Entwicklung und Validierung eines KI-basierten Markovagenten zur Interaktion mit Menschen in wiederholten nichtkooperativen Spielen

Zur Erlangung des akademischen Grades eines
Doktors der Wirtschaftswissenschaften

(Dr. rer. pol.)

von der KIT-Fakultät für Wirtschaftswissenschaften
des Karlsruher Instituts für Technologie (KIT)

genehmigte

DISSERTATION

von

M.Sc. Wi.-Ing. Florian Müller

Tag der mündlichen Prüfung:	27. September 2021
Referent:	Prof. Dr. Hagen Lindstädt
Korreferent:	Prof. Dr. Michael Wolff
Veröffentlichung:	2021

Inhaltsverzeichnis

Detailliertes Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungs- und Symbolverzeichnis

Abkürzungen

AgentM	Name des allgemeinen entwickelten Markovagenten
AgentM01	Markovagent mit Parametrisierung $\Omega^{j}_{AgentM01} = \{(0,1)\}$
AgentM11	Markovagent mit Parametrisierung $\Omega^{j}_{AgentM11} = \{(1,1)\}$
AgentMx1	Markovagent mit Parametrisierung $\Omega^{j}_{AgentMx1} = \{(0,1),(1,1)\}$
AI	Artificial Intelligence (siehe auch KI)
Ar	Assurance Game
Aufl.	Auflage
AWESOME	Adapt When Everybody is Stationary, Otherwise Move to Equilibrium
Ba	Battle Game
Bd.	Band
bspw.	beispielsweise
bzw.	beziehungsweise
C	*Kooperation* im Prisoner's Dilemma
Ch	Chicken Game
Cm	Compromise Game
Co	Coordination Game
D	*Abweichung* im Prisoner's Dilemma
DFA	Deterministischer Finiter Automat
Dl	Deadlock Game
Ges.	Gesamt
GIGA	Generalized Infinitesimal Gradient Ascent
Ha	Harmony Game
Hr	Hero Game
Hrsg.	Herausgeber
IGA	Infinitesimal Gradient Ascent
IP-EMS	Individual Prediction based on Estimation of Markov Strategies
IT-US-L*	Iterative Unsupervised-L*
KI	Künstliche Intelligenz
MAE	Mean Absolute Error

MAL	Multi-Agent Learning
N	Nash-Gleichgewicht
Nc	Concord/No Conflict Game
NvA	Normalverteilungsannahme
P	Paretoeffizienter Zustand
Pc	Peace Game
Pd	Prisoner's Dilemma
POI	Proportion of Inaccuracy
RMSE	Root Mean Squared Error
S.	Seite
SA	Symmetrieannahme
SAL	Single-Agent Learning
Sh	Stag Hunt
TFT	Tit-for-Tat
TFTr	Tit-for-Tat mit Rauschen
UA	Unabhänigkeitsannahme (Unabhängigkeit der Stichprobenbeobachtungen)
US-L*	Unsupervised-L*
vgl.	vergleiche
Wiss.	Wissenschaften
WoLF	Win or Learn Fast
z.B.	zum Beispiel

Alphanumerische Symbole

$A = A^1 \times ... \times A^n$	Aktionsraum des Spiels
$A^i = \{a^i_{1...\|A^i\|}\}$	Endliche Aktionsmenge für Spieler i (reine Strategien)
a^i	Spezifische Aktion für Spieler i
B^i	Auswahlfunktion zur Bestimmung der besten Antwort für s^j und U^i für Spieler i
$C_(z,\emptyset)$	Funktion zur Berechnung der absoluten Häufigkeit des Zustandes z in einer Stichprobe D
$C_(z,a)$	Funktion zur Berechnung der absoluten Häufigkeit von Aktion a in Zustand z in einer Stichprobe D
$D = X_1 - X_2$	Differenz zweier Zufallsvariablen
$D^j(h(t))$	Durch Spieler i beobachtete Stichprobe zu Spieler j
$d[M_Z, M_Z]$	Euklidische Distanz zwischen zwei Übergangsmatrizen
$\mathbf{E}[...]$	Erwartungswert einer Zufallsvariable
F_z	Funktion zur Berechnung der empirischen relativen Häufigkeit von Aktion a^j_2 in Zustand z in einer Stichprobe D

G	Stufenspiel in Normalform für zwei Spieler mit A^1, A^2, u^1, u^2
$G^{\#}$	Wiederholtes Spiel in Normalform für zwei Spieler mit S^1, S^2, U^1, U^2
H_0	Nullhypothese
$H(G^{\#})$	Menge der Historien für das wiederholte Spiel $G^{\#}$
$\hat{H}(G^{\#})$	Menge der Pfade für das wiederholte Spiel $G^{\#}$
$h(t)$	Historie des wiederholten Spiels in Runde t
$\hat{h}_{(s^1,s^2)}(t)$	Pfad des wiederholten Spiels für das Strategiepaar (s^1, s^2) in Runde t
I	Menge der simulierten Interaktionen
$i = 1..N$	Spielerindex
$j = 1..N : j \neq i$	Gegnerindex
k	Zählvariable
L^i	Der das Gegnermodell $\hat{s}^j$ bildende Lernalgorithmus von Spieler i
$ln(L)$	Log-Likelihood Schätzer
$M^i : Z^i \rightarrow \mathbf{P}[A^i]$	Übergangsmatrix einer Markovstrategie von Spieler i
$\hat{M}^j$	Modell des Spielers i der gegnerischen Übergangsmatrix M^j
max	Maximalwertoperator
min	Minimalwertoperator
N	Anzahl der Spieler
n	Stichprobengröße
$O^i = (o^i_i, o^i_j)$	Nach Spieler differenzierte Gedächtnistiefe oder Markovordnung von Spieler i
$\hat{O}^j = (\hat{o}^j_j, \hat{o}^j_i) \in \Omega^j$	Modell des Spielers i der gegnerischen Gedächtnistiefe O^j
o^i	Gedächtnistiefe oder Markovordnung von Spieler i
$\mathbf{P}[...]$	Wahrscheinlichkeit eines Ereignisses
p	p-Wert eines statistischen Prüfgröße
Q^i	Auswahlfunktion zur Bestimmung des glaubwürdigsten Gegnermodells $\hat{s}^j$ $in\hat{S}^j$ für Spieler i
$R^i(G)$	Auszahlungsvektor von Spiel G und Spieler i für jede Aktionskombinationen (a^i, a^j)
r^i_t	Realisierte Auszahlung für Spieler i in Runde t
S^i	Endliche Strategiemenge für Spieler i
s	Im Rahmen von statistischen Analysen: Standardabweichung der Stichprobe
$s^i \in S^i$	Strategie für Spieler i
$s^{*i}(s^j, U^i)$	Optimale Strategie für Spieler i
$\hat{s}^j$	Modell des Spielers i der Gegnerstrategie s^j
T	Anzahl der Runden eines wiederholten Spiels
$T_0 < T$	Anzahl der Runden einer Explorationsphase

