Representation-Induced Algorithmic Bias

An empirical assessment of behavioural equivalence over 14 reinforcement learning algorithms across 4 isomorphic gameform representations

Stanton, S. C.[†*], Dermoudy, J.[*], Ollington, R.[*]

AJCAI2021

Abstract

In conceiving of autonomous agents able to employ adaptive cooperative behaviours we identify the need to effectively assess the equivalence of agent behavior under conditions of external change. Reinforcement learning algorithms rely on input from the environment as the sole means of reifying internal state. This work investigates the assumption that isomorphic representations of environment will lead to equivalent behaviour. To test this *equivalence-of* assumption we analyse the variance between behavioural profiles in a set of agents using fourteen foundational reinforcement-learning algorithms across four isomorphic representations of the classical Prisoner's Dilemma gameform. A behavioural profile exists as the aggregated episode-mean distributions of the game outcomes CC, CD, DC, and DD generated from the symmetric selfplay repeated stage game across a two-axis sweep of input parameters: the principal learning rate, α, and the discount factor γ, which provides 100 observations of the frequency of the four game outcomes, per algorithm, per gameform representation. A measure of equivalence is indicated by a low variance displayed between any two behavioural profiles generated by any one single algorithm. ***Despite the representations being theoretically equivalent analysis reveals significant variance in the behavioural profiles of the tested algorithms at both aggregate and individual outcome scales.*** Given this result, we infer that the isomorphic representations tested in this study are not necessarily equivalent with respect to the induced reachable space made available to any specific algorithm, which in turn can lead to unexpected agent behaviour. Therefore, we conclude that structure-preserving operations applied to environmental reward signals may introduce a vector for algorithmic bias.

Introduction

An agent's relationship to its environment may change *in-situ* due to the environment being mutable, such that the process for derivation of reward signals being input *to* the agent may fluctuate. Also, an agent's internal state representation may transform as part of normal algorithmic operation. Similarly, an agent's method of utility extraction from reward signals may be altered by design, or computationally. Alternatively, an agent or algorithm may be entirely transplanted from one use case or application to another. **A generalised perspective on these processes captures those occasions in everyday life that entail an implicit cast between representations, for example, whenever we (in the real-world) fix a price-point to a preference, or conversely, when compromising on features of a purchase with a known price.** In other words, whenever we translate from scalar (or cardinal) values to ordinal preferences, and vice versa. This conversion impacts utility functions based in preference relations, which has relevance to how an agent may adapt to change in any external stimuli, that is, to any change in the reward signal representation as offered by the environment.

Ashlock and Kim (2008, p647) investigated the role of representation in an evolutionary computational context and found in one series of experiments that "all three representations sample the strategy space in a radically different manner", and, in another series of experiments concluded that "changing the payoff matrix, within the bounds permitted by the defining inequalities of prisoner's dilemma, yields different results" (Ashlock et al., 2010, p225). For various learning algorithms, Crandall et al. (2018b, pp8–10) note that "actual payoff values assigned to the ordinal preferences … can, and often do, impact the behaviors of some algorithms in repeated games", and in the discussion of an empirical investigation into a topological representation of Prisoner's Dilemma, Robinson and Goforth (2005, p156) assert that while a topology of ordinal games defines the relationships between these games, the topology "is insufficient for describing and predicting patterns of behaviour". Given that Robinson and Goforth also assert that "any ordinally equivalent game is also a Prisoner's Dilemma" (2005, p. 6) we are compelled to investigate how various reinforcement learning algorithms will perform over isomorphic representations of this canonical gameform.

Methods

Our initial hypothesis is that an algorithmic behavioural profile *will not vary between equivalent representations of the game.* The response variable of interest is *variance in the aggregated distribution of cooperative game outcomes between representations*, as observed in a symmetric selfplay paired parameter study of 14 foundational reinforcement learning algorithms listed in **Table 1.**

Table 1. Algorithms implemented and their parameters. Parameter superscript ¹² denotes paired parameters.

