision, accuracy or recall). Theoretical approaches were dominated by the use of Lyapunov stability theory. We further characterized the objectives of the stability, robustness and resilience behaviors. Quantitative approaches to measuring the behavior focused on the ability to handle differences in the operational environment, whereas the vast majority of theoretical approaches to determining the behavior did not specifically state an objective. However, the objective of the theoretical approaches can be implied by the use of Lyapunov stability theory, that is, to prove the stability of the system. Lyapunov was used regardless of whether the article was on stability or robustness.

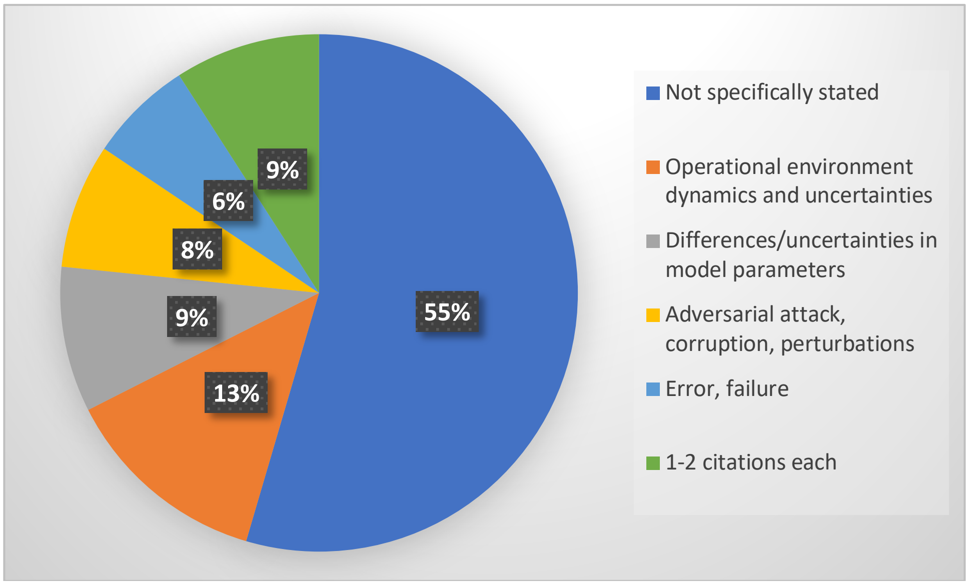


Figure 8. Theoretical <*behavior*> objectives

As an aid to deciding which metric to use, we developed a decision tree based on the information obtained in this literature review. It is a collapsible tree so that branches are not exposed unless selected and open branches can be closed or collapsed. There are several levels in the decision tree, starting with the i) behavior (stability, robustness or resilience); ii) the domain (see Table 3); iii) a list of quantitative and theoretical objectives (see Tables 9, 11); iv) the next level divides the metrics into external, internal and theoretical metrics (see Tables 5, 6); and v) the last level, i.e., the leaves, is the set of metrics for that branch of the decision tree (see Tables 7, 8). For example, suppose we want to find a suitable metric to measure robustness of a control system that is expected to face changes in the operational environment. From the metric decision tree shown in Figure 14, we see that the first selection is for a robustness metric. This selection displays the domains in which robustness metrics were described. Selecting the General Control domain reveals 5 quantitative objectives and 4 theoretical objectives, including the objective “Changes in the Operational Environment” in both the quantitative and theoretical objectives. An external metric found in the literature for this case is “blood glucose response” which is not applicable for this control system. The more appropriate metrics and approaches are Lyapunov stability theory and calculation, size of the stability region and value-based guarantees. One or all of these can be used to measure robustness of a general control system to changes in the operational environment.

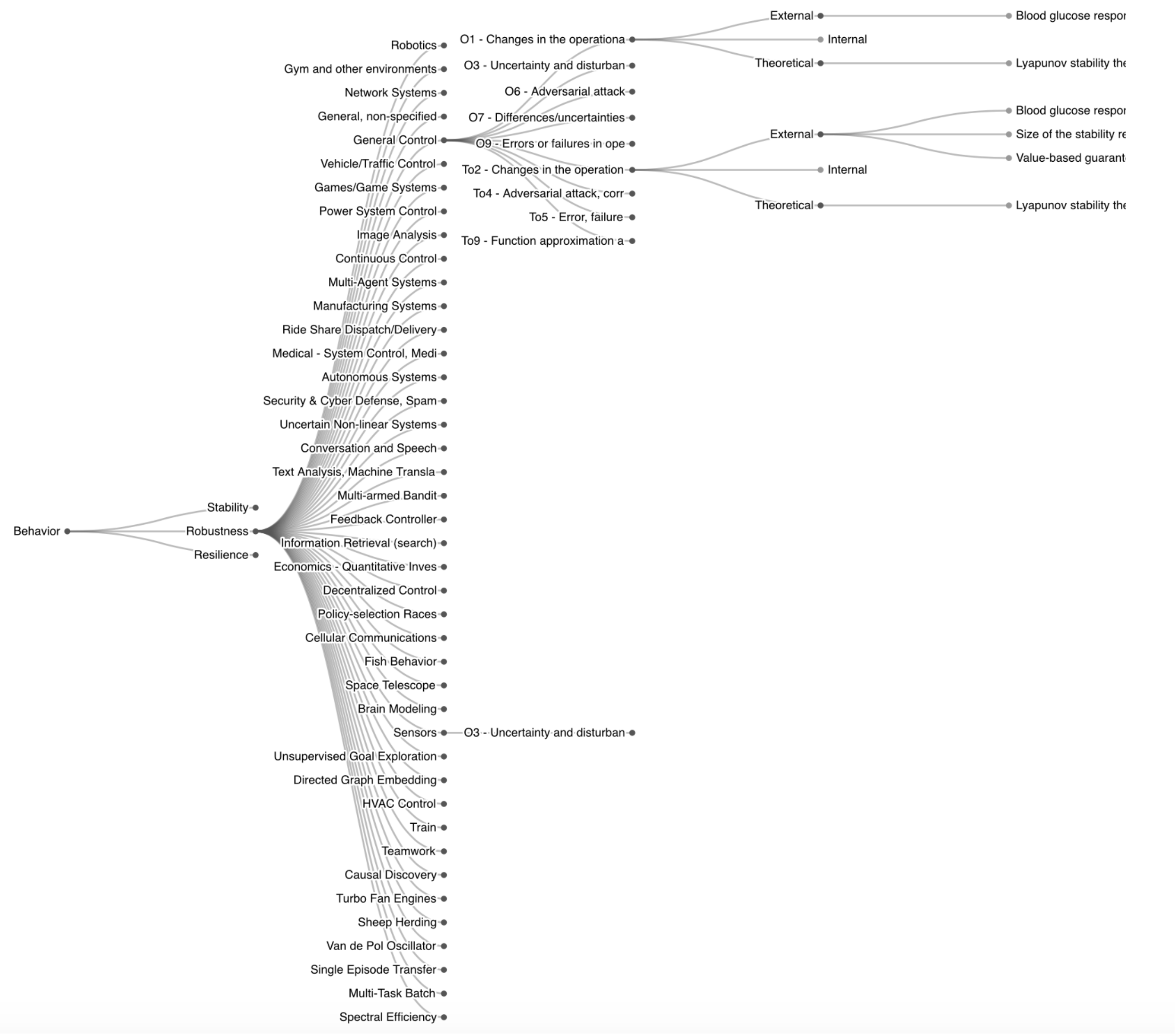


Figure 9. Metric Selection Decision Tree

5. Supporting Information

The databases in Table 12 were readily available to the author through Oak Ridge National Laboratory library subscriptions. Databases selected for this study are international in scope and ensure we investigate relevant studies that are global and cover a wide range of application domains, thereby eliminating the risk of bias from the author subject matter expertise and a Western perspective. However, we are bound by the content of these databases.

Many abstracting and indexing databases have very broad coverage, which results in individual databases duplicating content found in others. Duplicated bibliographic records were identified and removed. After duplicate entries were removed, the remaining citations were reviewed for completeness in terms of information needed to obtain the document from a library. If the citation did not include this information (e.g., author name(s), article title, and journal name), the missing information was obtained from the source database or other online sources and the citation was manually corrected.

Table 12. Information sources used in study

Database	**Dates of Coverage**	**Type of Database**
arXiv	1991-present	Publicly Available / Open Access
Scopus	1823-present	Subscription
Web of Science	1900-present	Subscription

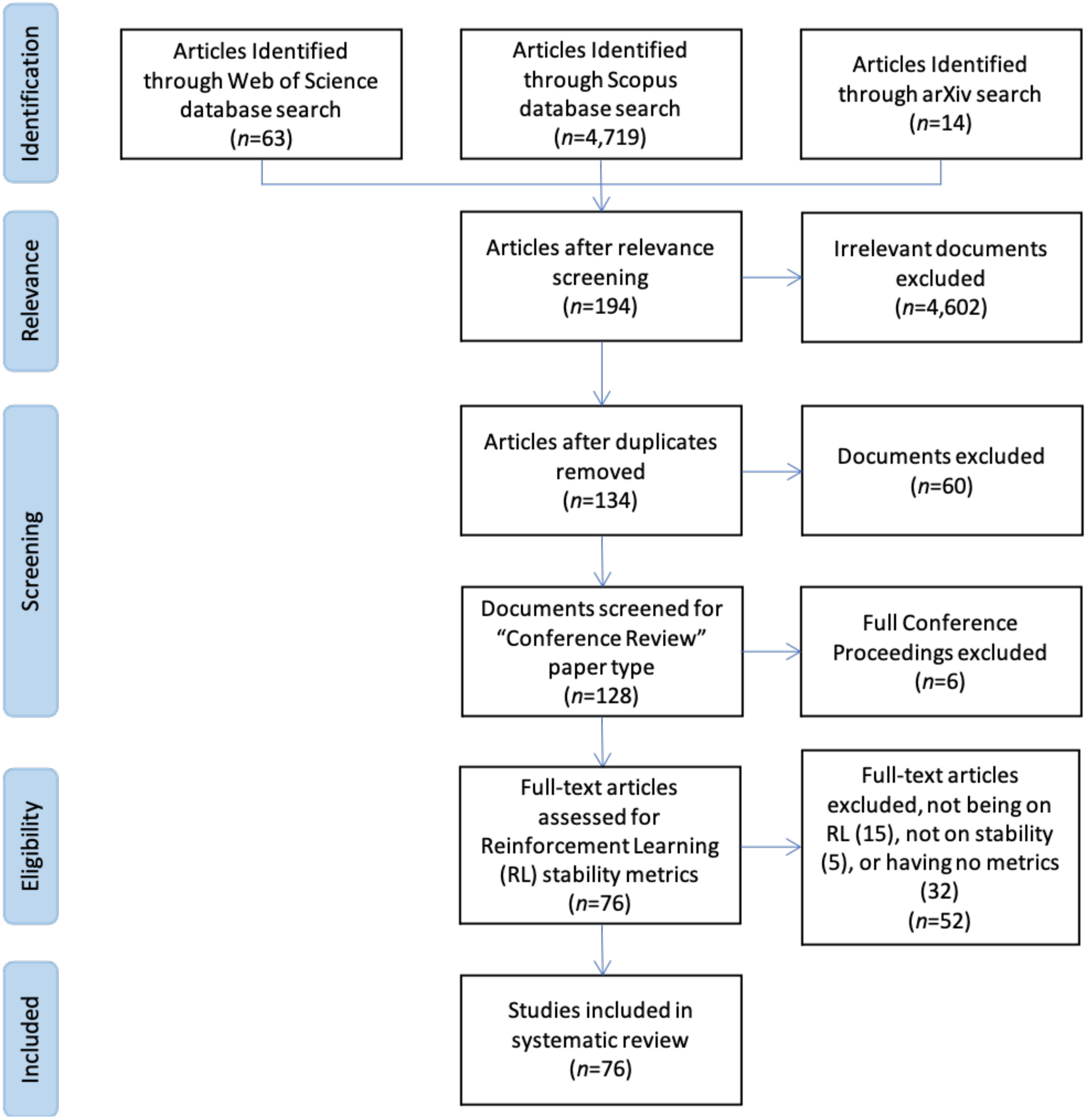

Identification
Articles Identified through Web of Science database search (*n*=63)
Articles Identified through Scopus database search (*n*=4,719)
Articles Identified through arXiv search (*n*=14)
Relevance
Articles after relevance screening (*n*=194)
Irrelevant documents excluded (*n*=4,602)
Screening
Articles after duplicates removed (*n*=134)
Documents excluded (*n*=60)
Documents screened for "Conference Review" paper type (*n*=128)
Full Conference Proceedings excluded (*n*=6)
Eligibility
Full-text articles assessed for Reinforcement Learning (RL) stability metrics (*n*=76)
Full-text articles excluded, not being on RL (15), not on stability (5), or having no metrics (32) (*n*=52)
Included
Studies included in systematic review (*n*=76)