t	Im Rahmen von statistischen Analysen: t-Statistik
t	Runde eines wiederholten Spiels
t_0	Erste Runde eines wiederholten Spiels
$U^i : S^1 \times S^2 \to \mathbb{R}$	Nutzenfunktion für Spieler i im wiederholten Spiel
$u^i : A^1 \times A^2 \to \mathbb{R}$	Nutzenfunktion für Spieler i im Stufenspiel
V	Testparameter Shapiro Wilk Tests
W	Testparameter des Shapiro Wilk Tests
X	Zufallsvariable
x	Hilfsvariable
$\bar{x}$	Mittelwert der Stichprobe
$\tilde{x}$	Median der Stichprobe
y	Hilfsvariable
Z^i	Endliche Menge möglicher Markovzustände für Spieler i
z	Im Rahmen von statistischen Analysen: z-Statistik
$z^i \in Z^i$	Markovzustand für Spieler i
γ_0	Gewichtungsfaktor des Priors $\hat{M}_0$
γ_F	Gewichtungsfaktor des der empirischen relativen Häufigkeit F_z
ε	Allgemeiner Fehlerterm
$\Theta^i(t)$	Kumulierte Zustandswahrscheinlichkeiten für i bis t
$\theta^i(t)$	Zustandswahrscheinlichkeiten für i in t
$\bar{\rho}^i$	Normierte durchschnittliche realisierte Auszahlung für Spieler i
σ^i	Initialisierungslogik einer Markovstrategie von Spieler i
$\hat{\sigma}^j$	Modell des Spielers i der gegnerischen Initialisierungslogik σ^j
τ^j	Aktionsspeicher als maximale Länge der Stichprobe $D^j_{O^j,\tau^j}(h(t))$ je Zustand z^j in $h(t)$
χ^2	χ^2-Statistik
Ψ^i	Transitionsmatrix aus Sicht von i
Ω^j	Menge der von Spieler i zur Modellierung von j in Betracht gezogene Gedächtnistiefen

Weitere Symbole

$\emptyset$	Leere Menge; beziehungsweise leeres Tupel
∞	Unendlichkeitssymbol
$(\ldots)^T$	Transponierte Matrix
$\|\|$	Verknüpfungsoperator
✓	Annehmen der Nullhypothese
$\times$	Verwerfen der Nullhypothese
†	Indikativ, knappes Nichterreichen des niedrigen Signifikanzniveaus

*	Erfüllen des niedrigen Signifikanzniveaus
**	Erfüllen des mittleren Signifikanzniveaus
***	Erfüllen des hohen Signifikanzniveaus

1 Einleitung

Im Anschluss einer kurzen thematischen Einordnung und Motivation der Arbeit, zeigt dieses Kapitel den Forschungsgegenstand auf und präsentiert abschließend den Aufbau des Werkes.

1.1 Hintergrund und Motivation

Interaktionen zwischen Akteuren im Kontext von ökonomischen Zielsetzungen sind zentraler Betrachtungsfokus der Wirtschaftswissenschaften, insbesondere der Spieltheorie. Von herausragendem Interesse sind dabei aufgrund ihrer Reichhaltigkeit solche Interaktionen, die sich über einen gewissen Zeitraum erstrecken und als wiederholte Spiele formalisiert werden können (vgl. Camerer, 2011). Zentrale und folgerichtige Kernfrage ist hierbei, wie sich Agenten in derartigen Interaktionen optimaler Weise verhalten sollen. Traditionell nähert sich die Forschung dieser Frage mit der auf Erwartungsnutzentheorie (vgl. Arrow, 1976; M. Friedman & Savage, 1948) basierenden Gleichgewichtsanalyse unter Annahme perfekter Rationalität (vgl. z.B. Rubinstein, 1979). Das klassische Nash-Gleichgewicht, Teilspielperfektheit und das Folk Theorem sind prominente Vertreter dieser Strömung (vgl. Fudenberg & Maskin, 1986; Nash, 1950; Selten, 1965). Die axiomatische Annahme *perfekter Rationalität* stößt bei empirischen Untersuchungen jedoch schnell an ihre Grenzen, da sie die Realität menschlichen Handelns übersteigt und sich daher weniger gut zu dessen Vorhersage eignet (vgl. Allais, 1979; Arthur, 1994; Ellsberg, 1965; Wright & Leyton-Brown, 2010). Häufig angeführte Gründe für das Scheitern perfekter Rationalität sind mangelnde kognitive Kapazität, die Abhängigkeit des Optimalitätsbegriffes von der Gegnerstrategie und die Mehrdeutigkeit von multiplen Gleichgewichten (vgl. Wright & Leyton-Brown, 2010, S. 901). Als Alternativkonzept trägt *beschränkte Rationalität* mit ihrer verhaltenswissenschaftlichen Ausrichtung daher der praktischen aber nicht logischen Unmöglichkeit Rechnung, perfekt rationale Entscheidungen zu fällen (vgl. Simon, 1955), wobei sich das Feld durch ein hohes Maß an Heterogenität auszeichnet (vgl. z.B. Kahneman & Tversky, 1979; Lindstädt, 2006; Roth, 1996; Simon, 1955; Tversky, 1996).

