Representation-Induced Algorithmic Bias

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

Simon C Stanton[1[0000-0003-0312-4407]] Julian Dermoudy[1[0000-0002-8556-5523]] Robert Ollington[1[0000-0001-7533-2307]]

[1] University of Tasmania
simon.stanton@utas.edu.au

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 informing and so reifying internal state. This paper 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 particular 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.

Keywords: Algorithmic Bias, Cooperative Behaviour, Game Theory

1 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. This is relevant to how the agent's internal state representation of the signals received are abstracted.

The context of this research is an investigation of a computational approach to agent dynamics in a complex system by adopting the Robinson and Goforth (2005) topological classification of gameforms. Complexity is understood here as the interplay of agents and environment, as per Arthur's definition: "systems responding to the context they create" (2019, p186). The Robinson and Goforth (2005) classification is defined in-part topologically, with a core set of group operations that link games into regions, families, and neighbours of similar and/or aligned game-theoretic properties. Furthermore, it absorbs much previous work on game theory typologies and taxonomies such as those developed by Rapoport, Gordon and Guyer (1976) and Brams (1994).

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 those 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.

The findings of this paper suggest that if differing representations induce a discrete and non-identical reachable space in respect to the values that internal agent state may take, then this will in turn circumscribe the possible behaviours that an agent can embody. Agents experiencing complex environments may display less predictable behaviour. We would regard such an outcome as a vector for *representation-induced algorithmic bias*.

Vectors for the introduction of bias—in algorithms specifically, and more generally in software—have been and are discussed widely. For example: bias via prejudice (Angwin et al., 2016; Patton et al., 2017); via both design and ethical formulations of systems (Hooker, 2021; Bryson, 2018; Winfield et al., 2019); through higher-order computation (Hooker et al., 2020; Waller & Waller, 2021); from inferential methods (Tversky and Kahneman, 1974); via our lack of understanding of animal intelligence (Herzing, 2014); and, perhaps non-intuitively, via a lack of consideration to non-human machine intelligence and behaviour—a view put forward by Rahwan et al. (2019).

Therefore, we wish to investigate the role of representation in gameforms to understand the importance of representation *to* learning algorithms. To encourage parsimony and tractability we adopt the highly constrained environment of the repeated stage game Prisoner's Dilemma as the experimental domain. We regard any derivation of a scalar- or ordinal-valued gameform to an isomorphic (structure-preserving) representation as equivalent in form, subject to satisfying the four social dilemma inequalities discussed below. The response variable of interest is the *variance in the aggregated distribution of cooperative game outcomes between representations*, as observed in a symmetric selfplay paired-parameter study of the 14 foundational reinforcement learning algorithms listed in **Table 1**. The aggregated distribution of outcomes attained from a single gameform representation is the *behavioural profile* of the algorithm under study, for that gameform representation. We also identify the peak cooperative outcome from each behavioural profile to identify the optimal performance of each algorithm, per gameform representation.

Our initial hypothesis then, is that an algorithmic behavioural profile *will not vary between equivalent representations of the game* (environment).

Table 1. Algorithms implemented in this study and their parameters. Parameter superscript [1,2] denotes paired parameters, computed in range (0,1] with increment of 0.1 giving 100 observation sets for each algorithm, per gameform representation.

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		$\varepsilon = 0.1$

Background. This work lies at the intersection of computational learning, cooperation, and game theory. Research in computational learning ranges from learning in cellular automata (Billard, 1996; Grim, 1996), to applications of associative learning automata (Barto et al., 1983; Barto and Anandan, 1985), and Reinforcement Learning (Sutton and Barto, 1998, 2018). Furthermore, Tan's predator-prey agents using Multi-Agent Reinforcement Learning (1993), and more recently, work by Leibo et al. on Deep Reinforcement Learning in sequential social dilemmas (2017) also inform this research. Extensive recent work by Crandall et al. involving a variety of learning algorithms (2018a, 2018b) over the Robinson and Goforth topology (2005) in a tournament setting has provided invaluable insight to the problem domain in a contemporary setting.

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**.