Figure 10. The PRISMA Flow Diagram for Stability

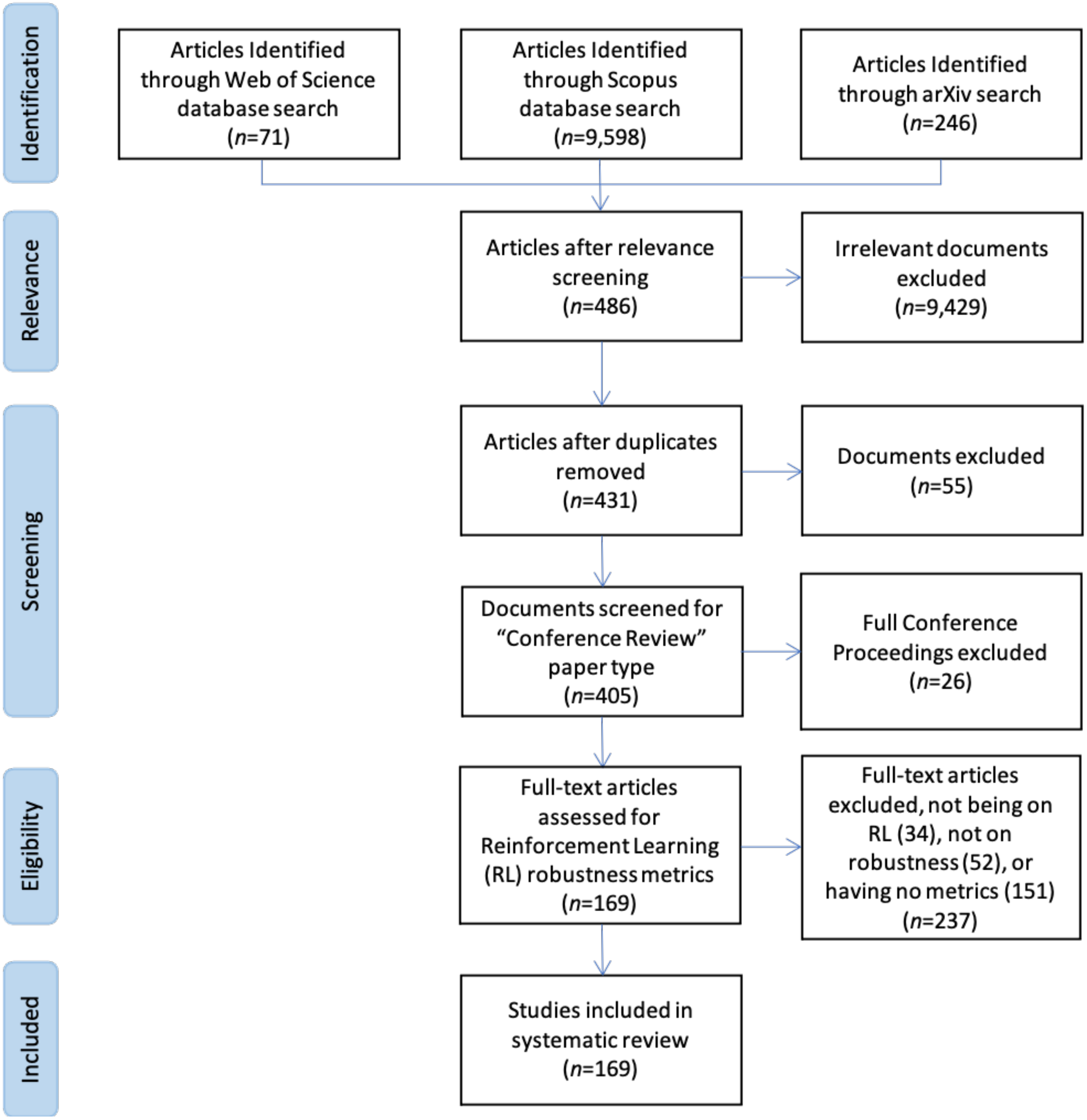

Identification
Articles Identified through Web of Science database search (n=71)
Articles Identified through Scopus database search (n=9,598)
Articles Identified through arXiv search (n=246)
Relevance
Articles after relevance screening (n=486)
Irrelevant documents excluded (n=9,429)
Screening
Articles after duplicates removed (n=431)
Documents excluded (n=55)
Documents screened for “Conference Review” paper type (n=405)
Full Conference Proceedings excluded (n=26)
Eligibility
Full-text articles assessed for Reinforcement Learning (RL) robustness metrics (n=169)
Full-text articles excluded, not being on RL (34), not on robustness (52), or having no metrics (151) (n=237)
Included
Studies included in systematic review (n=169)

Figure 11. The PRISMA Flow Diagram for Robustness

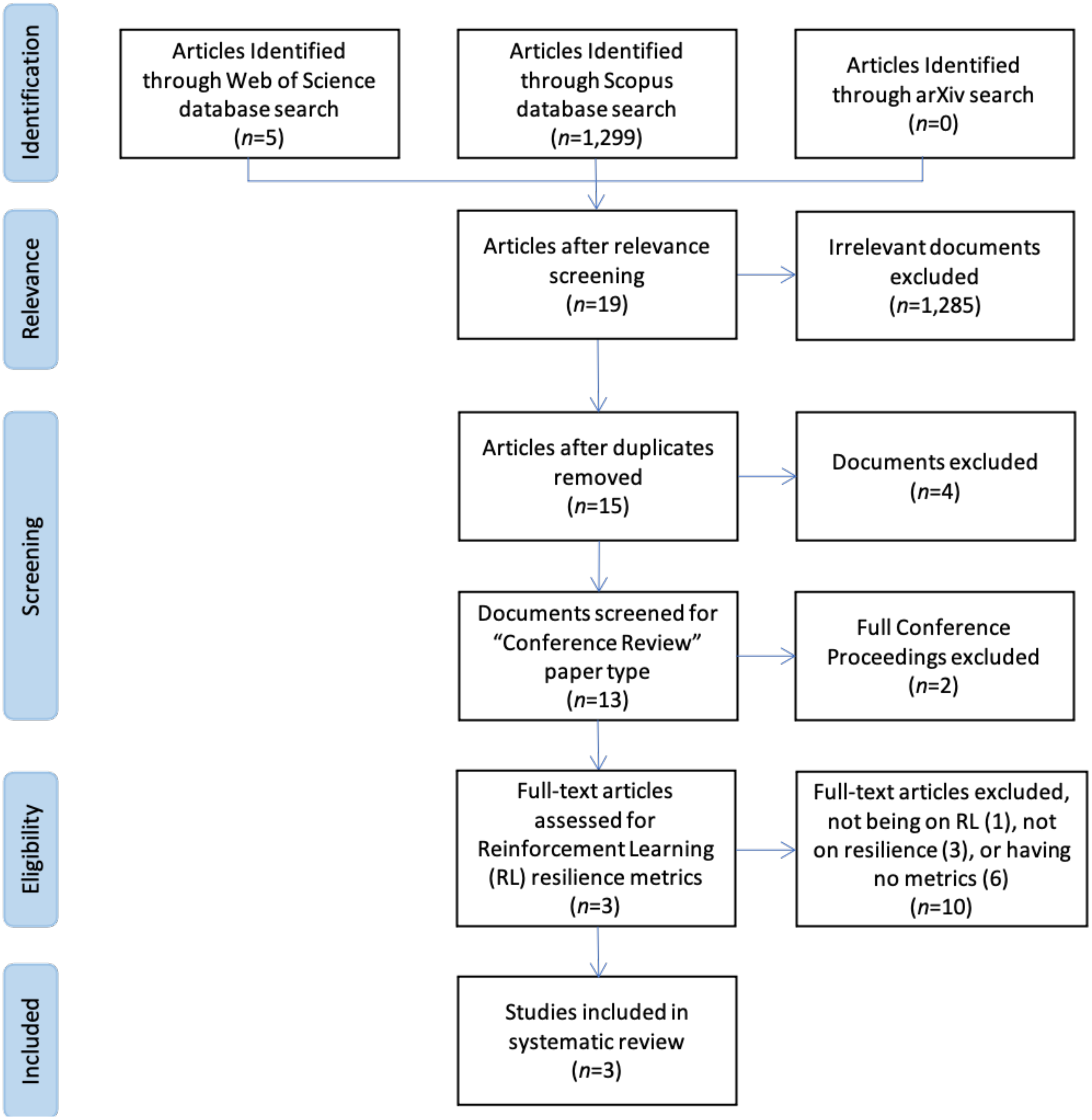
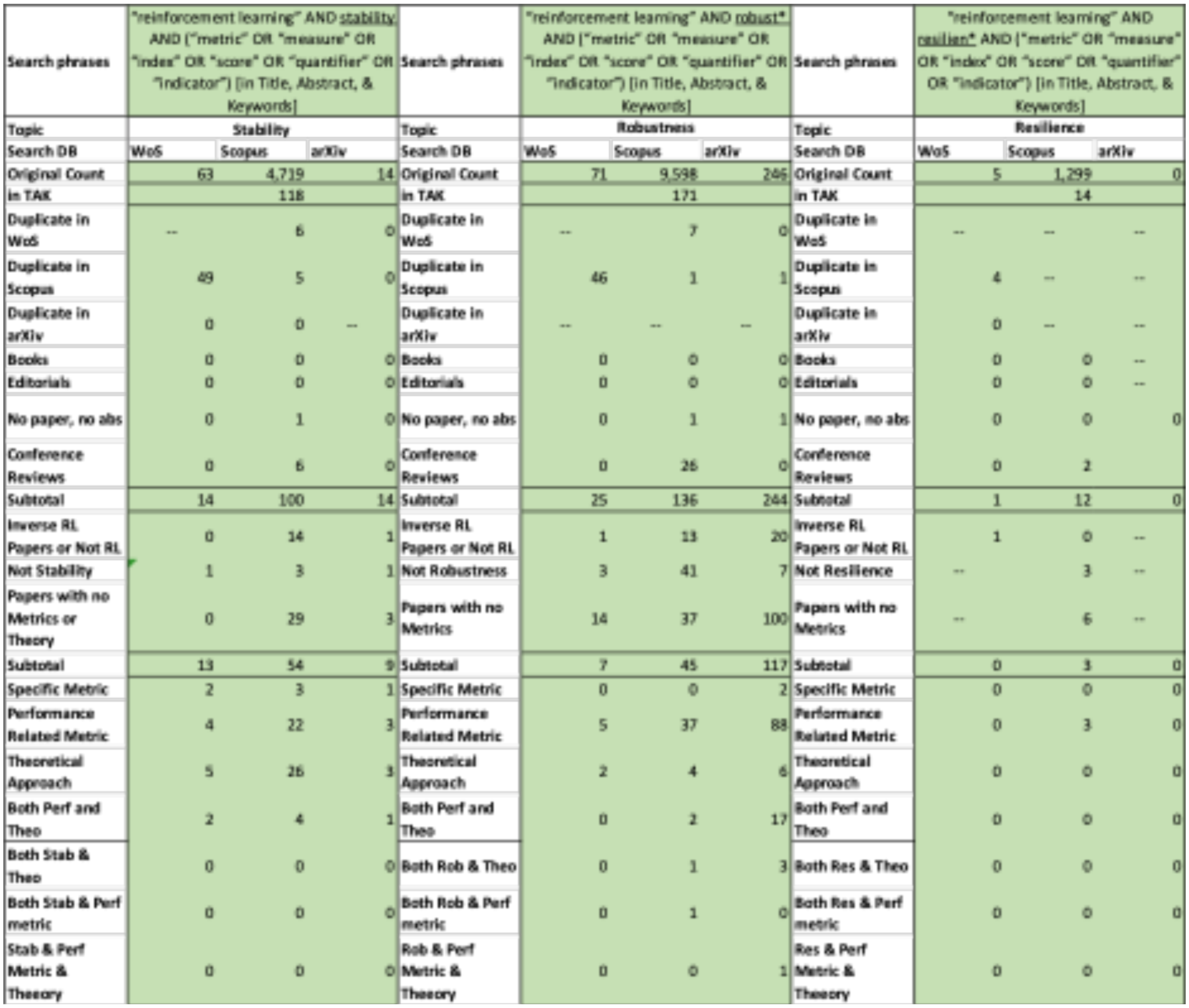