Menschliche Individuen verhalten sich in gleichen Situationen potentiell unterschiedlich (vgl. Erev & Roth, 1998, S. 859). Klar ist, dass eine identische Spielhistorie bei zwei unterschiedlichen Menschen zu einer unterschiedlichen nächsten Spielrunde führen kann. Ein wesentlich schwerer beherrschbarer Sachverhalt ist jedoch, dass eine identische Spielhistorie bei demselben menschlichen Individuum zu unterschiedlichen nächsten Spielrunden führen kann. Vor die-

sem Hintergrund gestaltet sich die Vorhersage menschlichen Verhaltens als nichttriviales Problem. Soll darüber hinaus eine performante Gegnerstrategie für die Interaktion mit menschlichen Agenten bestimmt werden, gipfelt die vorliegende Komplexität darin, dass zusätzlich die optimale Antwort auf eine spezifische aber unbekannte Gegnerstrategie selbst bei einer deterministischen Beziehung zwischen Spielhistorie und menschlichem Aktionsverhalten ex ante nicht bestimmt werden kann, da das optimale eigene Verhalten stets vom nur punktuell, nur über Zeit und nur unter Billigung von Lernkosten, beobachtbaren Gegnerverhalten abhängt. Vor diesem Hintergrund ist das Vorhaben, einen leistungsfähigen lernenden Agenten für das Spiel mit Menschen zu programmieren von einer vielschichtigen Problemstruktur gekennzeichnet.

Lernalgorithmen stellen einen grundsätzlich validen Lösungsansatz für die beschriebene Herausforderung dar. Getrieben durch zunehmende Rechenleistung sowie gesteigerte wissenschaftliche Relevanz drängen Lernalgorithmen zunehmend in die Debatte zur Untersuchung von Interaktionen zwischen (menschlichen) Akteuren (vgl. Shoham et al., 2007). Dabei stellt das *Multi Agent Learning* (MAL), also dem Lernen unter Berücksichtigung der Präsenz anderer lernender Akteure, aufgrund fachlicher Überschneidungen zwischen Informatik und Spieltheorie in Bezug auf Lösungs- und Gleichgewichtskonzepte eine forschungsrelevante Schnittstelle dar (vgl. Shoham & Powers, 2014a). Aufgrund des interdisziplinären Charakters des rapide wachsenden Feldes ergibt sich jedoch eine Plethora an verzweigenden Lösungskonzepten, deren Strukturierung aufgrund regelmäßig impliziten Zielsetzungen schwer fällt (vgl. T. Sandholm, 2007; Shoham et al., 2007).

Shoham et al. (2007) arbeiten die *nichtkooperative präskriptive Agenda* als eine von fünf MAL Strömungen als jene heraus, welche sich mit optimalem Verhalten in wiederholten Spielen und somit der eingangs formulierten Fragestellung aus einer informatischen Perspektive nähert. Der Sachverhalt, dass die Qualität der eigenen Strategie im Sinne eines Planes für alle Möglichkeiten bei Unkenntnis der generischen Strategie nicht beurteilt werden kann, wird im Kontext des MAL insbesondere dadurch verstärkt, dass auch andere Agenten einen Lernprozess durchlaufen, sodass die Lernumgebung nicht länger als stationär betrachtet werden kann (T. W. Sandholm & Crites, 1996). Die Modellierung sich verändernden Verhaltens ist aufgrund des unbegrenzten Möglichkeitenfeldes zukünftiger Verhaltensweisen herausfordernd (vgl. Albrecht & Stone, 2018) und weder die Problemstellung, noch die Algorithmen des MAL leiten sich aufgrund dieser Interdependenz direkt aus dem *Single Agent Learning*[1] (SAL) Fall ab (Shoham & Powers, 2014a). Das übergeordnete Ziel präskriptiver MAL Algorithmen ist Leistungsfähigkeit entsprechend einer definierten Zielgröße. Folglich ist es elementar, die Zielgröße klar zu formulieren. Die häufig herangezogene Konvergenz der empirischen Verteilung gegen Gleichgewichte scheint hier nicht zielführend, da diese in Abhängigkeit des Gegners suboptimale Aus-

[1] Single Agent Learning bezeichnet das Lernen in einer Umwelt, welche nicht durch die Präsenz anderer lernender Agenten charakterisiert ist.

zahlungsergebnisse produzieren kann und Stabilität in realen Interaktionen nicht notwendigerweise gegeben ist (T. Sandholm, 2007). Zusammenfassend sind präskriptive MAL Algorithmen nicht nur technisch, sondern auch konzeptionell herausfordernd und erfordern Klarheit bezüglich des untersuchten Problems und der herangezogenen Erfolgskriterien (vgl. Shoham et al., 2007). Dabei sind nichtkooperative Interaktionen dank der inhärenten Konflikte komplexer als kooperativ-koordinative Interaktionen und nichtkooperative Nichtnullsummenspiele wiederum eine größere Herausforderung als Konstantsummenspiele (vgl. Littman, 1994; Ortega & Legg, 2018).

In der Literatur wird eine breite Palette an Lösungskonzepten für nichtkooperative präskriptive Agenten angeboten, deren Leistung nicht immer zufriedenstellend ist. Die aus dem SAL stammenden Formate des Reinforcement Learnings lassen sich nicht ohne Weiteres auf Mehragentensysteme übertragen (vgl. Shoham et al., 2007).[2] Im MAL Kontext eignen sie sich zwar dennoch für Nullsummenspiele, da der Einfluss des Gegners auf Überlegungen bezüglich des eigenen Verhalten vergleichsweise gering ausfällt (vgl. Littman, 1994, 1996), wohingegen sie für Nichtnullsummenspiele (vgl. Greenwald & Hall, 2003; Hu & Wellman, 1998; Littman, 2001) weniger gut abschneiden.[3] Andere Algorithmen beschränken sich im Falle bedauernsminimierender Ansätze dagegen bei ihrer Zielfunktion auf die Absicherung einer Mindestauszahlung (vgl. Fudenberg & Levine, 1995, 1998), während andere Ansätze aufgrund von Gleichgewichtsbestrebungen nur implizite Aussagen über erzielbare Payoffs treffen (vgl. Bowling & Veloso, 2001). Lediglich der Lösungsvorschlag von Powers und Shoham (2005b) greift eine Maximierung der eigenen Auszahlung als Erfolgskriterium erfolgreich auf.