Algorithm	Parameters Learning Rate(s)	Discount	Trace	Action-Selection
Actor/Critic	$\alpha^1, \beta = 0.9$	γ^2		$softargmax$
Actor/Critic with eligibility traces	$\alpha^1, \beta = 0.9$	γ^2	$\lambda = 0.9$	$softargmax$
Actor-Critic with replacing traces	$\alpha^1, \beta = 0.9$	γ^2	$\lambda = 0.9$	$softargmax$
Q-Learning	α^1	γ^2		$\varepsilon = 0.1$
Double Q-Learning	α^1	γ^2		$\varepsilon = 0.1$
Expected SARSA	α^1	γ^2		$\varepsilon = 0.1$
R Learning	α^1, β^2			$\varepsilon = 0.1$
SARSA	α^1	γ^2		$\varepsilon = 0.1$
SARSA Lambda	α^1	γ^2	$\lambda = 0.9$	$\varepsilon = 0.1$
SARSA Lambda, with replacing traces	α^1	γ^2	$\lambda = 0.9$	$\varepsilon = 0.1$
Watkins (naïve) Q, Lambda	α^1	γ^2	$\lambda = 0.9$	$\varepsilon = 0.1$
Watkins (naïve) Q, Lambda, replacing traces	α^1	γ^2	$\lambda = 0.9$	$\varepsilon = 0.1$
Watkins Q, Lambda	α^1	γ^2	$\lambda = 0.9$	$\varepsilon = 0.1$
Watkins Q, Linear Function Approximation	α^1	γ^2		$\varepsilon = 0.1$

The experiments in the study are implemented as symmetric selfplay repeated stage games along two dimensions defined by two parameters that range from (0,1], in 0.1 increments; giving 100 observations for each algorithm. **Each observation is the terminal episode-mean distribution of the four cooperative game outcomes CC, CD, DC, and DD over 500 episodes of 1000 timesteps each.** The collection of experiments that is formed by running each algorithm in a single gameform representation gives a set of behavioural profiles. The combined grouping of two sets of behavioural profiles are referred to as experiment groups 1 through 4.

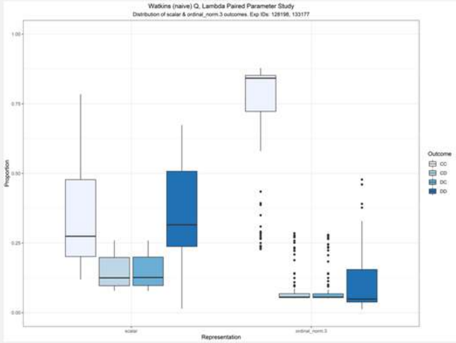
The 14 learning algorithms implemented are drawn from foundational work in Reinforcement Learning (Sutton and Barto, 1998, 2018; Szepesvári, 2010). Each algorithm is implemented with a representation of state by tabular data structures except for Watkins Q Linear Function Approximation (Sutton and Barto, 1998, 2018) where features are implemented as a mapping from the previous timestep outcome to a vector parameterised by a weight vector. **For each algorithm, the parameters varied are the principal learning rate parameter α, and, in all cases except R-Learning, the discount factor parameter γ.**

We adopt the following explicit constraints: firstly, that each episode is effectively independent, and in real-time, or *Online*; secondly, episodes have an *Infinite Horizon*, as such, T=∞; thirdly, *No Signaling*, in that there is no communication between agents; and lastly, the *Environment Boundary* is everything external to the agent and the only information crossing this boundary is the reward signal.

Each outcome distribution is tested for normality via the Shapiro-Wilk normality test, and then tested for equivalence via the Wilcoxon Signed Rank (Paired Treatment) test.