Identification
Articles Identified through Web of Science database search (n=5)
Articles Identified through Scopus database search (n=1,299)
Articles Identified through arXiv search (n=0)
Relevance
Articles after relevance screening (n=19)
Irrelevant documents excluded (n=1,285)
Screening
Articles after duplicates removed (n=15)
Documents excluded (n=4)
Documents screened for "Conference Review" paper type (n=13)
Full Conference Proceedings excluded (n=2)
Eligibility
Full-text articles assessed for Reinforcement Learning (RL) resilience metrics (n=3)
Full-text articles excluded, not being on RL (1), not on resilience (3), or having no metrics (6) (n=10)
Included
Studies included in systematic review (n=3)

Figure 12. The PRISMA Flow Diagram for Resilience

Table 13. Data Reduction Summary



Search phrases	"reinforcement learning" AND stability AND ("metric" OR "measure" OR "index" OR "score" OR "quantifier" OR "indicator") [in Title, Abstract, & Keywords]			Search phrases	"reinforcement learning" AND robust* AND ("metric" OR "measure" OR "index" OR "score" OR "quantifier" OR "indicator") [in Title, Abstract, & Keywords]			Search phrases	"reinforcement learning" AND resilien* AND ("metric" OR "measure" OR "index" OR "score" OR "quantifier" OR "indicator") [in Title, Abstract, & Keywords]		
Topic	Stability			Topic	Robustness			Topic	Resilience		
Search DB	WoS	Scopus	arXiv	Search DB	WoS	Scopus	arXiv	Search DB	WoS	Scopus	arXiv
Original Count	63	4,719	14	Original Count	71	9,598	246	Original Count	5	1,299	0
in TAK	118			in TAK	171			in TAK	14		
Duplicate in WoS	--	6	0	Duplicate in WoS	--	7	0	Duplicate in WoS	--	--	--
Duplicate in Scopus	49	5	0	Duplicate in Scopus	46	1	1	Duplicate in Scopus	4	--	--
Duplicate in arXiv	0	0	--	Duplicate in arXiv	--	--	--	Duplicate in arXiv	0	--	--
Books	0	0	0	Books	0	0	0	Books	0	0	--
Editorials	0	0	0	Editorials	0	0	0	Editorials	0	0	--
No paper, no abs	0	1	0	No paper, no abs	0	1	1	No paper, no abs	0	0	0
Conference Reviews	0	6	0	Conference Reviews	0	26	0	Conference Reviews	0	2	
Subtotal	14	100	14	Subtotal	25	136	244	Subtotal	1	12	0
Inverse RL Papers or Not RL	0	14	1	Inverse RL Papers or Not RL	1	13	20	Inverse RL Papers or Not RL	1	0	--
Not Stability	1	3	1	Not Robustness	3	41	7	Not Resilience	--	3	--
Papers with no Metrics or Theory	0	29	3	Papers with no Metrics	14	37	100	Papers with no Metrics	--	6	--
Subtotal	13	54	9	Subtotal	7	45	117	Subtotal	0	3	0
Specific Metric	2	3	1	Specific Metric	0	0	2	Specific Metric	0	0	0
Performance Related Metric	4	22	3	Performance Related Metric	5	37	88	Performance Related Metric	0	3	0
Theoretical Approach	5	26	3	Theoretical Approach	2	4	6	Theoretical Approach	0	0	0
Both Perf and Theo	2	4	1	Both Perf and Theo	0	2	17	Both Perf and Theo	0	0	0
Both Stab & Theo	0	0	0	Both Rob & Theo	0	1	3	Both Res & Theo	0	0	0
Both Stab & Perf metric	0	0	0	Both Rob & Perf metric	0	1	0	Both Res & Perf metric	0	0	0
Stab & Perf Metric & Theeory	0	0	0	Rob & Perf Metric & Theeory	0	0	1	Res & Perf Metric & Theeory	0	0	0

Table 13. The PRISMA Checklist [251]

Section/topic	#	Checklist item	Reported in Section
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	Introduction
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	Introduction
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	Introduction
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	Introduction
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	Introduction
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	Methods and Supporting Information (SI)
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	Methods and SI
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	SI
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	Methods

Table 13. The PRISMA Checklist [251]

Section/topic	#	Checklist item	Reported in Section
METHODS (Continued)			
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	Methods
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	Methods and SI
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	Methods
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	Methods
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	Methods
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	Methods
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	Results and SI
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	Results
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	Results
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	Results
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	Results
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	Results

Table 13. The PRISMA Checklist [251]

Section/topic	#	Checklist item	Reported in Section
RESULTS (Continued)			
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	Results
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	Methods
Section/topic	**#**	**Checklist item**	**Reported in Section**
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	Discussion
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	Discussion
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	Discussion
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	Acknowledgements

6. Acknowledgements

The author would like to acknowledge Rama Vasudevan, PhD of Oak Ridge National Laboratory for intellectual discussions and collaborative research on reinforcement learning. The author would also like to thank Nathan Martindale for assistance making the decision tree more readable and thus usable.

This work was funded initially by the AI Initiative at the Oak Ridge National Laboratory and subsequently funded by the US Department of Energy, National Nuclear Security Administration's Office of Defense Nuclear Nonproliferation Research and Development (NA-22).

Acronyms

A2C	Advantage Actor Critic
AODV	Ad-hoc On-demand Distance Vector
ARVI	Anytime Reduced Value Iteration
ATTN	Active Tracking Target Network
AUC	Area Under the receiver operating characteristic (ROC) Curve (AUC)
BMI	Brain Machine Interface
CLAC	Lyapunov-based Actor Critic
CPPO	Constraint-controlled Proximal Policy Optimization
DB	Database
DDPG	Deep Deterministic Policy Gradient
DOE	Department of Energy
DrAC	Data-regulated Actor-Critic
HVAC	Heating, Ventilation, and Air Conditioning
JSD	Jensen-Shannon Divergence
Max	Maximum
Min	Minimum
MOOSE	Model-based Offline policy Search with Ensembles
NESM	Normalized Energy Stability Margin
Perf	Performance
PPO	Proximal Policy Optimization
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RL	Reinforcement Learning
Res	Research
Rob	Robustness
ROC	Receiver Operating Characteristic

sec	seconds
Stab	Stability
SVD	Singular Value Decomposition
TAK	Title, Abstract and Keywords
Theo	Theoretical
US	United States
WoS	Web of Science

References

Resilience

[1] Behzadan, V., and Munir, A. 2018. Adversarial exploitation of emergent dynamics in smart cities. *Proceedings of the 2018 IEEE International Smart Cities Conference*, doi 10.1109/ISC2.2018.8656789.

[2] Enjalbert, S., and Vanderhaegen, F. 2017. A hybrid reinforced learning system to estimate resilience indicators. *Engineering Applications of Artificial Intelligence* 64:295-301.

[3] Bunyakitanon, M., Vasilakos, X., Nejabati, R., and Simeonidou D. 2020. End-to-End Performance-Based Autonomous VNF Placement with Adopted Reinforcement Learning. *IEEE Transactions on Cognitive Communications and Networking* 6(2):534-547. doi 10.1109/TCCN.2020.2988486.

Stability

[4] Dong, Zhe, X. Huang, Y. Dong, and Z. Zhang. 2020. Multilayer perception based reinforcement learning supervisory control of energy systems with application to a nuclear steam supply system. *Journal of Applied Energy* 259, Feb. 2020. doi 10.1016/j.apenergy.2019.114193.

[5] Wen, Guoxing, C.L. Philip Chen, Shuzhi Sam Ge, Hongli Yang, and Xiaoguang Liu. 2019. Optimized adaptive nonlinear tracking control using actor-critic reinforcement learning strategy. *IEEE Transactions on Industrial Informatics* 15(9):4969-4977.

[6] Muneeswari, B., and M.S.K. Manikandan. 2019. Energy efficient clustering and secure routing using reinforcement learning for three-dimensional mobile ad hoc networks. *IET Communications* 13(12):1828-1839.

[7] de Oliveira, Bernardo A.G., Carlos A.P. da S. Martins, Flavia Magalhaes, Luis Fabricio, and W. Goes. 2019. Difference based metrics for deep reinforcement learning algorithms. *IEEE Access* 7:159141-159149.

[8] Zhang, Kaiqing, Alec Koppel, Hao Zhu, and Tamer Bas. 2019. Policy search in infinite-horizon discounted reinforcement learning: advances through connections to non-convex optimization. *Proceedings: 53rd Annual Conference on Information Sciences and Systems (CISS)*, Baltimore, MD, Mar 20-22.

[9] Du, Zhijiang, Wei Wang, Zhiyuan Yan, Wei Dong, and Weidong Wang. 2017. Variable admittance control based on fuzzy reinforcement learning for minimally invasive surgery manipulator, *Sensors* 17(4).

[10] Jiang, He, Huaguang Zhang, Yanhong Luo, and Junyi Wang. 2016. Optimal tracking control for completely unknown nonlinear discrete-time Markov jump systems using data-based reinforcement learning method. *Neurocomputing* 194:176-182.

[11] Abdallah, Sherief. 2009. Why global performance is a poor metric for verifying convergence of multi-agent learning. arXiv:0904.2320v1 [cs.MA] 15 April 2009.

[12] Talele, Nihar, and Katie Byl. 2019. Mesh-based tools to analyze deep reinforcement learning policies for underactuated biped locomotion. arXiv:1903.12311v2 [cs.RO] 1 November 2019.

[13] Tuan, Yi-Lin, Jinzhi Zhang, Yujia Li, and Hung-yi Lee. 2018. Proximal policy optimization and its dynamic version for sequence generation. arXiv:1808.07982v1 [cs.CL] 24 August 2018.