Eine Gemeinsamkeit ist, dass die vorgenannten Konzepte nicht in der Lage sind, reichhaltigere Interaktionen im Sinne *sophistizierter Rationalität* im MAL Kontext abzubilden (vgl. Powers & Shoham, 2005a, 2005b). Sophistizierte Rationalität beschreibt, dass Menschen empirisch bisweilen einen Mittelweg zwischen perfekter und beschränkter Rationalität einzuschlagen scheinen. Dabei sind sie sich über adaptive Prozesse des Gegners und den Einfluss des eigenen Handelns auf das gegnerische Handeln bewusst und beziehen derartige Überlegungen in ihre eigene Entscheidungsfindung mit ein (vgl. Camerer, 1997; Milgrom & Roberts, 1991). Infolgedessen agieren menschliche Agenten häufig auf Basis von auf strategische Überlegungen zurückzuführende Entscheidungsregeln (vgl. Camerer, 2004), welche die Kapazität einfacher Lernmodelle auf Basis von Stationaritätsüberlegungen übersteigt (vgl. Erev & Haruvy, 2016, S. 684). Tatsächlich bedingen Menschen, wenn sie gebeten werden ihre Entscheidungsregeln explizit zu formulieren, regelmäßig auf eine kürzliche Partition der Aktionshistorie (vgl. Dal Bo & Fre-

[2] Es sei beispielsweise auf Graf (2018) verwiesen, der exemplarisch zeigt, wie Q-Learner grundsätzlich die beste Antwort auf eine stationäre Gegnerstrategie lernen können, wenn der Zustandsraum des Q-Learners adäquat vordefiniert wurde. Die Ergebnisse deuten darauf hin, dass sich die Anwendung jenseits stationärer Gegner aufgrund einer unverhältnismäßig langen Lernperiode als ineffizient gestaltet.

[3] In Nullsummenspielen ist jeder Zustand paretoeffizient.

chette, 2013, S. 1). Diese Eigenschaft wird jedoch nur von Powers und Shoham (2005a) mit dem *Manipulator* Algorithmus und von Carmel und Markovitch (1996, 1998) mit dem *IT-US-L** Algorithmus aufgegriffen, wobei letzterer jedoch ausschließlich deterministisches Gegnerverhalten abbilden kann. Weiterhin zeichnen sich die beiden Lösungen gerade dadurch aus, dass sie erstens lediglich gegen algorithmische Spieler getestet wurden und zweitens eine hochgradig impraktikable und lange Explorationsphase benötigen.

Zusammenfassend gestalten sich Modellierung und Bewertung des reziproken oder interaktiven Lernprozesses in einer MAL Umgebung komplex. Der Literaturkorpus zu dem Thema ist aufgrund variierender Lösungsansätze und häufig impliziten Zielsetzungen und Kriterien der Erfolgsmessung bisweilen unübersichtlich (vgl. Shoham & Powers, 2014a). Der aktuelle Forschungsstand bietet eine herausragende Gelegenheit für Untersuchungen, welche sich der Entwicklung und Validierung effizienter Methoden der Künstlichen Intelligenz (KI) zur Interaktion mit menschlichen Akteuren bei Präsenz von Interessenskonflikten, welche Aspekte wie sophistizierte Rationalität und Kosten einer Explorationsphase zielführend berücksichtigt. Die Relevanz der Thematik wird aufgrund zunehmender Mensch-Maschine-Interaktionen und fortschreitender Informationstechnologie unterstrichen.

1.2 Zielsetzung und Methodik

Zielsetzung dieser Arbeit ist die Entwicklung eines präskriptiven interaktiven Agenten für nichtkooperative Interaktionen mit Menschen, welcher in der Lage ist, reichhaltige Interaktionen sophistizierter Rationalität informationseffizient abzubilden, ohne auf überproportionale Explorationsphase zurückgreifen zu müssen. Insbesondere soll der Agent eine über verschiedene Spiele hinweg anwendbare Lösung und im Gegensatz zu z.B. Tit-for-Tat keine Speziallösung für bestimmte 2x2 Spiele darstellen (vgl. Axelrod, 1984). Wesentliches Augenmerk liegt dabei auf der umfassenden Validierung der Auszahlungsleistung der entwickelten Lösung, welche in der MAL Literatur häufig nicht explizit, nur formaltheoretisch oder empirisch nur gegen algorithmische Spieler stattfindet (vgl. Powers & Shoham, 2005a; T. Sandholm, 2007; Shoham et al., 2007). Daher strebt diese Arbeit eine Bewertung sowohl im Laborexperiment gegen Menschen, als auch analog zu Axelrod (1980) im Turnier gegen Algorithmen und in Bezug auf formale Kriterien an.[4] Axelrod (1980) beschränkte sich lediglich auf Turniere mit Algorithmen. Die Logik des Agenten gliedert sich bei in zwei Teilschritte:

1. Vorhersage des Gegnerverhaltens

[4] Es sei herausgestellt, dass die Zielfunktion zur Leistungsbewertung *nicht* darauf abzielt, eine höhere Auszahlung *als der Gegner* zu erzielen. Stattdessen ist wesentliches Leistungsmerkmal, die eigene Auszahlung in einem Turnier zu maximieren. Zentrale Eigenschaft einer derartigen Betrachtung ist, dass die erzielbare Auszahlung von der Grundgesamtheit der Turnierteilnehmer abhängt. Eine von der Gegnerpopulation losgelöste empirische Leistungsbetrachtung ist insofern nicht möglich.

2. Bestimmung der eigenen Antwort

Für den ersten Teilschritt bieten die herausragenden empirischen Ergebnisse von Müller (2018) zur effizienten Vorhersage von individuellen menschlichen Verhalten auf Basis von Markovstrategien (vgl. Breitmoser, 2015) einen vielversprechenden Anknüpfungspunkt für einen modellbasierten generalisierbaren MAL Agenten für wiederholte Spiele, der weiterentwickelt werden soll. Bezüglich des zweiten Teilschritts stellt diese Arbeit eine simulationsbasierte Lösung vor.