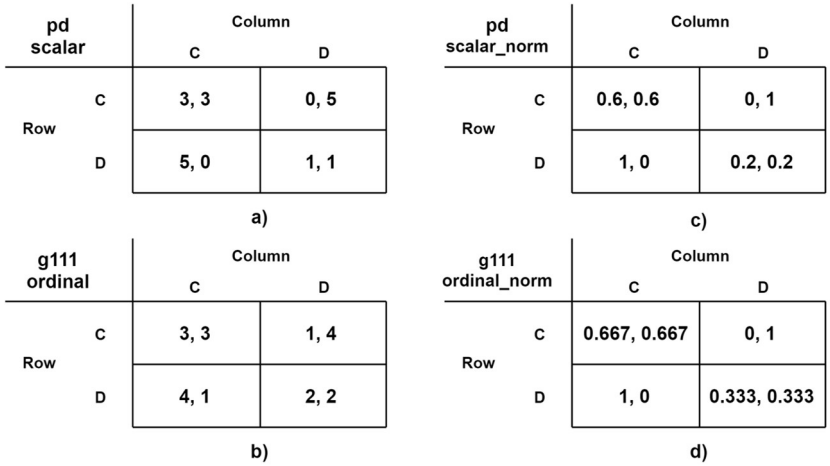

Fig. 1. Four gameform representations: a) canonical Prisoner's Dilemma with scalar values; b) scalar values are replaced by the payoff's ordinal preference; c) scalar values are normalised; d) ordinal preference values are normalised.

To assess formal equivalence, we restrict analysis in this paper to a formal definition of social dilemma inequalities as they provide a means to assert the preservation of the structure of the game representation under various transformations while retaining the meaning attributed to the dilemma. Each of the four inequalities relate to the four game outcomes as shown in **Fig. 2**.

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:

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

Fig. 2. Semantic labels attributed to outcomes in the Prisoner's Dilemma. R (Reward), S (Sucker), T (Temptation), and P (Punishment) single-letter labels are outcomes from the perspective of the Row player. Labels in parentheses are Row, Column ordered pair where D signifies Defect and C signifies Cooperate.

$$R > P \tag{1}$$

$$R > S \tag{2}$$

$$2R > T + S \tag{3}$$

$$T > R, or\ P > S \tag{4}$$

We establish the formal equivalence of **pd:scalar** with the three alternative representations under examination in these experiments 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)	$T > R$	$5 > 3$	$1 > 0.6$	$4 > 3$	$1 > 0.667$	$True$
	$P > S$	$1 > 0$	$0.2 > 0$	$2 > 1$	$0.333 > 0$	$True$

2 Methods

The experiments in this 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.

As shown in **Fig. 1**, this work examines four representations of the Prisoner's Dilemma gameform: *scalar*, *normalised scalar*, *ordinal*, and *normalised ordinal*. Given the infinite space of possible transformation functions, we do not examine any beyond these four common mappings.

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 is referred to as *experiment groups 1* through *4* in the remainder of this work.

The 14 learning algorithms implemented are listed in **Table 1**. These algorithms 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 with the exception of 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. Given that R-Learning does not use γ, we test on both learning rate parameters—α and β—instead. Fixed values for other parameters in each algorithm are as commonly found in the literature and are detailed in **Table 1**.

An algorithm's memory depth (history of outcomes from previous timesteps) is configured by a *memory_depth* option which is set to 1 in all algorithms (however, for Watkins Q Linear Function Approximation this is implemented as a feature vector of the outcomes from the last timestep only, plus an associated weight vector). Some algorithms incorporate *trace* data structures, which can be thought of as weights over states on longer timescales than just the immediate past. In other cases, the algorithm's data structures—such as Q-tables—are limited to handling state representation with memory of only the last timestep. This restriction on memory depth has implications for performance. However, the intent of this study is not to optimize all available parameters—nor to necessarily obtain the absolute optimal frequency of cooperation—but to examine the behaviour of the algorithms over a subset of their possible states. Implicitly, each algorithm is attempting to maximise its reward, and therefore, frequency of the CC outcome.

The data from each run is analysed in a two-stage process. The first stage takes the raw episodic data and calculates aggregated and summary statistics on episode, and timestep, actions and rewards; on game outcomes; and also extracts CPU-time and memory usage metadata. Subsequent analysis performs statistical tests on the outputs of the first pass. 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.

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, $T = \infty$; 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.

3 Results

The results of analysis of all fourteen algorithms across each of the gameform representations are presented as a set of four experiment groups, as described in **Methods**. 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).

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

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.

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

Algorithm	Peak % CC		Wilcoxon			
	S	O	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

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.

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

Algorithm	Peak % CC		Wilcoxon			
	S	NS	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

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.