[14] Serhani, Abdellatif, Najib Naja, and Abdellah Jamali. 2020. AQ-Routing: mobility-, stability-aware adaptive routing protocol for data routing in MANET–IoT systems. *Cluster Computing* 23:13–27. doi 10.1007/s10586-019-02937-x.

[15] Dong, Zhe, Xiaojin Huang, Yujie Dong, and Zuoyi Zhang. 2020. Multilayer perception based reinforcement learning supervisory control of energy systems with application to a nuclear steam supply system. *Applied Energy* 259 (2020) 114193. doi 10.1016/j.apenergy.2019.114193.

[16] Zhang, Huaguang, Kun Zhang, Yuliang Cai, and Jian Han. 2019. Adaptive fuzzy fault-tolerant tracking control for partially unknown systems with actuator faults

via integral reinforcement learning method. *IEEE Transactions on Fuzzy Systems* 27(10). doi 10.1109/TFUZZ.2019.2893211.

[17] Cohen, Daniel, Scott M. Jordan, and W. Bruce Croft. 2019. Learning a Better Negative Sampling Policy with Deep Neural Networks for Search. In the *2019 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '19)*, October 2–5, 2019, Santa Clara, CA, USA. ACM, New York, NY, USA. doi 10.1145/3341981.3344220.

[18] Mu, Chaoxu, Qian Zhao, Zhongke Gao, and Changyin Sun. 2019. Q-learning solution for optimal consensus control of discrete-time multiagent systems using reinforcement learning. *Journal of the Franklin Institute* 356:6946–6967. doi 10.1016/j.jfranklin.2019.06.0070016-0032.

[19] Tang, Xiaocheng, Zhiwei (Tony) Qin, Fan Zhang, Zhaodong Wang, Zhe Xu, Yintai Ma, Hongtu Zhu, and Jieping Ye. 2019. A deep value-network based approach for multi-driver order dispatching. *KDD 19*, August 4–8, 2019, Anchorage, AK, USA. doi 10.1145/3292500.3330724.

[20] Abouheaf, Mohammed, and Wail Gueaieb. 2019. Model-free adaptive control approach using integral reinforcement learning. *IEEE International Symposium on Robotic and Sensors Environments – Proceedings.*

[21] Seo, Donghyeon, Harin Kim, and Donghan Kim. 2019. Push recovery control for humanoid robot using reinforcement learning. *Third IEEE International Conference on Robotic Computing (IRC)*. doi 10.1109/IRC.2019.00102.

[22] Lv, Y., X. Ren, and J. Na. 2019. Online Nash-optimization tracking control of multi-motor driven load system with simplified RL scheme. *ISA Transactions.* doi 10.1016/j.isatra.2019.08.025.

[23] Tang, Li, Yan-Jun Liu, and C. L. Philip Chen. 2018. Adaptive critic design for pure-feedback discrete-time MIMO systems preceded by unknown backlashlike hysteresis. *IEEE Transactions on Neural Networks and Learning Systems.* 29(11), November. doi 10.1109/TNNLS.2018.2805689.

[24] Mertikopoulos, Panayotis, and William H. Sandholm. 2018. Riemannian game dynamics. *Journal of Economic Theory* (177):315-364. doi 10.1016/j.jet.2018.06.002.

[25] Liu, D., and G.-H. Yang. 2018. Model-free adaptive control design for nonlinear discrete-time processes with reinforcement learning techniques. *International*

Journal of Systems Science 49(11):2298-2308. doi 10.1080/00207721.2018.1498557.

[26] Bentaleb, Abdelhak, Ali C. Begen, and Roger Zimmermann. 2018. ORL-SDN: Online reinforcement learning for SDN-enabled HTTP adaptive streaming. *ACM Trans. Multimedia Comput. Commun. Appl.* 14(3) Article 71 (August), 28 pages. doi 10.1145/3219752.

[27] Hu Y., and B. Si. 2018. A reinforcement learning neural network for robotic manipulator control. *Neural Computation* 30(7):1983-2004. doi 10.1162/neco_a_01079.

[28] Mei, Y., G.-Z. Tan, Z.-T. Liu, and H. Wu. 2018. Chaotic time series prediction based on brain emotional learning model and self-adaptive genetic algorithm. *Acta Physica Sinica* 67(8). doi 10.7498/aps.67.20172104.

[29] Hong, Zhang-Wei, Shih-Yang Su, Tzu-Yun Shann, Yi-Hsiang Chang, and Chun-Yi Lee. 2018. A deep policy inference Q-network for multi-agent systems. In *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018)*, Stockholm, Sweden, July 10–15, 2018.

[30] Xiong, Yanhai, Haipeng Chen, Mengchen Zhao, and Bo An. 2018. HogRider: Champion agent of Microsoft Malmo collaborative AI challenge. *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pp. 4767-4774.

[31] Wu, Weiguo, and Liyang Gao. 2017. Posture self-stabilizer of a biped robot based on training platform and reinforcement learning. *Robotics and Autonomous Systems* 98:42-55. doi 10.1016/j.robot.2017.09.001.

[32] Boushaba, Mustapha, Abdelhakim Hafid, and Michel Gendreau. 2017. Node stability-based routing in wireless mesh networks. *Journal of Network and Computer Applications* 93:1–12. doi 10.1016/j.jnca.2017.02.010.

[33] Chasparis, Georgios C. 2017. Stochastic stability analysis of perturbed learning Automata with constant step-size in strategic-form games. *American Control Conference (ACC)*, Seattle, WA, 2017, pp. 4607-4612.

[34] Prins, Noeline W., Justin C. Sanchez and Abhishek Prasad. 2017. Feedback for reinforcement learning based brain-machine interfaces using confidence metrics. *Journal of Neural Engineering* 14 036016. doi 10.1088/1741-2552/aa6317.

[35] Song, Ruizhuo, Frank L. Lewis, and Qinglai Wei. 2017. Off-policy integral reinforcement learning method to solve nonlinear continuous-time multiplayer

nonzero-sum games. *IEEE Transactions on Neural Networks and Learning Systems* 28(3). doi 10.1109/TNNLS.2016.2582849.

[36] Yousefian, Reza, Amireza Sahami, and Sukumar Kamalasadan. 2017. Hybrid transient energy function-based real-time optimal wide-area damping controller. *IEEE Transactions on Industry Applications* 53(2). doi 10.1109/TIA.2016.2624264.

[37] Tatari, Farzaneh, Mohammad-Bagher Naghibi-Sistani, and Kyriakos G. Vamvoudakis. 2015. Distributed learning algorithm for non-linear differential graphical games. *Transactions of the Institute of Measurement and Control* 1–10.

[38] Lu, C., J. Huang, and J. Gong. 2016. Reinforcement learning for ramp control: an analysis of learning parameters. *Promet – Traffic & Transportation* 28(4):371-381.

[39] Vamvoudakis, Kyriakos G. 2016. Optimal trajectory output tracking control with a Q-learning algorithm. *Proceedings of the American Control Conference*, pp. 5752-5757. doi 10.1109/ACC.2016.7526571.

[40] Rêgo, Patrícia Helena Moraes, João Viana da Fonseca Neto, and Ernesto M. Ferreira. 2013. Convergence of the standard RLS method and UDUT factorisation of covariance matrix for solving the algebraic Riccati equation of the DLQR via heuristic approximate dynamic programming. *International Journal of Systems Science*. doi 10.1080/00207721.2013.844283.

[41] Liu, Derong, and Qinglai Wei. 2014. Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems. *IEEE Transactions on Neural Networks and Learning Systems* 25(3).

[42] Alharbi, Amal, Abdullah Al-Dhalaan, and Miznah Al-Rodhaan. 2014. Q-routing in cognitive packet network routing protocol for MANETs. In *Proceedings of the International Conference on Neural Computation Theory and Applications (NCTA-2014)*, pp. 234-243. doi 10.5220/0005082902340243.

[43] Yousefian, Reza, and Sukumar Kamalasadan. 2014. An approach for real-time tuning of cost functions in optimal system-centric wide area controller based on adaptive critic design. *IEEE Power and Energy Society General Meeting*, October 2014. doi 10.1109/PESGM.2014.6939224.

[44] Dong, Bo, and Yuanchun Li. 2013. Decentralized reinforcement learning robust optimal tracking control for time varying constrained reconfigurable modular

robot based on ACI and Q-function. *Mathematical Problems in Engineering* 2013, Article 387817, 16 pages. doi 10.1155/2013/387817.

[45] Teixeira, Carlos, Lino Costa, and Cristina Santos. 2014. Biped locomotion - improvement and adaptation. *IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*, May 14-15, Espinho, Portugal.

[46] Luy, N.T., N.T. Thanh, H.M. Tri. 2014. Reinforcement learning-based intelligent tracking control for wheeled mobile robot. *Transactions of the Institute of Measurement and Control* 36(7):868–877. doi 10.1177/0142331213509828.

[47] Hager, Louw vS., Kenneth R. Uren, and George van Schoor. 2014. Series-Parallel Approach to On-line Observer based Neural Control of a Helicopter System. *Proceedings of the 19th World Congress the International Federation of Automatic Control*, Cape Town, South Africa. August 24-29, 2014.

[48] Wei, Qinglai, and Derong Liu. 2014. Data-driven neuro-optimal temperature control of water-gas shift reaction using stable iterative adaptive dynamic programming. *IEEE Transactions on Industrial Electronics* 61(11).

[49] Zhao, D., B. Wang, and D. Liu. 2013. A supervised Actor-Critic approach for adaptive cruise control. *Soft Comput* 17:2089–2099. doi 10.1007/s00500-013-1110-y.

[50] Kashki, M., M.A. Abido, and Y.L. Abdel-Magid. 2013. Power system dynamic stability enhancement using optimum design of PSS and static phase shifter based stabilizer. *Arab J Sci Eng* 38:637–650. doi 10.1007/s13369-012-0325-z.

[51] Li, Cai, Robert Lowe, and Tom Ziemke. 2013. Humanoids learning to walk: a natural CPG-actor-critic architecture. *Frontiers in Neurorobotics* 7, Article 5. doi 10.3389/fnbot.2013.00005.

[52] Vamvoudakis, K.G., F.L. Lewis, and G.R. Hudas. 2012. Multi-agent differential graphical games: Online adaptive learning solution for synchronization with optimality. *Automatica* 48(8):1598-1611. doi 10.1016/j.automatica.2012.05.074.

[53] Moradi, P., M.E. Shiri, A.A. Rad, A. Khadivi, and M. Hasler. 2012. Automatic skill acquisition in reinforcement learning using graph centrality measures. *Intelligent Data Analysis* 16(1):113-135. doi 10.3233/IDA-2011-0513.

[54] Bhasin, Shubhendu, Nitin Sharma, Parag Patre, and Warren Dixon. 2011. Asymptotic tracking by a reinforcement learning-based adaptive critic controller. *J Control Theory Appl* 9(3):400-409. doi 10.1007/s11768-011-0170-8.

[55] Hafner, Roland, and Martin Riedmiller. 2011. Reinforcement learning in feedback control: Challenges and benchmarks from technical process control. *Machine Learning* 84:137-169. doi 10.1007/s10994-011-5235-x.

[56] Luy, Nguyen Tan. 2012. Reinforcement learning-based tracking control for wheeled mobile robot. *IEEE International Conference on Systems, Man, and Cybernetics*, October 14-17, 2012, Seoul, Korea.

[57] Shih, Peter, Brian C. Kaul, Sarangapani Jagannathan, and James A. Drallmeier. 2009. Reinforcement-learning-based output-feedback control of nonstrict nonlinear discrete-time systems with application to engine emission control. *IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics* 39(5).