Primäres Ziel ist folglich die Entwicklung eines auf Markovstrategien zurückgreifenden modellbasierten KI Agenten sowie die Untersuchung seiner Leistung im Spiel gegen menschliche Gegner. Methodisch soll der deskriptive Markovansatz von Müller (2018) auf das interaktive Anwendungsfeld präskriptiver Algorithmen angepasst und in Hinblick auf performantes Spielverhalten optimiert werden. Dazu fällt das Augenmerk zunächst auf die Erarbeitung einer konzeptionellen Lösung, wodurch eine größtmögliche Transparenz gewährleistet werden soll. Maßgebliche Eigenschaft des zu entwickelnden Markovagenten soll sein, dass dessen Leistungsfähigkeit spielunabhängig ist und er in jedem beliebigen 2x2 Spiel performant interagieren kann. Es soll sich um keine, auf einen engen Anwendungsraum zugeschnittene, Speziallösung handeln. Der entwickelte Markovagent soll daher anschließend einer umfassenden experimentellen Validierung im Rahmen ausgewählter relevanter 2x2 Spiele geprüft werden. Nachfolgende, auf Basis statistischer Auswertungen abgeleitete Ergebnisse sollen Aufschluss über die Leistung des entwickelten Agenten geben. Zentraler Aspekt dabei ist die Frage nach einer relativen Bewertungsskala, anhand derer die Auszahlungsleistung des Markovspielers bewertet werden kann. Unter Berücksichtigung der Abhängigkeit des eigenen erzielbaren Payoffs von der Strategie des Gegnerspielers wird die Leistung des Algorithmus daher relativ zu menschlichen Referenzspielern bewertet, welche gegen den gleichen zweiten menschlichen Akteur als Sparringspieler wie der Markovagent gespielt hat.

Sekundäres Ziel ist darüber hinaus im Rahmen einer umfassenden Validierung die Prüfung formaler Leistungskriterien sowie die Durchführung eines Strategieturniers gegen algorithmische Gegner. Powers und Shoham (2005b) formulieren als erste ein formales Kriterienset für nichtkooperative präskriptive MAL Algorithmen, welches sowohl Mindest-, als auch Optimalitätsanforderungen an dessen erzielbare Auszahlungen stellt. Die Erfüllung dieser Kriterien durch die hier zu entwickelnde Lösung ist daher von unmittelbarem Interesse der Arbeit. Um die Robustheit der empirischen Ergebnisse gegen Menschen weiter zu untermauern sollen diese außerdem als gängige Praxis der spieltheoretischen MAL Forschung (vgl. z.B. Carmel & Markovitch, 1998; Powers & Shoham, 2005a) in einem Turnier gegen algorithmische Gegner überprüft werden.

In Summe gliedert sich der Beitrag dieser Dissertation somit in zwei Aspekte. Erstens handelt es sich bei dem Markovagenten um die erste dem Verfasser bekannte KI Methode, welche sich

zur effizienten Interaktion mit Menschen eignet und in der Lage ist, an reichhaltigen strategischen Spielverläufen teilzunehmen, ohne auf eine kostspielige Explorationsphase zurückgreifen zu müssen. Zweitens möchte die Arbeit durch den Dreiklang aus empirisch-menschlicher, empirisch-algorithmischer und formallogischer Validierung einen integrierenden Bewertungsrahmen für präskriptive MAL Algorithmen bieten.

1.3 Aufbau der Arbeit

Auf Basis des beschriebenen Forschungsziels gestaltet sich die Gliederung dieser Arbeit wie folgt. Das sich anschließende Kapitel 2 führt dabei zunächst strukturierend an das Themenfeld präskriptiver MAL Algorithmen heran und leitet abschließend die zu untersuchenden Forschungsfragen sowie eine Vorgehensskizze zu deren Beantwortung ab. Darauffolgend erarbeitet Kapitel 3 Grundlagen von Markovstrategien und setzt diese in einem Formalmodell für den interaktiven Markovagenten um. Kapitel 4 beschäftigt sich dann im Rahmen der Methodik mit dem Modus der Laboruntersuchung. Dabei werden Aspekte wie die Auswahl relevanter Spiele, das Design und die Umsetzung der Experimente sowie die Auswertung der Daten thematisiert. Abschließend findet eine Validierung der Laborbedingungen anhand eines Probelaufes statt. Anschließend greift Kapitel 5 die empirischen Hauptergebnisse der Arbeit auf. Im Rahmen von ausgewählten wiederholten 2x2 Spielen findet eine statistische Auswertung der Auszahlungsleistung des KI Algorithmus gegen Menschen statt. Die Auswertung greift neben einer deskriptiven Analyse auf Hypothesentests und Regressionsanalysen zurück. Darauffolgend reflektiert Kapitel 6 die Implikationen der Ergebnisse und setzt diese mit dem Literaturkörper in Bezug. Insbesondere geht der Abschnitt dabei auf die Erfüllung konzeptioneller Leistungskriterien ein und präsentiert die Turnierergebnisse des Markovspielers gegen algorithmische Spieler. Zuletzt fasst Kapitel 7 die gewonnenen Erkenntnisse zusammen, zeigt deren Limitationen auf und bietet Ansatzpunkte für weiterführende Forschung an.

2 Grundlagen

Nachfolgend findet nach einer Eingrenzung und Strukturierung der dem zu untersuchenden Sachverhalt zugrundeliegenden Theorie eine Einführung in Ansätze zur Leistungsbewertung von nichtkooperativen präskriptiven Agenten statt. Das Kapitel orientiert sich dabei an existierenden Literaturüberblicken (T. Sandholm, 2007; Shoham et al., 2007). Im Anschluss an eine kritische Würdigung der Thematik werden dann Forschungsfragen und das weitere Vorgehen abgeleitet.