Conclusion

Our results indicate that the behaviour of the algorithms can vary substantially as a product of the input representation of an otherwise equivalent gameform. This suggests that we cannot unconditionally generalise learning algorithms between representations, either *a priori*, or *in-situ*, without possibly introducing bias. We also assert that the character of the behavioural profile can be altered considerably, as is evident with *Watkins (naïve) Q Lambda* in experiment group three, shown in **Fig. 3**.



We conclude that the null hypothesis is not supported, as shown in **Table 7.** Only *Watkins Q Linear Function Approximation* exhibits non-significant variance in all experiment groups, however the result in experiment group four (V=44057, p-value=.07136) for this algorithm is somewhat weak.

Table 7. Algorithms that do not exhibit significant variance between behavioural profiles.

Experiment Group	p-value > .05	Experiment Group	p-value > .05
One	Actor/Critic Actor/Critic with Replacing Traces **Watkins Q, Linear Function Approximation**	Three	Double Q-Learning **Watkins Q, Linear Function Approximation**
Two	Actor/Critic Actor/Critic with Replacing Traces Q-Learning Double Q-Learning Expected SARSA R Learning SARSA SARSA Lambda, with Replacing Traces Watkins Q, Lambda **Watkins Q, Linear Function Approximation**	Four	Actor/Critic with Replacing Traces SARSA Lambda SARSA Lambda, with Replacing Traces Watkins (naïve) Q, Lambda Watkins (naïve) Q, Lambda, Replacing Traces Watkins Q, Lambda **Watkins Q, Linear Function Approximation**

Results

Experiment Group One

In experiment group one the mapping between **pd:scalar** and **g111:ordinal** is assessed. Results of the Wilcoxon tests are shown in **Table 3**. Of the fourteen algorithms, eleven have a p-value < .05, which indicates that these algorithms do not exhibit equivalence of behaviour between the two representations in this experiment group. Only three algorithms—*Actor/Critic*, *Actor/Critic with Eligibility Traces*, and *Watkins Q Linear Function Approximation*—do not exhibit variance that can be regarded as significant.

Experiment Group Two

Results of the Wilcoxon Signed Rank tests for experiment group two are shown in **Table 4**. It is apparent that of the algorithms in this experiment group, four have a p-value < .05. This indicates that these four algorithms do not exhibit equivalence between scalar and normalised scalar representations.

Experiment Group Three

Results of the Wilcoxon Signed Rank tests for experiment group three are shown in **Table 5**. Twelve of the algorithms have p-value < .05 indicating that these algorithms do not exhibit equivalence of behaviour between the two representations in this experiment group.

Experiment Group Four

Results of the Wilcoxon Signed Rank tests for experiment group four are shown in **Table 6**. Of the fourteen algorithms, seven have p-values < .05 indicating that these algorithms do not exhibit equivalence of behaviour between the two representations in this experiment group. The set of algorithms that return significant p-values differs from those that exhibit this property in experiment groups one, two, and three, as detailed in **Table 7**.

In each experiment group we highlight the episode-mean frequency of the peak cooperative outcome to show that where some algorithms appear to achieve near-parity in optimal behaviour between representations, other algorithms exhibit substantial disparity.

The aggregate distribution of each behavioural profile returns a non-normal distribution as determined by a Shapiro-Wilk Normality test; for example, the *Watkins-Q Linear Function Approximation* algorithm's aggregated behavioural profile for the normalised ordinal representation returns a Shapiro-Wilk score of W=0.96486, p-value=.00000003331).

Table 3. Exp Group One: Scalar (S) ~ Ordinal (O) Aggregated Distribution.