[58] Boada, M.J.L., B.L. Boada, A.G. Babé, J.A. Calvo Ramos, and V.D. López. 2009. Active roll control using reinforcement learning for a single unit heavy vehicle. International *Journal of Heavy Vehicle Systems* 16(4):412-430. doi 10.1504/IJHVS.2009.027413.

[59] Guo, L., Y. Zhang, and J.-L. Hu. 2007. Adaptive HVDC supplementary (lamping controller based on reinforcement learning. *Electric Power Automation Equipment* 27(10):87-91.

[60] Lin, C.-K. 2003. A reinforcement learning adaptive fuzzy controller for robots. *Fuzzy Sets and Systems* 137(3):339-352. doi 10.1016/S0165-0114(02)00299-3.

[61] Jagannathan, S. 2002. Adaptive critic neural network-based controller for nonlinear systems. *Proceedings of the 2002 IEEE International Symposium on Intelligent Control*, Vancouver, Canada, October 27-30, 2002.

[62] Kaygisiz, Burak H., Aydan M. Erkmen, and Ismet Erkmen. 2002. Smoothing stability roughness of fractal boundaries using reinforcement learning. *IFAC Proceedings Volumes* 15(1):481-485.

[63] Li, J.N., J.L. Ding, T.Y. Chai, and F.L. Lewis. 2020. Nonzero-Sum Game Reinforcement Learning for Performance Optimization in Large-Scale Industrial Processes. *IEEE Transactions on Cybernetics* 50(9):4132-4145. Doi 10.1109/TCYB.2019.2950262.

[64] Zhang, K., H.G. Zhang, Y.L. Cai, and R. Su. 2020. Parallel Optimal Tracking Control Schemes for Mode-Dependent Control of Coupled Markov Jump Systems via Integral RL Method. *IEEE Transactions on Automation Science and Engineering* 17(3):1332-1342. Doi 10.1109/TASE.2019.2948431.

[65] Zhang, Qian, Kui Wu, and Yang Shi. 2020. Route Planning and Power Management for PHEVs With Reinforcement Learning. *IEEE Transactions on Vehicular Technology* 69(5):4751-4762. doi 10.1109/TVT.2020.2979623.

[66] Serhani, Abdellatif, Najib Naja, and Abdellah Jamali. 2020. AQ-Routing: mobility-, stability-aware adaptive routing protocol for data routing in MANET-IoT systems. *Cluster Computing* 23(1):13-27. Doi 10.1007/s10586-019-02937-x.

[67] Dong, Zhe, Xiaojin Huang, Yujie Dong, and Zuoyi Zhang. 2020. Multilayer perception based reinforcement learning supervisory control of energy systems with application to a nuclear steam supply system. *Applied Energy* 259. doi 10.1016/j.apenergy.2019.114193.

[68] Wang, Qingling. 2020. Integral Reinforcement Learning Control for a Class of High-Order Multivariable Nonlinear Dynamics with Unknown Control Coefficients. *IEEE Access* 8:86223-86229. Doi 10.1109/ACCESS.2020.2993265.

[69] Zhang J., Z. Peng, J. Hu, Y. Zhao, R. Luo, B.K. Ghosh. 2020. Internal reinforcement adaptive dynamic programming for optimal containment control of unknown continuous-time multi-agent systems. *Neurocomputing* 413:85-95. doi 10.1016/j.neucom.2020.06.106.

[70] Mitriakov, A., P. Papadakis, S.M. Nguyen, and S. Garlatti. 2020. Staircase traversal via reinforcement learning for active reconfiguration of assistive robots. *Proceedings IEEE International Conference on Fuzzy Systems*. doi 10.1109/FUZZ48607.2020.9177581.

[71] Lv, Y., X. Ren, and J. Na. 2020. Online Nash-optimization tracking control of multi-motor driven load system with simplified RL scheme. *ISA Transactions* 98:251-262. doi 10.1016/j.isatra.2019.08.025.

[72] Thornton, C.E., M.A. Kozy, R.M. Buehrer, A.F. Martone, and K.D. Sherbondy. 2020. Deep Reinforcement Learning Control for Radar Detection and Tracking in Congested Spectral Environments. *IEEE Transactions on Cognitive Communications and Networking*. doi 10.1109/TCCN.2020.3019605.

[73] Prasanna, John Deva, John Aravindhar, P. Sivasankar, and K. Perumal. 2020. Reinforcement learning based virtual backbone construction in manet using connected dominating sets. *Journal of Critical Reviews* 7(9):146-152. doi 10.31838/jcr.07.09.28.

[74] Pongfai, J., X. Su, H. Zhang, and W. Assawinchaichote. 2020. PID controller autotuning design by a deterministic Q-SLP algorithm. *IEEE Access* 8:50010-50021. doi 10.1109/ACCESS.2020.2979810.

[75] Hoppe, Sabrina, and Marc Toussaint. 2020. Q graph-bounded Q-learning: Stabilizing Model-Free O-Policy Deep Reinforcement Learning. arXiv:2007.07582v1 [cs.LG] 15 Jul 2020.

[76] Osinenko, Pavel, Lukas Beckenbach, Thomas Göhrt, and Stefan Streif. 2020. A reinforcement learning method with closed-loop stability guarantee. arXiv:2006.14034v1 [math.OC] 24 Jun 2020.

[77] Han, Minghao, Lixian Zhang, Jun Wang, and Wei Pan. 2020. Actor-Critic Reinforcement Learning for Control with Stability Guarantee. arXiv:2004.14288v3 [cs.RO] 15 Jul 2020.

[78] Khader, Shahbaz A., Hang Yin, Pietro Falco and Danica Kragic. 2020. Stability-Guaranteed Reinforcement Learning for Contact-rich Manipulation. arXiv:2004.10886v2 [cs.RO] 27 Sep 2020.

[79] Han, Minghao, Yuan Tian, Lixian Zhang, Jun Wang, and Wei Pan. 2020. H infinity Model-free Reinforcement Learning with Robust Stability Guarantee. arXiv:1911.02875v3 [cs.LG] 25 Jul 2020.

[80] Tessler, Chen, Nadav Merlis, and Shie Mannor. Stabilizing Deep Reinforcement Learning with Conservative Updates. arXiv:1910.01062v2 [cs.LG] 9 Feb 2020.

Robustness

[81] Abuzainab, Nof, Tugba Erpek, Kemal Davaslioglu, Yalin E. Sagduyu, Yi Shi, Sharon J. Mackey, Mitesh Patel, Frank Panettieri, Muhammad A. Qureshi, Volkan Isler, and Aylin Yener. 2019. QoS and jamming-aware wireless networking using deep reinforcement learning. arXiv:1910.05766v1 [cs.NI] 13 October 2019.

[82] Ahn, Michael. 2019. ROBEL: Robotics benchmarks for learning with low-cost robots. arXiv:1909.11639v3 [cs.RO] 16 Dec 2019.

[83] Dhiman, Vikas, Shurjo Banerjee, Brent Griffin, Jeffrey M Siskind, and Jason J Corso. 2019. A critical investigation of deep reinforcement learning for navigation. arXiv:1802.02274v2 [cs.RO] 4 Jan 2019.

[84] Naderializadeh, Navid, Jaroslaw Sydir, Meryem Simsek, Hosein Nikopour, and Shilpa Talwar. 2019. When multiple agents learn to schedule: a distributed radio resource management framework. arXiv:1906.08792v1 [cs.LG] 20 Jun 2019.

[85] Nguyen, Khanh, Hal Daumé III, and Jordan Boyd-Graber. 2017. Reinforcement learning for bandit neural machine translation with simulated human feedback. arXiv:1707.07402v4 [cs.CL] 11 Nov 2017.

[86] Talele, Nihar, and Katie Byl. 2019. Mesh-based tools to analyze deep reinforcement learning policies for underactuated biped locomotion. arXiv:1903.12311v2 [cs.RO] 1 Nov 2019.

[87] Turchetta, Matteo, Andreas Krause, and Sebastian Trimpe. 2019. Robust model-free reinforcement learning with multi-objective Bayesian optimization. arXiv:1910.13399v1 [cs.RO] 29 Oct 2019.

[88] Yuan, Ye, and Kris Kitani. 2019. Ego-pose estimation and forecasting as real-time PD control. arXiv:1906.03173v2 [cs.CV] 4 Aug 2019.

[89] Muneeswari, B., and M.S.K. Manikandan. 2019. Energy efficient clustering and secure routing using reinforcement learning for three-dimensional mobile ad hoc networks. *IET Communications* 13(12):1828-1839. doi 10.1049/iet-com.2018.6150.

[90] Zhao, B., D. Wang, G. Shi, D.R. Liu, and Y.C. Li. 2018. Decentralized control for large-scale nonlinear systems with unknown mismatched interconnections via policy iteration. *IEEE Transactions on Systems Man Cybernetics-Systems* 48(10).

[91] Zhang, Y., Y. Yang, S.X. Ding, and L.L. Li. 2016. Optimal design of residual-driven dynamic compensator using iterative algorithms with guaranteed convergence. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 46(4). doi 10.1109/TSMC.2015.2450203.

[92] Tokic, M. 2010. Adaptive epsilon-greedy exploration in reinforcement learning based on value differences. *Lecture Notes in Artificial Intelligence*, *33rd Annual German Conference on Artificial Intelligence*, Karlsruhe, Germany, Sep 21-24, 2010.

[93] Xiong, Y., L. Guo, Y. Huang, and L. Chen. 2020. Intelligent thermal control strategy based on reinforcement learning for space telescope. *Journal of Thermophysics and Heat Transfer* 34(1):37-44.

[94] Isa-Jara, R.F., G.J. Meschino, and V.L. Ballarin. 2020. A comparative study of reinforcement learning algorithms applied to medical image registration. *IFMBE Proceedings*, pp. 281-289.

[95] Guo, F., X. Zhou, J. Liu, Y. Zhang, D. Li, and H. Zhou. 2019. A reinforcement learning decision model for online process parameters optimization from offline data in injection molding. *Applied Soft Computing Journal* 85. doi 10.1016/j.asoc.2019.105828.

[96] Li, S., C. He, M. Liu, Y. Wan, Y. Gu, J. Xie, S. Fu, and K. Lu. 2019. Design and implementation of aerial communication using directional antennas: Learning control in unknown communication environments. *IET Control Theory and Applications* 13(17):2906-2916. doi 10.1049/iet-cta.2018.6252.

[97] Tang, X., Z. Qin, F. Zhang, Z. Wang, Z. Xu, Y. Ma, H. Zhu, and J. Ye. 2019. A deep value-network based approach for multi-driver order dispatching. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1780-1790. doi 10.1145/3292500.3330724.

[98] Chowdhury, A., S.A. Raut, and H.S. Narman. 2019. DA-DRLS: Drift adaptive deep reinforcement learning based scheduling for IoT resource management. *Journal of Network and Computer Applications* 138:51-65. doi 10.1016/j.jnca.2019.04.010.

[99] Wang, X., C. Li, L. Yu, L. Han, X. Deng, E. Yang, and P. Ren. 2019. UAV first view landmark localization with active reinforcement learning. *Pattern Recognition Letters* 125:549-555. doi 10.1016/j.patrec.2019.03.011.

[100] Lütjens, B., M. Everett, and J.P. How. 2019. Safe reinforcement learning with model uncertainty estimates. *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 8662-8668. doi 10.1109/ICRA.2019.8793611.

[101] Balakrishnan, A., and J.V. Deshmukh. 2019. Structured reward functions using STL. *Proceedings of the 2019 22nd ACM International Conference on Hybrid Systems: Computation and Control*, pp. 270-271. doi 10.1145/3302504.3313355.

