



Universiteit  
Leiden  
The Netherlands

## Searching by learning: Exploring artificial general intelligence on small board games by deep reinforcement learning

Wang, H.

### Citation

Wang, H. (2021, September 7). *Searching by learning: Exploring artificial general intelligence on small board games by deep reinforcement learning*. Retrieved from <https://hdl.handle.net/1887/3209232>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/3209232>

**Note:** To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle <https://hdl.handle.net/1887/3209232> holds various files of this Leiden University dissertation.

**Author:** Wang, H.

**Title:** Searching by learning: Exploring artificial general intelligence on small board games by deep reinforcement learning

**Issue Date:** 2021-09-07

# Searching by Learning: Exploring Artificial General Intelligence on Small Board Games by Deep Reinforcement Learning

**Proefschrift**

ter verkrijging van  
de graad van doctor aan de Universiteit Leiden,  
op gezag van rector magnificus prof.dr.ir. H. Bijl,  
volgens besluit van het college voor promoties  
te verdedigen op dinsdag 7 september 2021  
klokke 16.15 uur

door

**Hui Wang**

geboren te Anhui, China

in 1992

---

## Promotiecommissie

Promotors: Dr. Michael Emmerich  
Prof. Dr. Aske Plaat

Co-promotor: Dr. Mike Preuss

Overige leden: Prof. Dr. Thomas Bäck  
Prof. Dr. Marcello Bonsangue  
Prof. Dr. Joost Batenburg  
Prof. Dr. Mark Winands Maastricht University  
Dr. Mitra Baratchi  
Dr. Thomas Moerland  
Dr. Ingo Schwab Karlsruhe University of Applied Science

Copyright © 2021 Hui Wang All Rights Reserved

ISBN: 978-94-6419-253-7

Het onderzoek beschreven in dit proefschrift is uitgevoerd aan het Leiden Institute of Advanced Computer Science (LIACS, Universiteit Leiden).

This Research is financially supported by the China Scholarship Council (CSC), CSC No. 201706990015.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Background . . . . .	1
1.2	Research Questions . . . . .	5
1.3	Dissertation Outline . . . . .	7
<b>2</b>	<b>Classical Q-learning in GGP</b>	<b>11</b>
2.1	Introduction . . . . .	11
2.2	Related Work and Preliminaries . . . . .	12
2.2.1	GGP . . . . .	12
2.2.2	Reinforcement Learning . . . . .	13
2.2.3	Q-learning . . . . .	13
2.3	Design . . . . .	14
2.3.1	Classical Q-learning for Two-Player Games . . . . .	14
2.3.2	Dynamic $\epsilon$ Enhancement . . . . .	15
2.3.3	QM-learning Enhancement . . . . .	16
2.4	Experiments and Results . . . . .	17
2.4.1	Dynamic $\epsilon$ Enhancement . . . . .	17
2.4.2	QM-learning Enhancement . . . . .	20
2.5	Summary . . . . .	23
<b>3</b>	<b>Hyper-Parameters for AlphaZero-like Self-play</b>	<b>25</b>
3.1	Introduction . . . . .	25
3.2	Related work . . . . .	26
3.3	Test Game . . . . .	27
3.4	AlphaZero-like Self-play . . . . .	28
3.4.1	The Base Algorithm . . . . .	28
3.4.2	Loss Function . . . . .	30
3.4.3	Bayesian Elo System . . . . .	30

## CONTENTS

---

3.4.4	Time Cost Function . . . . .	31
3.5	Experimental Setup . . . . .	32
3.5.1	Hyper-Parameter Sweep . . . . .	32
3.5.2	Hyper-Parameters Correlation Evaluation . . . . .	33
3.6	Experimental Results . . . . .	33
3.6.1	Hyper-Parameter Sweep Results . . . . .	34
3.6.2	Hyper-Parameter Correlation Evaluation Results . . . . .	38
3.7	Summary . . . . .	41
<b>4</b>	<b>Loss Functions of AlphaZero-like Self-play</b>	<b>43</b>
4.1	Introduction . . . . .	43
4.2	Related Work . . . . .	45
4.3	Test Games . . . . .	45
4.4	Loss Function . . . . .	47
4.4.1	Minimization Targets . . . . .	47
4.5	Experimental Setup . . . . .	47
4.5.1	Measurements . . . . .	48
4.6	Experiment Results . . . . .	48
4.6.1	Training Loss . . . . .	49
4.6.2	Training Elo Rating . . . . .	53
4.6.3	The Final Best Player Tournament Elo Rating . . . . .	55
4.7	Summary . . . . .	57
<b>5</b>	<b>Warm-Starting AlphaZero-like Self-Play</b>	<b>59</b>
5.1	Introduction . . . . .	59
5.2	Related Work . . . . .	61
5.3	AlphaZero-like Self-play Algorithms . . . . .	62
5.3.1	The Algorithm Framework . . . . .	62
5.3.2	MCTS . . . . .	64
5.3.3	MCTS Enhancements . . . . .	64
5.4	Initial Experiment: MCTS(RAVE) vs. RHEA . . . . .	66
5.5	Full Length Experiment . . . . .	67
5.5.1	Experiment Setup . . . . .	67
5.5.2	Results . . . . .	68
5.6	Summary . . . . .	69
<b>6</b>	<b>Adaptive Warm-Start AlphaZero-like Self-play</b>	<b>73</b>
6.1	Introduction . . . . .	73
6.2	Related Work . . . . .	75

6.3	Warm-Start AlphaZero Self-play . . . . .	76
6.3.1	The Algorithm Framework . . . . .	76
6.3.2	MCTS . . . . .	76
6.3.3	MCTS enhancements . . . . .	77
6.4	Adaptive Warm-Start Switch Method . . . . .	78
6.5	Experimental Setup . . . . .	79
6.6	Results . . . . .	80
6.6.1	MCTS vs MCTS Enhancements . . . . .	80
6.6.2	Fixed $I'$ Tuning . . . . .	81
6.6.3	Adaptive Warm-Start Switch . . . . .	83
6.7	Summary . . . . .	86
<b>7</b>	<b>Ranked Reward Reinforcement Learning</b>	<b>89</b>
7.1	Introduction . . . . .	89
7.2	Related Work . . . . .	90
7.3	Morpion Solitaire . . . . .	91
7.4	Ranked Reward Reinforcement Learning . . . . .	92
7.5	Experiment Setup . . . . .	94
7.6	Result and Analysis . . . . .	95
7.7	Summary . . . . .	97
<b>8</b>	<b>Conclusion</b>	<b>99</b>
8.1	Contributions . . . . .	100
8.2	Outlook . . . . .	102
<b>A</b>		<b>105</b>
A.1	Symbols . . . . .	105
A.2	Abbreviations . . . . .	106
A.3	Algorithms . . . . .	107
A.4	Elo Computation . . . . .	115
	<b>Bibliography</b>	<b>119</b>
	<b>English Summary</b>	<b>131</b>
	<b>Nederlandse Samenvatting</b>	<b>133</b>
	<b>Acknowledgements</b>	<b>135</b>
	<b>Curriculum Vitae</b>	<b>137</b>

## CONTENTS

---