Algorithm	Peak % CC S	O	Wilcoxon V	p-value	CI L	CI U
Actor/Critic	32.1	30.2	35742	.07125	-0.0176	0.0006
Actor/Critic with Eligibility Traces	46.3	44.9	40978	.7047	-0.0036	0.0043
Actor/Critic with Replacing Traces	51.5	42.6	45337	.01835	0.0019	0.0200
Q-Learning	80.1	81.5	46086	.009683	0.0010	0.0061
Double Q-Learning	68.9	32.7	34431	.01428	-0.0059	-0.0006
Expected SARSA	75.7	77.3	27735	$9.08e-08$	-0.0185	-0.0070
R Learning	25.6	28.3	45520	.01916	0.0024	0.0177
SARSA	79.9	80.3	46848	.003545	0.0012	0.0053
SARSA Lambda	68.8	86.7	49435	$5.47e-05$	0.0171	0.0411
SARSA Lambda, with Replacing Traces	89.0	89	45731	.01495	0.0005	0.0057
Watkins (naïve) Q, Lambda	78.4	88.1	50043	$1.73e-05$	0.0270	0.0505
Watkins (naïve) Q, Lambda, Replacing Traces	76.4	89.1	49320	$6.76e-05$	0.0193	0.0487
Watkins Q, Lambda	87.2	87.8	49355	$6.35e-05$	0.0029	0.0089
Watkins Q, Linear Function Approximation	47.1	46.2	38262	.4775	-0.0013	0.0006

Table 4. Exp Group Two: Scalar (S) ~ Normalised Scalar (NS) Aggregated Distribution.

Algorithm	Peak % CC S	NS	Wilcoxon V	p-value	CI L	CI U
Actor/Critic	32.1	14.0	41239	.6227	-0.0049	0.0074
Actor/Critic with Eligibility Traces	46.3	33.7	46744	.004087	0.0026	0.0120
Actor/Critic with Replacing Traces	51.5	44.9	39599	.8286	-0.0110	0.0091
Q-Learning	80.1	80.2	40091	.9971	-0.0004	0.0004
Double Q-Learning	68.9	68.5	39339	.7424	-0.0005	0.0004
Expected SARSA	75.7	75.7	37225	.214	-0.0006	0.0001
R Learning	25.6	23.7	39866	.9196	-0.0004	0.0004
SARSA	79.9	80.2	40180	.9037	-0.0003	0.0004
SARSA Lambda	68.8	86.6	49945	$2.09e-05$	0.0123	0.0308
SARSA Lambda, with Replacing Traces	89.0	89.1	39177	.6901	-0.0005	0.0003
Watkins (naïve) Q, Lambda	78.4	87.2	49853	$2.50e-05$	0.0273	0.0488
Watkins (naïve) Q, Lambda, Replacing Traces	76.4	89.2	48331	$2.55e-05$	0.0120	0.0382
Watkins Q, Lambda	87.2	87.2	39692	.9281	-0.0005	0.0004
Watkins Q, Linear Function Approximation	47.1	24.9	39939	.9445	-0.0004	0.0004

Table 5. Exp Group Three: Scalar (S) ~ Normalised-Ordinal (NO) Aggregated Distribution.

Algorithm	Peak % CC S	NO	Wilcoxon V	p-value	CI L	CI U
Actor/Critic	32.1	19.7	47086	.002535	0.0048	0.0195
Actor/Critic with Eligibility Traces	46.3	37.4	48229	.0003028	0.0048	0.0163
Actor/Critic with Replacing Traces	51.5	39.0	44761	.04398	0.0003	0.0253
Q-Learning	80.1	80.4	47276	.001377	0.0023	0.0075
Double Q-Learning	68.9	68.9	43233	.1758	-0.0006	0.0034
Expected SARSA	75.7	76.9	49542	$2.88e-05$	0.0021	0.0049
R Learning	25.6	24.1	45745	.0147	0.0089	0.0242
SARSA	79.9	80.9	47300	.001862	0.0019	0.0068
SARSA Lambda	68.8	87.0	49205	$8.33e-05$	0.0122	0.0311
SARSA Lambda, with Replacing Traces	89.0	89.2	48050	.0005908	0.001	0.0033
Watkins (naïve) Q, Lambda	78.4	87.8	49985	$1.94e-05$	0.0269	0.0496
Watkins (naïve) Q, Lambda, Replacing Traces	76.4	89.1	49091	.000102	0.0154	0.0404
Watkins Q, Lambda	87.2	86.9	50420	$5.03e-06$	0.0020	0.0047
Watkins Q, Linear Function Approximation	47.1	46.8	38471	.4815	-0.0007	0.0004