[102] Tang, C., W. Zhu, and X. Yu. 2019. Deep hierarchical strategy model for multi-source driven quantitative investment. *IEEE Access* 7:79331-79336. doi 10.1109/ACCESS.2019.2923267.

[103] Cheng, Q., X. Wang, Y. Niu, and L. Shen. 2019. Reusing source task knowledge via transfer approximator in reinforcement transfer learning. *Symmetry* 11(1). doi 10.3390/sym11010025.

[104] Jeon, Y.-S., H. Lee, and N. Lee. 2018. Robust MLSD for wideband SIMO systems with one-bit ADCs: reinforcement-learning approach. *Proceedings - 2018 IEEE*

International Conference on Communications Workshops, ICC Workshops, pp. 1-6. doi 10.1109/ICCW.2018.8403665.

[105] Yang, X., and H. He. 2018. Self-learning robust optimal control for continuous-time nonlinear systems with mismatched disturbances. *Neural Networks* 99:19-30. doi 10.1016/j.neunet.2017.11.022.

[106] Jiang, H., H. Zhang, Y. Cui, and G. Xiao. 2018. Robust control scheme for a class of uncertain nonlinear systems with completely unknown dynamics using data-driven reinforcement learning method. *Neurocomputing* 273:68-77. doi 10.1016/j.neucom.2017.07.058.

[107] Shayeghi, H., and A. Younesi. 2017. An online Q-learning based multi-agent LFC for a multi-area multi-source power system including distributed energy resources. *Iranian Journal of Electrical and Electronic Engineering* 13(4):385-398. doi 10.22068/IJEEE.13.4.385.

[108] Zhao, D., Y. Ma, Z. Jiang, and Z. Shi. 2017. Multiresolution airport detection via hierarchical reinforcement learning saliency model. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 10(6):2855-2866. doi 10.1109/JSTARS.2017.2669335.

[109] Tow, A.W., S. Shirazi, J. Leitner., N. Sünderhauf, M. Milford, and B. Upcroft. 2016. A robustness analysis of deep Q networks. *Australasian Conference on Robotics and Automation*, pp. 116-125.

[110] Hatami, E., and H. Salarieh. 2015. Adaptive critic-based neuro-fuzzy controller for dynamic position of ships. *Scientia Iranica* 22(1):272-280.

[111] Xiang, J., and Z. Chen. 2015. Adaptive traffic signal control of bottleneck subzone based on grey qualitative reinforcement learning algorithm. *ICPRAM 2015 - 4th International Conference on Pattern Recognition Applications and Methods, Proceedings*, 2:295-301.

[112] Padmanabhan, R., N. Meskin, and W.M. Haddad. 2015. Closed-loop control of anesthesia and mean arterial pressure using reinforcement learning. *Biomedical Signal Processing and Control* 22:54-64. doi 10.1016/j.bspc.2015.05.013.

[113] Bruno, R., A. Masaracchia, and A. Passarella. 2014. Robust adaptive modulation and coding (AMC) selection in LTE systems using reinforcement learning. *IEEE Vehicular Technology Conference*. doi 10.1109/VTCFall.2014.6966162.

[114] Jamali, N., P. Kormushev, S.R. Ahmadzadeh, and D.G. Caldwell. 2014. Covariance analysis as a measure of policy robustness. *OCEANS 2014 – Taipei.* doi 10.1109/OCEANS-TAIPEI.2014.6964339.

[115] Tati, S., S. Silvestri, T. He, and T.L. Porta. 2014. Robust network tomography in the presence of failures. *Proceedings - International Conference on Distributed Computing Systems*, pp. 481-492. doi 10.1109/ICDCS.2014.56.

[116] Luy, N.T., N.T. Thanh, and H.M. Tri. 2013. Reinforcement learning-based robust adaptive tracking control for multi-wheeled mobile robots synchronization with optimality. *Proceedings of the 2013 IEEE Workshop on Robotic Intelligence in Informationally Structured Space*, pp. 74-81. doi 10.1109/RiiSS.2013.6607932.

[117] Kashki, M., M.A. Abido, and Y.L. Abdel-Magid. 2013. Power system dynamic stability enhancement using optimum design of PSS and static phase shifter based stabilizer. *Arabian Journal for Science and Engineering* 38(3):637-650. doi 10.1007/s13369-012-0325-z.

[118] Lopes, M., T. Lang, M. Toussaint, and P.-Y. Oudeyer. 2012. Exploration in model-based reinforcement learning by empirically estimating learning progress. *Advances in Neural Information Processing Systems* 1:206-214.

[119] Llorente, M.S., and S.E. Guerrero. 2012. Increasing retrieval quality in conversational recommenders. *IEEE Transactions on Knowledge and Data Engineering* 24(10):1876-1888. doi 10.1109/TKDE.2011.116.

[120] Maes, F., L. Wehenkel, and D. Ernst. 2012. Learning to play K-armed bandit problems. *Proceedings of the 4th International Conference on Agents and Artificial Intelligence*, pp. 74-81.

[121] Bhasin, S., N. Sharma, P. Patre, and W. Dixon. 2011. Asymptotic tracking by a reinforcement learning-based adaptive critic controller. *Journal of Control Theory and Applications* 9(3):400-409. doi 10.1007/s11768-011-0170-8.

[122] Tjahjadi, A., S. Sendari, S. Mabu, and K. Hirasawa. 2010. Robustness analysis of genetic network programming with reinforcement learning. *Proceedings Joint 5th International Conference on Soft Computing and Intelligent Systems and 11th International Symposium on Advanced Intelligent Systems*, pp. 594-601.

[123] Kulkarni, S.A., and G.R. Rao. 2010. Vehicular ad hoc network mobility models applied for reinforcement learning routing algorithm. *Communications in Computer and Information Science*, pp. 230-240. doi 10.1007/978-3-642-14825-5_20.

[124] Molina, C., N.B. Yoma, F. Huenupán, C. Garretón, and J. Wuth. 2010. Maximum entropy-based reinforcement learning using a confidence measure in speech recognition for telephone speech. *IEEE Transactions on Audio, Speech and Language Processing* 18(5):1041-1052. doi 10.1109/TASL.2009.2032618.

[125] Luy, N.T., N.D. Thanh, N.T. Thanh, and N.T.P. Ha. 2010. Robust reinforcement learning-based tracking control for wheeled mobile robot. *The 2nd International Conference on Computer and Automation Engineering*, pp. 171-176. doi 10.1109/ICCAE.2010.5451973.

[126] Heidrich-Meisner, Verena, and Christian Igel. 2009. Hoeffding and Bernstein races for selecting policies in evolutionary direct policy search. *Proceedings of the 26th International Conference on Machine Learning*, pp. 401-408.

[127] Satoh, H. 2008. A nonlinear approach to robust routing based on reinforcement learning with state space compression and adaptive basis construction. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* 7:1733-1740. doi 10.1093/ietfec/e91-a.7.1733.

[128] Conn, K., and R.A. Peters. 2007. Reinforcement learning with a supervisor for a mobile robot in a real-world environment. *Proceedings of the 2007 IEEE International Symposium on Computational Intelligence in Robotics and Automation*, pp. 73-78. doi 10.1109/CIRA.2007.382878.

[129] Wang, X.-S., Y.-H. Cheng, and W. Sun. 2007. A proposal of adaptive PID controller based on reinforcement learning. *Journal of China University of Mining and Technology* 17(1):40-44. doi 10.1016/S1006-1266(07)60009-1.

[130] Leem, JoonBum, and Ha Young Kim. 2020. Action-specialized expert ensemble trading system with extended discrete action space using deep reinforcement learning. *PLOS One* 15(7). doi 10.1371/journal.pone.0236178.

[131] Xiong, Yan, Liang Guo, Yong Huang, and Liheng Chen. 2020. Intelligent Thermal Control Strategy Based on Reinforcement Learning for Space Telescope. *Journal of Thermophysics and Heat Transfer* 34(1):37-44. doi 10.2514/1.T5774.

[132] Balakrishnan, Anand, and Jyotirmoy V. Deshmukh. 2019. Structured Reward Functions using STL. *Proceedings of the 2019 22nd ACM International Conference on Hybrid Systems: Computation and Control* (HSCC '19), pp. 270-271. doi 10.1145/3302504.3313355.

[133] Chen, G., S. Yao, J. Ma, L. Pan, Y. Chen, P. Xu, J. Ji, and X. Chen. 2020. Distributed Non-Communicating Multi-Robot Collision Avoidance via Map-Based Deep Reinforcement Learning. *Sensors* 20(17). doi 10.3390/s20174836.

[134] Sun, C., X. Li, and C. Belta. 2020. Automata Guided Semi-Decentralized Multi-Agent Reinforcement Learning. *Proceedings of the American Control Conference*, pp. 3900-3905. doi 10.23919/ACC45564.2020.9147704.

[135] Wang, X., and X. Ye. 2020. Optimal Robust Control of Nonlinear Uncertain System via Off-Policy Integral Reinforcement Learning. *Proceedings of the Chinese Control Conference*, pp. 1928-1933. doi 10.23919/CCC50068.2020.9189626.

[136] Yan, Z., J. Ge, Y. Wu, L. Li, and T. Li. 2020. Automatic virtual network embedding: A deep reinforcement learning approach with graph convolutional networks. IEEE Journal on Selected Areas in Communications 38(6):1040-1057. doi 10.1109/JSAC.2020.2986662.

[137] Alhazmi, K., and S.M. Sarathy. 2020. Continuous Control of Complex Chemical Reaction Network with Reinforcement Learning. *Proceedings of the European Control Conference 2020*, ECC 2020, pp. 1066-1068.

[138] Ghasemkhani, A., A. Darvishi, I. Niazazari, A. Darvishi, H. Livani, and L. Yang. 2020. DeepGrid: Robust Deep Reinforcement Learning-based Contingency Management. *Proceedings of the 2020 IEEE Power and Energy Society Innovative Smart Grid Technologies Conference*, ISGT 2020. doi 10.1109/ISGT45199.2020.9087633.

[139] Pitti, A., M. Quoy, C. Lavandier, and S. Boucenna. 2020. Gated spiking neural network using Iterative Free-Energy Optimization and rank-order coding for structure learning in memory sequences (INFERNO GATE). *Neural Networks* 121:242-258. doi 10.1016/j.neunet.2019.09.023.

[140] Vecerik, Mel, Jean-Baptiste Regli, Oleg Sushkov, David Barker, Rugile Pevceviciute, Thomas Rothörl, Christopher Schuster, Raia Hadsell, Lourdes Agapito, and Jonathan Scholz. 2020. S3K: Self-Supervised Semantic Keypoints for Robotic Manipulation via Multi-View Consistency. arXiv:2009.14711.

[141] Jiao, Yusheng, Feng Ling, Sina Heydari, Nicolas Heess, Josh Merel, and Eva Kanso. 2020. Learning to swim in potential flow. arXiv:2009.14280.

[142] Almási, Péter, Róbert Moni, and Bálint Gyires-Tóth. 2020. Robust Reinforcement Learning-based Autonomous Driving Agent for Simulation and Real World. arXiv:2009.11212.

[143] Ding, Wenhao, Baiming Chen, Bo Li, Kim Ji Eun, and Ding Zhao. 2020. Multimodal Safety-Critical Scenarios Generation for Decision-Making Algorithms Evaluation. arXiv:2009.08311.

[144] Schamberg, Gabe, Marcus Badgeley, and Emery N. Brown. 2020. Controlling Level of Unconsciousness by Titrating Propofol with Deep Reinforcement Learning. arXiv:2008.12333.