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

Algorithm	Peak % CC		Wilcoxon			
	S	NO	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

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.**

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

Algorithm	Peak % CC		Wilcoxon			
	O	NO	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

4 Discussion

The results of these experiments indicate that the behaviour of the algorithms studied can vary substantially as a product of the input representation of an otherwise equivalent gameform. Given the common use of semantic interpretations of repeated game outcomes the validity of the expectation that conclusions drawn over varying representations—even those that conform to the social dilemma inequalities—should be invariant with respect to behaviour is thus unclear. Rather, it may be that the domain of possible values, or reachable space, accessible to the internal state of each algorithm varies under different representations.

With regard to practical impacts on the use of learning algorithms, these results suggest that we cannot unconditionally generalise learning algorithms between representations, either *a priori*, or *in-situ*, without possibly introducing bias. Within a representation we assert that we can compare algorithms—with the proviso that the given representation may define the depth and location of local minima. In addition, we are not assured that the behaviours available to an algorithm, given the reachable space, will conform to intuitive expectations derived from a semantic interpretation of the dilemma. Furthermore, we may ask how much the variance in behaviour actually affects an algorithm—does the range of an outcome's frequency shift only slightly, or is the effect more pronounced? We assert that the character of the behavioural profile can be altered considerably, as is evident in the algorithm Watkins (naïve) Q Lambda in experiment group three, shown in **Fig. 3**.

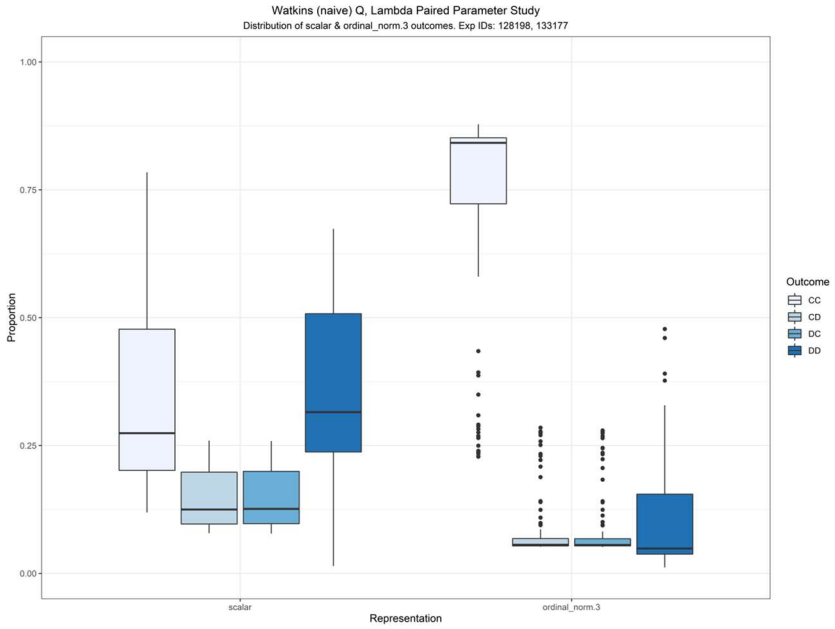

Fig. 3. Grouped boxplot of experiment group three outcomes for Watkins (naive) Q, Lambda.

We conclude that the null hypothesis—that the behavioural profile will not vary between equivalent representations of the game—is not supported for the majority of the learning algorithms, as can be seen in **Table 7**, where only Watkins Q Linear Function Approximation exhibits non-significant variance in all experiment groups.

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	Three	Double Q-Learning
	Actor/Critic with Replacing Traces		**Watkins Q, Linear Function Approximation**
	Watkins Q, Linear Function Approximation		
Two	Actor/Critic	Four	Actor/Critic with Replacing Traces
	Actor/Critic with Replacing Traces		SARSA Lambda
	Q-Learning		SARSA Lambda, with Replacing Traces
	Double Q-Learning		Watkins (naïve) Q, Lambda
	Expected SARSA		Watkins (naïve) Q, Lambda, Replacing Traces
	R Learning		Watkins Q, Lambda
	SARSA		**Watkins Q, Linear Function Approximation**
	SARSA Lambda, with Replacing Traces		
	Watkins Q, Lambda		
	Watkins Q, Linear Function Approximation		

The result for Watkins Q Linear Function Approximation is somewhat weak in experiment group four (V=44057, p-value=.07136). Regardless, the ability of Watkins

Q Linear Function Approximation to maintain equivalence across these four representations directs our future work in this area towards contemporary policy-gradient reinforcement learning algorithms.