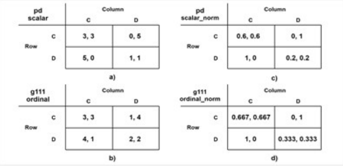
Table 6. Exp Group Four: Ordinal (O) ~ Normalised Ordinal (NO) Aggregated Distribution.

Algorithm	Peak % CC O	NO	Wilcoxon V	p-value	CI L	CI U
Actor/Critic	30.2	19.7	48983	.0001235	0.0285	0.0712
Actor/Critic with Eligibility Traces	44.9	37.4	45950	.01147	0.0015	0.0147
Actor/Critic with Replacing Traces	42.6	39.0	43710	.1188	-0.0016	0.0141
Q-Learning	81.5	80.4	48454	.0003057	0.0022	0.0066
Double Q-Learning	32.7	68.9	51223	$1.53e-06$	0.0067	0.0140
Expected SARSA	77.3	76.9	52508	$8.21e-08$	0.0113	0.0248
R Learning	28.3	24.1	49944	$1.32e-05$	0.0011	0.0026
SARSA	80.3	80.9	49862	$2.45e-05$	0.0039	0.0090
SARSA Lambda	86.7	87.0	36226	.0941	-0.0036	0.0002
SARSA Lambda, with Replacing Traces	89	89.2	39935	.9431	-0.0019	0.0017
Watkins (naïve) Q, Lambda	88.1	87.8	42598	.2804	-0.0005	0.0017
Watkins (naïve) Q, Lambda, Replacing Traces	89.1	89.1	38132	.3951	-0.0020	0.0008
Watkins Q, Lambda	87.8	86.9	35790	.07462	-0.0037	0.0001
Watkins Q, Linear Function Approximation	46.2	46.8	44057	.07136	-0.0001	0.0014

Formal Equivalence

Formal Equivalence of Representations via Inequality Rules

We examine four representations of the Prisoner's Dilemma in normal form as shown in **Fig. 1**. Robinson and Goforth (2005) refer to the ordinal Prisoner's Dilemma as game **g111**. Each of the four inequalities relate to the four game outcomes as shown in **Fig. 2**.



a)

pd scalar		Column C	Column D
Row	C	3, 3	0, 5
Row	D	5, 0	1, 1

c)

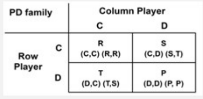
pd scalar_norm		Column C	Column D
Row	C	0.6, 0.6	0, 1
Row	D	1, 0	0.2, 0.2

b)

g111 ordinal		Column C	Column D
Row	C	3, 3	1, 4
Row	D	4, 1	2, 2

d)

g111 ordinal_norm		Column C	Column D
Row	C	0.667, 0.667	0, 1
Row	D	1, 0	0.333, 0.333



PD family		Column Player C	Column Player D
Row Player	C	R (C,C) (R,R)	S (C,D) (S,T)
Row Player	D	T (D,C) (T,S)	P (D,D) (P, P)

Fig 1 (left) Four gameform representations of canonical Prisoner's Dilemma. **Fig 2** (above) Semantic labels attributed to outcomes in the Prisoner's Dilemma. R (Reward), S (Sucker), T (Temptation), and P (Punishment).