[145] Pang, Bo, and Zhong-Ping Jiang. 2020. Robust Reinforcement Learning: A Case Study in Linear Quadratic Regulation. arXiv:2008.11592.

[146] Kobayashi, Taisuke, and Wendyam Eric Lionel Ilboudo. 2020. t-Soft Update of Target Network for Deep Reinforcement Learning. arXiv:2008.10861.

[147] Zavoli, Alessandro, and Lorenzo Federici. 2020. Reinforcement Learning for Low-Thrust Trajectory Design of Interplanetary Missions. arXiv:2008.08501.

[148] Limoyo, Oliver, Bryan Chan, Filip Marić, Brandon Wagstaff, Rupam Mahmood, and Jonathan Kelly. 2020. Heteroscedastic Uncertainty for Robust Generative Latent Dynamics. arXiv:2008.08157.

[149] Zhao, Wenshuai, Jorge Peña Queralta, Li Qingqing, and Tomi Westerlund. 2020. Towards Closing the Sim-to-Real Gap in Collaborative Multi-Robot Deep Reinforcement Learning. arXiv:2008.07875.

[150] Qu, Xinghua, Yew-Soon Ong, Abhishek Gupta, and Zhu Sun. 2020. Defending Adversarial Attacks without Adversarial Attacks in Deep Reinforcement Learning. arXiv:2008.06199.

[151] Swazinna, Phillip, Steffen Udluft, and Thomas Runkler. 2020. Overcoming Model Bias for Robust Offline Deep Reinforcement Learning. arXiv:2008.05533.

[152] Ahmed, Ibrahim, Hamed Khorasgani, and Gautam Biswas. 2020. Comparison of Model Predictive and Reinforcement Learning Methods for Fault Tolerant Control. arXiv:2008.04403.

[153] Kovač, Grgur, Adrien Laversanne-Finot, and Pierre-Yves Oudeyer. 2020. GRIMGEP: Learning Progress for Robust Goal Sampling in Visual Deep Reinforcement Learning. arXiv:2008.04388.

[154] Zhu, Jianxin Li, Hao Peng, Senzhang Wang, Philip S. Yu, and Lifang He. 2020. Adversarial Directed Graph Embedding. arXiv:2008.03667.

[155] Ma, Xiao, Siwei Chen, David Hsu, and Wee Sun Lee. 2020. Contrastive Variational Model-Based Reinforcement Learning for Complex Observations. arXiv:2008.02430.

[156] Oikarinen, Tuomas, Tsui-Wei Weng, and Luca Daniel. 2020. Robust Deep Reinforcement Learning through Adversarial Loss. arXiv:2008.01976.

[157] Vinitsky, Eugene, Yuqing Du, Kanaad Parvate, Kathy Jang, Pieter Abbeel, and Alexandre Bayen. 2020. Robust Reinforcement Learning using Adversarial Populations. arXiv:2008.01825.

[158] Park, Hwangpil, Ri Yu, Yoonsang Lee, Kyungho Lee, and Jehee Lee. 2020. Understanding the Stability of Deep Control Policies for Biped Locomotion. arXiv:2007.15242.

[159] Steverson, Kai, Jonathan Mullin, and Metin Ahiskali. 2020. Adversarial Robustness for Machine Learning Cyber Defenses Using Log Data. arXiv:2007.14983.

[160] Chen, Xinwei, Tong Wang, Barrett W. Thomas, and Marlin W. Ulmer. 2020. Same-Day Delivery with Fairness. arXiv:2007.09541.

[161] Chen, Xin, Yawen Duan, Zewei Chen, Hang Xu, Zihao Chen, Xiaodan Liang, Tong Zhang, and Zhenguo Li. 2020. CATCH: Context-based Meta Reinforcement Learning for Transferrable Architecture Search. arXiv:2007.09380.

[162] Zhang, Lin, Hao Xiong, Ou Ma, and Zhaokui Wang. 2020. Multi-robot Cooperative Object Transportation using Decentralized Deep Reinforcement Learning. arXiv:2007.09243.

[163] Tan, Kai Liang, Yasaman Esfandiari, Xian Yeow Lee, Aakanksha, and Soumik Sarkar. 2020. Robustifying Reinforcement Learning Agents via Action Space Adversarial Training. arXiv:2007.07176.

[164] Stooke, Adam, Joshua Achiam, and Pieter Abbeel. 2020. Responsive Safety in Reinforcement Learning by PID Lagrangian Methods. arXiv:2007.03964.

[165] Abe, Kenshi, and Yusuke Kaneko. 2020. Off-Policy Exploitability-Evaluation and Equilibrium-Learning in Two-Player Zero-Sum Markov Games. arXiv:2007.02141.

[166] Wang, Xiao, Saasha Nair, and Matthias Althoff. 2020. Falsification-Based Robust Adversarial Reinforcement Learning. arXiv:2007.00691.

[167] Lee, Heunchul, Maksym Girnyk, and Jaeseong Jeong. 2020. Deep reinforcement learning approach to MIMO precoding problem: Optimality and Robustness. arXiv:2006.16646.

[168] Xu, Duo, Mohit Agarwal, Ekansh Gupta, Faramarz Fekri, and Raghupathy Sivakumar. 2020. Accelerating Reinforcement Learning Agent with EEG-based Implicit Human Feedback. arXiv:2006.16498.

[169] Yu, Liang, Yi Sun, Zhanbo Xu, Chao Shen, Dong Yue, Tao Jiang, and Xiaohong Guan. 2020. Multi-Agent Deep Reinforcement Learning for HVAC Control in Commercial Buildings. arXiv:2006.14156.

[170] Gleave, Adam, Michael Dennis, Shane Legg, Stuart Russell, and Jan Leike. 2020. Quantifying Differences in Reward Functions. arXiv:2006.13900.

[171] Raileanu, Roberta, Max Goldstein, Denis Yarats, Ilya Kostrikov, and Rob Fergus. 2020. Automatic Data Augmentation for Generalization in Deep Reinforcement Learning. arXiv:2006.12862.

[172] Liu, Haotian, and Wenchuan Wu. 2020. Online Multi-agent Reinforcement Learning for Decentralized Inverter-based Volt-VAR Control. arXiv:2006.12841.

[173] Zou, Yayi, and Xiaoqi Lu. 2020. Gradient-EM Bayesian Meta-learning. arXiv:2006.11764.

[174] Panaganti, Kishan, and Dileep Kalathil. 2020. Model-Free Robust Reinforcement Learning with Linear Function Approximation. arXiv:2006.11608.

[175] Zhang, Amy, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. 2020. Learning Invariant Representations for Reinforcement Learning without Reconstruction. arXiv:2006.10742.

[176] Rahman, Arrasy, Niklas Hopner, Filippos Christianos, and Stefano V. Albrecht. 2020. Open Ad Hoc Teamwork using Graph-based Policy Learning. arXiv:2006.10412.

[177] Jeong, Heejin, Hamed Hassani, Manfred Morari, Daniel D. Lee, and George J. Pappas. 2020. Learning to Track Dynamic Targets in Partially Known Environments. arXiv:2006.10190.

[178] Ning, Kun-Peng, and Sheng-Jun Huang. 2020. Reinforcement Learning with Supervision from Noisy Demonstrations. arXiv:2006.07808.

[179] Dou, Yingtong, Guixiang Ma, Philip S. Yu, and Sihong Xie. 2020. Robust Spammer Detection by Nash Reinforcement Learning. arXiv:2006.06069.

[180] Huang, Xiaoshui, Fujin Zhu, Lois Holloway, and Ali Haidar. 2020. Causal Discovery from Incomplete Data using An Encoder and Reinforcement Learning. arXiv:2006.05554.

[181] Chow, Yinlam, Brandon Cui, MoonKyung Ryu, and Mohammad Ghavamzadeh. 2020. Variational Model-based Policy Optimization. arXiv:2006.05443.

[182] Jafferjee, Taher, Ehsan Imani, Erin Talvitie, Martha White, and Micheal Bowling. 2020. Hallucinating Value: A Pitfall of Dyna-style Planning with Imperfect Environment Models. arXiv:2006.04363.

[183] Tian, Yuan, Manuel Arias Chao, Chetan Kulkarni, Kai Goebel, and Olga Fink. 2020. Real-Time Model Calibration with Deep Reinforcement Learning. arXiv:2006.04001.

[184] Kallus, Nathan, and Masatoshi Uehara. 2020. Efficient Evaluation of Natural Stochastic Policies in Offline Reinforcement Learning. arXiv:2006.03886.

[185] Hou, Linfang, Liang Pang, Xin Hong, Yanyan Lan, Zhiming Ma, and Dawei Yin. 2020. Robust Reinforcement Learning with Wasserstein Constraint. arXiv:2006.00945.

[186] Zhi, Jixuan, and Jyh-Ming Lien. 2020. Learning to Herd Agents Amongst Obstacles: Training Robust Shepherding Behaviors using Deep Reinforcement Learning. arXiv:2005.09476.

[187] Chandak, Yash, Georgios Theocharous, Shiv Shankar, Martha White, Sridhar Mahadevan, and Philip S. Thomas. 2020. Optimizing for the Future in Non-Stationary MDPs. arXiv:2005.08158.

[188] Ding, Yiming, Ignasi Clavera, and Pieter Abbeel. 2020. Mutual Information Maximization for Robust Plannable Representations. arXiv:2005.08114.

[189] Totaro, Simone, Ioannis Boukas, Anders Jonsson, and Bertrand Cornélusse. 2020. Lifelong Control of Off-grid Microgrid with Model Based Reinforcement Learning. arXiv:2005.08006.

[190] Xie, Zhaoming, Hung Yu Ling, Nam Hee Kim, and Michiel van de Panne. 2020. ALLSTEPS: Curriculum-driven Learning of Stepping Stone Skills. arXiv:2005.04323.

[191] Singh, Rahul, Qinsheng Zhang, and Yongxin Chen. 2020. Improving Robustness via Risk Averse Distributional Reinforcement Learning. arXiv:2005.00585.

[192] Kostrikov, Ilya, Denis Yarats, and Rob Fergus. 2020. Image Augmentation Is All You Need: Regularizing Deep Reinforcement Learning from Pixels. arXiv:2004.13649.

[193] Chen, Jerry Zikun. 2020. Reinforcement Learning Generalization with Surprise Minimization. arXiv:2004.12399.

[194] Ngo, Phuong D., and Fred Godtliebsen. 2020. Data-Driven Robust Control Using Reinforcement Learning. arXiv:2004.07690.

[195] Everett, Michael, Bjorn Lutjens, and Jonathan P. How. 2020. Certified Adversarial Robustness for Deep Reinforcement Learning. arXiv:2004.06496.

[196] Koren, Mark, and Mykel J. Kochenderfer. 2020. Adaptive Stress Testing without Domain Heuristics using Go-Explore. arXiv:2004.04292.

[197] Anahtarci, Berkay, Can Deha Kariksiz, and Naci Saldi. 2020. Q-Learning in Regularized Mean-field Games. arXiv:2003.12151.

[198] Lindenberg, Björn, Jonas Nordqvist, and Karl-Olof Lindahl. 2020. Distributional Reinforcement Learning with Ensembles. arXiv:2003.10903.

[199] Shen, Qianli, Yan Li, Haoming Jiang, Zhaoran Wang, and Tuo Zhao. 2020. Deep Reinforcement Learning with Robust and Smooth Policy. arXiv:2003.09534.

[200] Zhang, Huan, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh. 2020. Robust Deep Reinforcement Learning against Adversarial Perturbations on State Observations. arXiv:2003.08938.