2.1 Eingrenzung des Betrachtungsgegenstandes

Die Eingrenzung des Betrachtungsgenstandes der Ausarbeitung geht zunächst auf die Spezifika von KI an der Schnittstelle zwischen Spieltheorie und Informatik ein. Dabei liegt ein besonderes Augenmerk auf Mehragentensystemen. Anschließend findet eine Strukturierung der MAL Forschungslandschaft statt.

2.1.1 Spieltheoretischer und informatischer Kontext

Zum Zwecke der Fokussierung des Forschungsvorhabens soll der Betrachtungsgenstand dieser Arbeit eingegrenzt werden. Grundsätzlich nimmt dieses Werk die Perspektive der angewandten Spieltheorie auf Methoden der Forschung zu künstlicher Intelligenz ein, weshalb lediglich KI Verfahren für spieltheoretische Anwendungen diskutiert werden. Der festgelegte Betrachtungsschwerpunkt würdigt somit das gestiegene Forschungsinteresse der Schnittstelle Spieltheorie und KI (vgl. Greenwald & Littman, 2007a; Shoham et al., 2007). Dabei werden spieltheoretische Grundlagen als bekannt vorausgesetzt. Einen zugänglichen Überblick zu entsprechendem Hintergrundwissen liefern zum Beispiel Leyton-Brown und Shoham (2008). Für eine Diskussion der Gemeinsamkeiten zwischen künstlicher Intelligenz und Spieltheorie sei auf Rezek et al. (2008) verwiesen.

Das Feld der künstlichen Intelligenz hat seinen Ursprung in Überlegungen zu *Single Agent Learning*. SAL beschäftigt sich mit Systemen mit einem lernenden Agenten, der sich in einer unbekannten Umgebung zurechtfinden muss (vgl. Mitchell, 1997; Sutton & Barto, 1998). In der Spieltheorie hingegen liegt der Fokus auf der Interaktion zwischen mehreren Agenten untereinander, statt ausschließlich mit deren Umwelt. Auch die KI Forschung nimmt diesen Aspekt

im Rahmen von *Multi Agent Learning* seit Anfang der Jahrtausendwende vermehrt als Forschungsgegenstand wahr (vgl. Shoham & Powers, 2014a).[5] Das Hauptunterscheidungsmerkmal zwischen MAL und SAL ist, dass im MAL Fall mehrere Agenten gleichzeitig lernen und so die interaktiven Prozesse des *Lehrens* und des *Lernens* der Agenten nicht mehr voneinander getrennt werden können. Durch diese Interdependenz ist MAL inhärent komplexer als SAL, da das *Lernen* das *zu lernende* verändert (vgl. Shoham & Powers, 2014a; Young, 2007). Das Lernergebnis des einen Agenten kann eine Veränderung in dessen Verhalten verursachen, was wiederum andere Agenten dazu veranlassen kann, ihr Verhalten ebenfalls anzupassen (vgl. T. Sandholm, 2007; Shoham & Powers, 2014a).

Tabelle 2.1: Normalform des Stackelberg Spiels mit den Auszahlungen des Stufenspiels für den Zeilenspieler gefolgt von denen des Spaltenspielers. Quelle: Eigene Darstellung.

	a_1^2	a_2^2
a_1^1	1/0	3/2
a_2^1	2/1	4/0

Der Sachverhalt wird im Stackelberg Spiel (siehe Tabelle 2.1) deutlich. Der Zeilenspieler besitzt im Stufenspiel die dominante Strategie a_2^1, kann sich jedoch im wiederholten Spiel mit dem Spaltenspieler gemeinsam auf $a_1^1 a_2^2$ als eine für beide Parteien bessere Position heben. Der Zeilenspieler kann nun versuchen dem Spaltenspieler zu lehren, dass eine Kooperationslösung möglich ist. Gleichzeitig lernt er über das Verhalten des Spaltenspielers und möchte sein Verhalten möglicherweise als Reaktion darauf anpassen. Für den Spaltenspieler ergibt sich eine für seine Auszahlungsfunktion analoge Dynamik (vgl. Shoham et al., 2007).

Das vorangegangene Beispiel verdeutlicht, dass der Begriff der *optimalen Politik* aus der KI Forschung zu SAL im Rahmen von Mehragentensystemen in seiner Bedeutung erodiert, da das Verhalten der Agenten durch nicht länger zu trennende und iterativ ineinander verschachtelte Prozesse der Reziprozität und des gegenseitigen Lehrens und Lernens beeinflusst werden (vgl. Erev & Roth, 2007; Shoham, 2008). Infolgedessen wird in dieser Arbeit Lernen als Prozess verstanden, in welchem lernende Agenten sich implizit oder explizit auf Basis des Spielverlaufs aufeinander einstellen, um ihrer Zielfunktion gerecht zu werden.

Gleichgewichte sind traditionell das Ergebnis einer Analyse von als bekannt angenommenen Wissen durch die Spieler zu Spielregeln, den Auszahlungen und Präferenzen, sowie der Rationalität, beziehungsweise der Spielerstrategien. Fudenberg und Levine (1998) sehen hierin kon-

[5] In der Spieltheorie ist MAL auch als *interaktives Lernen* bekannt (vgl. Shoham & Powers, 2014a). Für Schnittstellenliteratur zwischen KI und Spieltheorie sei beispielsweise auf Young (2004) sowie Fudenberg und Levine (1998) aber auch Tennenholtz (2002) verwiesen, die spieltheoretische Überlegungen für die Anwendung im MAL Kontext erarbeiten. Weiterhin gibt Greenwald und Littman (2007b) anhand einer Diskussion von diversen Algorithmen einen strukturierenden Überblick über die MAL Forschung.