Social dilemma inequalities can be found in slightly differing but effectively comparable forms in the literature (Ashlock and Kim, 2008), however the form as given by Macy and Flache (2002, p. 7229) and Leibo et al. (2017) is presented here:

$$R > P \quad (1)$$

$$R > S \quad (2)$$

$$2R > T + S \quad (3)$$

$$T > R, or\ P > S \quad (4)$$

We establish the formal equivalence of **pd:scalar** with the alternative representations by observing that the social dilemma inequalities that define Prisoner's Dilemma *do hold*, as detailed in **Table 2**, illustrating that these four representations of Prisoner's Dilemma are structurally equivalent.

Table 2. Social Dilemma Inequalities

Inequality		Scalar	Scalar Normalised	Ordinal	Ordinal Normalised	Equivalent
(1)	$R > P$	$3 > 1$	$0.6 > 0.2$	$3 > 2$	$0.667 > 0.333$	*True*
(2)	$R > S$	$3 > 0$	$0.6 > 0$	$3 > 1$	$0.667 > 0$	*True*
(3)	$2R > T + S$	$6 > 5 + 0$	$1.2 > 1 + 0$	$6 > 4 + 1$	$1.334 > 1 + 0$	*True*
(4)	$T > R$	$5 > 3$	$1 > 0.6$	$4 > 3$	$1 > 0.667$	*True*
	$P > S$	$1 > 0$	$0.2 > 0$	$2 > 1$	$0.333 > 0$	*True*

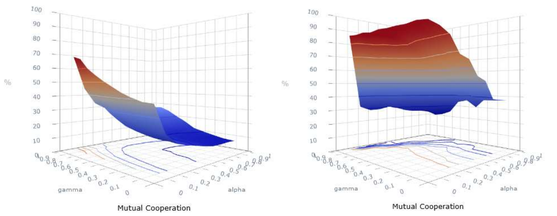


Fig. 4: SARSA Lambda behavioural profiles under two representations: left, scalar; and right, ordinal.

14/11/2021

References

Ashlock, D., & Kim, E.-Y. (2008). *Fingerprinting: Visualization and Automatic Analysis of Prisoner's Dilemma Strategies.* IEEE Transactions on Evolutionary Computation, 12(5), 647–659. doi:10.1109/TEVC.2008.920675

Ashlock, D., Kim, E.-Y., & Ashlock, W. (2010). *A fingerprint comparison of different Prisoner's Dilemma payoff matrices.* Proceedings of the 2010 IEEE Conference on Computational Intelligence and Games, 219–226. doi:10.1109/ITW.2010.5593352

Crandall, J. W., Oudah, M., Tennom, Ishowo-Oloko, F., Abdallah, S., Bonnefon, J.-F., Cebrian, M., Shariff, A., Goodrich, M. A., & Rahwan, I. (2018b). *Supplementary Material—Cooperating with Machines*. Nature Communications, 9(1). https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-017-02597-8/MediaObjects/41467_2017_2597_MOESM1_ESM.pdf, Accessed 2020-01-07

Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., & Graepel, T. (2017). *Multi-agent Reinforcement Learning in Sequential Social Dilemmas.* Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems, 464–473.

Macy, M. W., & Flache, A. (2002). *Learning dynamics in social dilemmas.* Proceedings of the National Academy of Sciences, 99 (suppl 3), 7229–7236, doi:10.1073/pnas.092080099

Robinson, D., & Goforth, D. (2005). *The Topology of the 2x2 games: A New Periodic Table.* Routledge. doi:10.4324/9780203340271

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction* (1st ed.). The MIT Press.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction (2nd ed.).* The MIT Press.

Szepesvári, C. (2010). Morgan and Claypool Publishers. *Algorithms for Reinforcement Learning.*

[†]simon.stanton@utas.edu.au [*]University of Tasmania **Acknowledgements**: We would like to acknowledge the use of the high-performance computing facilities provided by Digital Research Services, University of Tasmania. This research is supported by an Australian Government Research Training Program (RTP) Scholarship. **Code Availability:** A repository of code used in this study is available at https://github.com/simoncstanton/equivalence_study.