[201] Guo, Xin, Anran Hu, Renyuan Xu, and Junzi Zhang, 2020. A General Framework for Learning Mean-Field Games. arXiv:2003.06069.

[202] Touati, Ahmed, Adrien Ali Taiga, and Marc G. Bellemare. 2020. Zooming for Efficient Model-Free Reinforcement Learning in Metric Spaces. arXiv:2003.04069.

[203] Gao, Shen, Peihao Dong, Zhiwen Pan, and Geoffrey Ye Li. 2020. Reinforcement Learning Based Cooperative Coded Caching under Dynamic Popularities in Ultra-Dense Networks. arXiv:2003.03758.

[204] Lin, Jieyu, Kristina Dzeparoska, Sai Qian Zhang, Alberto Leon-Garcia, and Nicolas Papernot. 2020. On the Robustness of Cooperative Multi-Agent Reinforcement Learning. arXiv:2003.03722.

[205] Derman, Esther, and Shie Mannor. 2020. Distributional Robustness and Regularization in Reinforcement Learning. arXiv:2003.02894.

[206] Spooner, Thomas, and Rahul Savani. Robust Market Making via Adversarial Reinforcement Learning. arXiv:2003.01820.

[207] Chancán, Marvin, and Michael Milford. 2020. MVP: Unified Motion and Visual Self-Supervised Learning for Large-Scale Robotic Navigation. arXiv:2003.00667.

[208] Ilboudo, Wendyam Eric Lionel, Taisuke Kobayashi, and Kenji Sugimoto. 2020. TAdam: A Robust Stochastic Gradient Optimizer. arXiv:2003.00179.

[209] Tschantz, Alexander, Beren Millidge, Anil K. Seth, and Christopher L. Buckley. 2020. Reinforcement Learning through Active Inference. arXiv:2002.12636.

[210] Kuutti, Sampo, Saber Fallah, and Richard Bowden. 2020. Training Adversarial Agents to Exploit Weaknesses in Deep Control Policies. arXiv:2002.12078.

[211] Nguyen, Ngoc Duy, Thanh Thi Nguyen, and Saeid Nahavandi. 2020. A Visual Communication Map for Multi-Agent Deep Reinforcement Learning. arXiv:2002.11882.

[212] Yang, Chao-Han Huck, Jun Qi, Pin-Yu Chen, Yi Ouyang, I-Te Danny Hung, Chin-Hui Lee, and Xiaoli Ma. 2020. Enhanced Adversarial Strategically-Timed Attacks against Deep Reinforcement Learning. arXiv:2002.09027.

[213] Sun, Tao, Han Shen, Tianyi Chen, and Dongsheng Li. 2020. Adaptive Temporal Difference Learning with Linear Function Approximation. arXiv:2002.08537.

[214] Naderializadeh, Navid, Jaroslaw Sydir, Meryem Simsek, and Hosein Nikopour. 2020. Resource Management in Wireless Networks via Multi-Agent Deep Reinforcement Learning. arXiv:2002.06215.

[215] Kamalaruban, Parameswaran, Yu-Ting Huang, Ya-Ping Hsieh, Paul Rolland, Cheng Shi, and Volkan Cevher. 2020. Robust Reinforcement Learning via Adversarial training with Langevin Dynamics. arXiv:2002.06063.

[216] Kallus, Nathan, and Masatoshi Uehara. 2020. Statistically Efficient Off-Policy Policy Gradients. arXiv:2002.04014.

[217] Lee, Gilwoo, Brian Hou, Sanjiban Choudhury, and Siddhartha S. Srinivasa. 2020. Bayesian Residual Policy Optimization: Scalable Bayesian Reinforcement Learning with Clairvoyant Experts. arXiv:2002.03042.

[218] Pacelli, Vincent, and Anirudha Majumdar. 2020. Learning Task-Driven Control Policies via Information Bottlenecks. arXiv:2002.01428.

[219] Yao, Jiahao, Marin Bukov, and Lin Lin. 2020. Policy Gradient based Quantum Approximate Optimization Algorithm. arXiv:2002.01068.

[220] Nishio, Daichi, Daiki Kuyoshi, Toi Tsuneda, and Satoshi Yamane. 2020. Discriminator Soft Actor Critic without Extrinsic Rewards. arXiv:2001.06808.

[221] Dai, Tianhong, Kai Arulkumaran, Tamara Gerbert, Samyakh Tukra, Feryal Behbahani, and Anil Anthony Bharath. 2020. Analysing Deep Reinforcement Learning Agents Trained with Domain Randomisation. arXiv:1912.08324.

[222] Zhang, Xinglong, Jiahang Liu, Xin Xu, Shuyou Yu, and Hong Chen. 2020. Learning-based Predictive Control for Nonlinear Systems with Unknown Dynamics Subject to Safety Constraints. arXiv:1911.09827.

[223] Lykouris, Thodoris, Max Simchowitz, Aleksandrs Slivkins, and Wen Sun. 2020. Corruption robust exploration in episodic reinforcement learning. arXiv:1911.08689.

[224] Salter, Sasha, Dushyant Rao, Markus Wulfmeier, Raia Hadsell, and Ingmar Posner. 2020. Attention-Privileged Reinforcement Learning. arXiv:1911.08363.

[225] Han, Minghao. Yuan Tian, Lixian Zhang, Jun Wang, and Wei Pan. 2020. H∞ Model-free Reinforcement Learning with Robust Stability Guarantee. arXiv:1911.02875.

[226] Lütjens, Björn, Michael Everett, and Jonathan P. How. 2020. Certified Adversarial Robustness for Deep Reinforcement Learning. arXiv:1910.12908.

[227] Uehara, Masatoshi, Jiawei Huang, and Nan Jiang. 2020. Minimax Weight and Q-Function Learning for Off-Policy Evaluation. arXiv:1910.12809.

[228] Li, Shuo, and Osbert Bastani. 2020. Robust Model Predictive Shielding for Safe Reinforcement Learning with Stochastic Dynamics. arXiv:1910.10885.

[229] Slaoui, Reda Bahi, William R. Clements, Jakob N. Foerster, and Sébastien Toth. 2020. Robust Visual Domain Randomization for Reinforcement Learning. arXiv:1910.10537.

[230] Zhang, Kaiqing, Bin Hu, and Tamer Başar. 2020. Policy Optimization for H2 Linear Control with H∞ Robustness Guarantee: Implicit Regularization and Global Convergence. arXiv:1910.09496.

[231] Liu, Zhuang, Xuanlin Li, Bingyi Kang, and Trevor Darrell. 2020. Regularization Matters in Policy Optimization. arXiv:1910.09191.

[232] Yang, Jiachen, Brenden Petersen, Hongyuan Zha, and Daniel Faissol. 2020. Single Episode Policy Transfer in Reinforcement Learning. arXiv:1910.07719.

[233] Chen, Shuhang, Adithya M. Devraj, Fan Lu, Ana Bušić, and Sean P. Meyn. 2020. Zap Q-Learning with Nonlinear Function Approximation. arXiv:1910.05405.

[234] Schwartz, Erez, Guy Tennenholtz, Chen Tessler, and Shie Mannor. 2020. Language is Power: Representing States Using Natural Language in Reinforcement Learning. arXiv:1910.02789.

[235] Yarats, Denis, Amy Zhang, Ilya Kostrikov, Brandon Amos, Joelle Pineau, and Rob Fergus. 2020. Improving Sample Efficiency in Model-Free Reinforcement Learning from Images. arXiv:1910.01741.

[236] Kalweit, Gabriel, Maria Huegle, and Joschka Boedecker. 2020. Composite Q-learning: Multi-scale Q-function Decomposition and Separable Optimization. arXiv:1909.13518.

[237] Ryu, Moonkyung, Yinlam Chow, Ross Anderson, Christian Tjandraatmadja, and Craig Boutilier. 2020. CAQL: Continuous Action Q-Learning. arXiv:1909.12397.

[238] Li, Jiachen, Quan Vuong, Shuang Liu, Minghua Liu, Kamil Ciosek, Henrik Iskov Christensen, and Hao Su. 2020. Multi-task Batch Reinforcement Learning with Metric Learning. arXiv:1909.11373.

[239] Shen, Macheng, and Jonathan P. How. 2020. Robust Opponent Modeling via Adversarial Ensemble Reinforcement Learning in Asymmetric Imperfect-Information Games. arXiv:1909.08735.

[240] Kallus, Nathan, Masatoshi Uehara. 2020. Double Reinforcement Learning for Efficient Off-Policy Evaluation in Markov Decision Processes. arXiv:1908.08526.

[241] Roy, Julien, Paul Barde, Félix G. Harvey, Derek Nowrouzezahrai, and Christopher Pal. 2020. Promoting Coordination through Policy Regularization in Multi-Agent Deep Reinforcement Learning. arXiv:1908.02269.

[242] Urakami, Yusuke, Alec Hodgkinson, Casey Carlin, Randall Leu, Luca Rigazio, and Pieter Abbeel. 2020. DoorGym: A Scalable Door Opening Environment and Baseline Agent. arXiv:1908.01887.

[243] Wang, Qisheng, Keming Feng, Xiao Li, and Shi Jin. 2020. PrecoderNet: Hybrid Beamforming for Millimeter Wave Systems with Deep Reinforcement Learning. arXiv:1907.13266.

[244] Bogdanovic, Miroslav, Majid Khadiv, Ludovic Righetti. 2020. Learning Variable Impedance Control for Contact Sensitive Tasks. arXiv:1907.07500.

[245] Mankowitz, Daniel J., Nir Levine, Rae Jeong, Yuanyuan Shi, Jackie Kay, Abbas Abdolmaleki, Jost Tobias Springenberg, Timothy Mann, Todd Hester, and Martin Riedmiller. 2020. Robust Reinforcement Learning for Continuous Control with Model Misspecification. arXiv:1906.07516.

[246] Li, Alexander C., Carlos Florensa, Ignasi Clavera, and Pieter Abbeel. Sub-policy Adaptation for Hierarchical Reinforcement Learning. arXiv:1906.05862.

[247] Assran, Mahmoud, Joshua Romoff, Nicolas Ballas, Joelle Pineau, and Michael Rabbat. 2020. Gossip-based Actor-Learner Architectures for Deep Reinforcement Learning. arXiv:1906.04585.

[248] Gravell, Benjamin, Peyman Mohajerin Esfahani, and Tyler Summers. 2020. Learning robust control for LQR systems with multiplicative noise via policy gradient. arXiv:1905.13547.

[249] Francis, Anthony, Aleksandra Faust, Hao-Tien Lewis Chiang, Jasmine Hsu, J. Chase Kew, Marek Fiser, and Tsang-Wei Edward Lee. 2020. Long-Range Indoor Navigation with PRM-RL. arXiv:1902.09458.

[250] Wang, Jingkang, Yang Liu, and Bo Li. 2020. Reinforcement Learning with Perturbed Rewards. arXiv:1810.01032.

Other

[251] Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. 2009. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Med* 6(7): e1000097. doi 10.1371/journal.pmed1000097

[252] Liu, Y., P. Ramachandran, Q. Liu, and J. Peng, Stein Variational Policy Gradient. arXiv:1704.02399v1 [cs.LG] 7 April 2017.

[253] 5. Liang, G., Zhu, X., Zhang, C. An Empirical Study of Bagging Predictors for Different Learning Algorithms. *AAAI*, pp. 1802–1803 (2011)

[254] Brockman, Greg and Cheung, Vicki and Pettersson, Ludwig and Schneider, Jonas and Schulman, John and Tang, Jie and Zaremba, Wojciech. Openai gym, arXiv preprint arXiv:1606.01540, 2016.