zeptionelle und empirische Mängel. Klassische spieltheoretische Methoden zur Erreichung von Gleichgewichtszuständen und Schätzung von beispielsweise gemischten Nash-Strategieprofilen eignen sich für stationäre Umgebungen. Eine Anpassung an nicht-stationäre Interaktionen im Allgemeinen und interdependente Lernprozesse im Speziellen fällt schwer, sodass die praktische Anwendung im dynamischen Kontext bei unbekannter Gegnerstrategie stark eingeschränkt ist (vgl. Rezek et al., 2008). Weiterhin wird die prädiktive oder präskriptive Potenz von Gleichgewichtsanalysen im Rahmen von menschlicher Interaktion durch das komplexe Ausmaß des Strategieraums eines wiederholten Spiels als Abbild der Spielhistorie auf gemischte Strategien im Stufenspiel eingeschränkt. Die Annahme, dass Spieler über ihren gesamten Strategieraum oder gar den des Gegners reflektieren scheint unangemessen (vgl. Shoham et al., 2007). Vielmehr stellen Idealisierungen wie unbeschränkte kognitive Kapazitäten und unendliche gegenseitige Rekursion der Agentenmodelle eine womöglich ineffektive Grundlage für anwendungsorientierte informatisch-spieltheoretische Modelle da, sodass diese Arbeit den Blickwinkel der beschränkten Rationalität einnimmt (vgl. Shoham, 2008). Der Fokus liegt hierbei, aufgrund der Eignung für dynamische Gegebenheiten auf adaptiven Ansätzen, welche eine bessere Grundlage für praktische Anwendungen darstellen, als auf stationäre Bedingungen ausgelegte klassische spieltheoretische Gleichgewichtskonzepte (Rezek et al., 2008). Ein modellbasierter *adaptiver Spieler* passt sein Gegnermodell während des Spiels auf Basis der Historie kontinuierlich an. Wann immer sich ein neues Gegnermodell ergibt, sucht der adaptive Spieler nach einer potentiell neuen *besten Antwort-Strategie* unter Berücksichtigung seiner Nutzenfunktion (vgl. Carmel & Markovitch, 1996).

Für eine wohl strukturierte Bewertung von KI Agenten an der Schnittstelle von Spieltheorie und Informatik haben sich wiederholte 2x2 Spiele in der Literatur als geeigneter Rahmen für eine formalisierte Untersuchung der Interaktion zweier lernender Agenten herauskristallisiert (vgl. Carmel & Markovitch, 1996; Shoham et al., 2007). Diese Arbeit konzentriert sich, den vorangegangenen Überlegungen entsprechend, auf wiederholte Spiele, deren Verlauf vollständig beobachtet werden kann.

2.1.2 Struktur der Multi Agent Learning (MAL) Forschungslandschaft

Die Forschungslandschaft zu MAL im Kontext der Spieltheorie gestaltet sich heterogenen und stellenweise unübersichtlich (vgl. Shoham & Powers, 2014a), sodass die Zeitschrift *Artificial Intelligence* sich mit einer dezidierten Sonderausgabe der Strukturierung des Themenfeldes widmet (R. Vohra & Wellman, 2007). R. V. Vohra und Wellman (2007) stellen die Literatur an der Schnittstelle zwischen Spieltheorie und Informatik bezüglich unterschiedlicher Ausgangspunkten und Zielsetzungen zur Diskussion und attestieren das Fehlen einer einheitlich definierten Problemstellung. Insbesondere erschwert die stellenweise vage bis implizite Natur der Forschungsziele einen stringenten Vergleich der Ansätze (Shoham et al., 2007). Infolgedessen

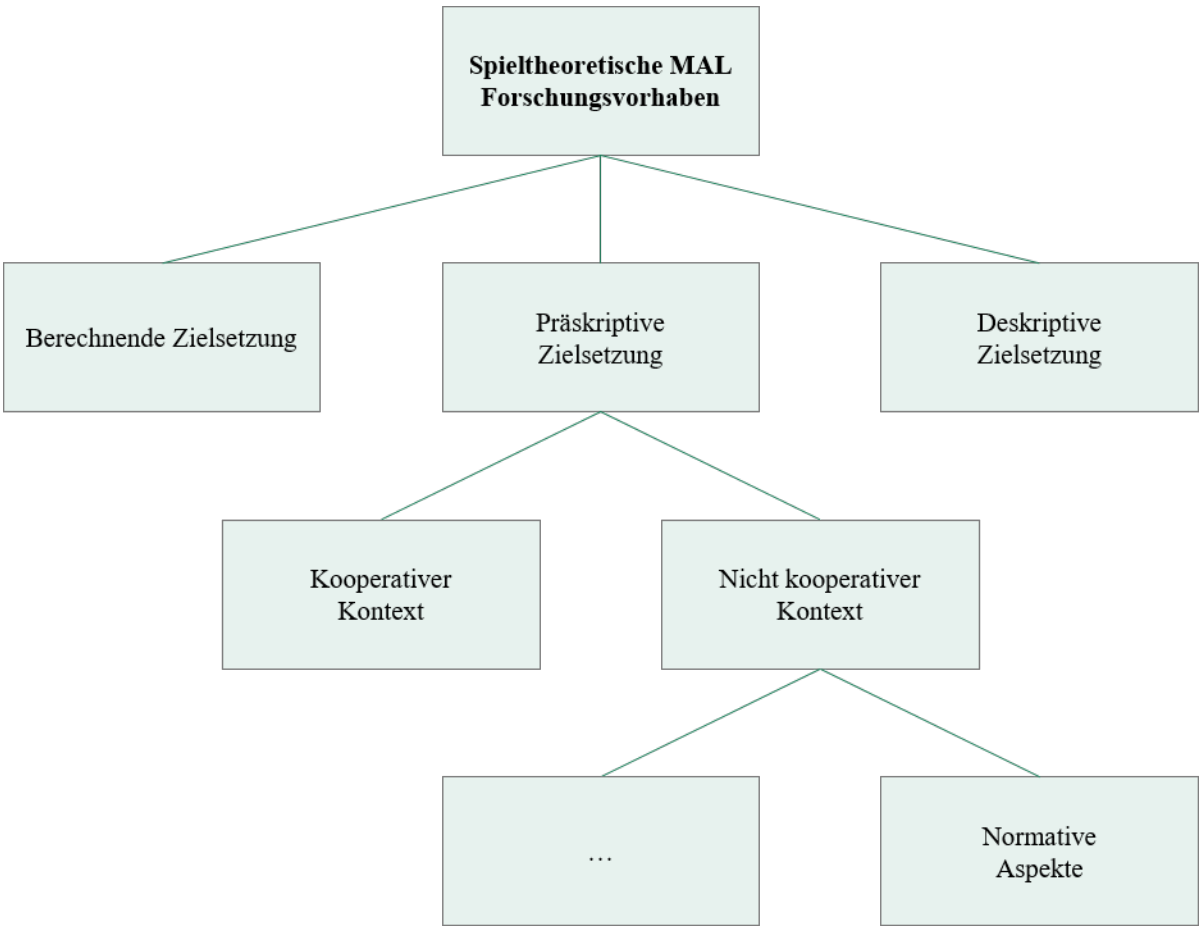


Abbildung 2.1: Struktur der Zielvorhaben von MAL Forschung. Quelle: Eigene Darstellung in Anlehnung an T. Sandholm (2007, S. 383).