Acknowledgements. We would like to acknowledge the use of the high-performance computing facilities provided by the Tasmanian Partnership for Advanced Computing (TPAC) funded and hosted by the 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, and further supplementary material, is available at https://github.com/simoncstanton/equivalence_study.

References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). *Machine Bias.* ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, Accessed 2021-07-05

Arthur, B. F. (2019). *Complexity Economics: Why Does Economics Need This Different Approach?* Complexity Economics: Proceedings of the Santa Fe Institute's 2019 Fall Symposium. Santa Fe Institute 2019 Fall Symposium, Santa Fe Institute.

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

Barto, A. G., & Anandan, P. (1985). *Pattern-Recognizing Stochastic Learning Automata.* doi:10.1109/tsmc.1985.6313371

Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). *Neuronlike adaptive elements that can solve difficult learning control problems.* IEEE Transactions on Systems, Man, and Cybernetics, SMC-13(5), 834–846. doi:10.1109/TSMC.1983.6313077

Billard, E. A. (1996). *Adaptation in a stochastic prisoner's dilemma with delayed information.* Biosystems, 37(3), 211–227. doi:10.1016/0303-2647(95)01560-4

Brams, S. J. (1994). *Theory of Moves.* Cambridge University Press.

Bryson, J. J. (2018). *Patiency is not a virtue: The design of intelligent systems and systems of ethics.* Ethics and Information Technology, 20(1), 15–26. doi:10.1007/s10676-018-9448-6

Crandall, J. W., Oudah, M., Tennom, Ishowo-Oloko, F., Abdallah, S., Bonnefon, J.-F., Cebrian, M., Shariff, A., Goodrich, M. A., & Rahwan, I. (2018a). *Cooperating with machines.* Nature Communications, 9(1), 233. doi:10.1038/s41467-017-02597-8

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

Grim, P. (1996). *Spatialization and greater generosity in the stochastic Prisoner's Dilemma.* Biosystems, 37(1), 3–17. doi:10.1016/0303-2647(95)01541-8

Herzing, D. L. (2014). *Profiling nonhuman intelligence: An exercise in developing unbiased tools for describing other "types" of intelligence on earth.* Acta Astronautica, 94(2), 676–680. doi:10.1016/j.actaastro.2013.08.007

Hooker, S. (2021). *Moving beyond "algorithmic bias is a data problem".* Patterns, 2(4). doi:10.1016/j.patter.2021.100241

Hooker, S., Moorosi, N., Clark, G., Bengio, S., & Denton, E. (2020). *Characterising Bias in Compressed Models.* https://arxiv.org/abs/2010.03058v2, Accessed 2021-06-28

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

Patton, D. U., Brunton, D.-W., Dixon, A., Miller, R. J., Leonard, P., & Hackman, R. (2017). *Stop and Frisk Online: Theorizing Everyday Racism in Digital Policing in the Use of Social Media for Identification of Criminal Conduct and Associations.* Social Media + Society, 3(3), 2056305117733344. doi:10.1177/2056305117733344

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A., Roberts, M. E., Shariff, A., Tenenbaum, J. B., Wellman, M. (2019). *Machine behaviour.* Nature, 568(7753), 477. doi:10.1038/s41586-019-1138-y

Rapoport, A., Guyer, M., & Gordon, D. G. (1976). *The 2 X 2 game.* University of Michigan Press.

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). *Algorithms for Reinforcement Learning.* Morgan and Claypool Publishers.

Tan, M. (1993). *Multi-Agent Reinforcement Learning: Independent vs. Cooperative Agents.* In Proceedings of the Tenth International Conference on Machine Learning, 330–337.

Tversky, A., & Kahneman, D. (1974). *Judgment under Uncertainty: Heuristics and Biases.* Science, 185(4157), 1124–1131. doi:10.1126/science.185.4157.1124

Waller, R. R., & Waller, R. (2021). *The machine mind: Beyond transparent biases.* Paper presented at Kinds of Intelligence Workshop Series: Cognitive Science Beyond the Human, Leverhulme Centre for the Future of Intelligence. http://lcfi.ac.uk/projects/kinds-of-intelligence/, June 25, 2021.

Winfield, A. F., Michael, K., Pitt, J., & Evers, V. (2019). *Machine Ethics: The Design and Governance of Ethical AI and Autonomous Systems.* Proceedings of the IEEE, 107(3), 509–517. doi:10.1109/JPROC.2019.2900622