soll nachfolgend die Literaturlandschaft entsprechend der Taxonomie von Shoham et al. (2007) sowie deren Erweiterung durch T. Sandholm (2007) gegliedert und im Anschluss auf für diese Arbeit relevante Aspekte eingrenzt werden. Insgesamt werden vier sich teilweise überlappende Forschungsstränge zu spieltheoretischem MAL unterschieden (siehe Abbildung 2.1), welche nachfolgend vorgestellt werden (T. Sandholm, 2007; Shoham & Powers, 2014a; Shoham et al., 2007):

1. **Berechnend:** Ziel der berechnenden Agenda ist, Spieleigenschaften wie Lösungskonzepte durch Lernalgorithmen iterativ zu bestimmen.

2. **Deskriptiv:** Ziel der deskriptiven Agenda ist, Lernprozessen natürlicher Agenten wie Menschen, Tiere und Organisationen durch Formalmodelle möglichst gut nachzubilden.

3. **Kooperativ präskriptiv:** Ziel der kooperativ präskriptiven Agenda ist, lernende Agenten für ein möglichst effektives kooperieren zur Erreichung eines gemeinsamen Ziels zu entwickeln.

4. **Nichtkooperativ präskriptiv:** Ziel der nichtkooperativ präskriptiven Agenda ist, möglichst effektive lernende Agenten im Kontext von Anreizdifferenzen zwischen den Akteuren zu entwickeln.[6]

Nachfolgend sollen die beschriebenen Strömungen kurz unter Einbezug beispielhafter Lösungskonzepte charakterisiert werden.

Berechnende Forschungsvorhaben nutzen Lernalgorithmen als Ansatz zur Bestimmung von Lösungskonzepten für Spiele. *Fictitious Play* wurde ursprünglich von Brown (1951) zur Ermittlung von Nash-Gleichgewichten in Nullsummenspielen entwickelt, während andere Ansätze beispielsweise auf die Berechnung von Gleichgewichten in Spielen mit potentiell asymmetrischem lokalem Effekt eines Spielers auf den Payoff des anderen Spielers abzielen (vgl. Leyton-Brown & Tennenholtz, 2003). Generell sind MAL Algorithmen in der Gleichgewichtssuche nicht effizient, zeichnen sich jedoch durch eine leicht verständliche Logik und einfach Implementierung aus (vgl. T. Sandholm, 2007; Shoham & Powers, 2014a; Shoham et al., 2007). Gleichwohl können MAL Algorithmen wie *No-Regret Learning* für spezifische Problemstellungen in Nullsummenspielen eine valide Alternative zu linearer Programmierung darstellen (vgl. Auer et al., 2002, S. 73) und eignen sich weiterhin zur Gleichgewichtsberechnung, wenn die Agenten die Struktur des Spiels im Sinne der Auszahlungen nicht kennen (vgl. T. Sandholm, 2007).

Deskriptive Forschungsvorhaben haben das grundlegende Ziel, Lernprozesse natürlicher Agenten anhand von Formalmodellen zu beschreiben, deren Ergebnisse sich mit tatsächlich beobachtetem Verhalten decken (vgl. Shoham & Powers, 2014a). Beispiele hierfür sind Kalai und Lehrer (1993) mit einem bayesianischen Ansatz, Erev und Roth (1998, 2007) mit ihrer Reinforcement Learning Lösung, Camerer et al. (2002) mit einem Hybrid aus Reinforcement Learning und Fictitious Play oder Müller (2018) mit einer markovbasierten Modellierung.[7] Versuche realistische Lernprozesse zu modellieren, die im Spiel zu Gleichgewichten führen, wenn Agenten spezifischen Lernregeln folgen und so die Validität der spieltheoretischen Gleichgewichtskonzepte untermauern sollen,[8] werden aufgrund Bedenken bezüglich der zugrundeliegenden Annahmen des Vorgehens nicht weiter berücksichtigt (vgl. T. Sandholm, 2007; Shoham et al., 2007).

Präskriptive Forschungsvorhaben im kooperativen Kontext finden in der KI Literatur insbesondere mit dem Ziel, Kontrollsysteme zu dezentralisieren große Beachtung. Hier wird angestrebt, dass eine Gruppe von Agenten kooperativ in einer potentiell unbekannten Aktionsumgebung lernen. Erfolgsgröße ist dabei häufig die Maximierung des gemeinsamen Payoffs (vgl. z.B.

[6] Weiterhin existiert die *normative Agenda* als Subkategorie der nichtkooperativ präskriptiven Strömung mit dem Ziel, Lernalgorithmen bezüglich ihrer wechselseitigen Gleichgewichtseigenschaften zu untersuchen. Diese wird in Kapitel 2.2 aufgegriffen.

[7] Für eine ausführliche Diskussion von deskriptiv motivierten Lernalgorithmen sei auf Müller (2018) verwiesen.

[8] Siehe auch Arrow (1986).

Chang et al., 2004; Guestrin et al., 2002). Trotz der geradlinig erscheinenden Aufgabenstellung, ein gemeinsames Optimum zu koordinieren, gestaltet sich das Lernproblem als nichttrivial, wenn das Spiel selbst unbekannt ist oder wenn verbündete Agenten nicht entsprechend der eigenen Lerntechnik vorgehen. In einem solchen Umfeld ist die Lernaufgabe in Hinblick auf das Lernen des Spiels selbst, das Lernen über die Mitspieler und das Lernen über das eigene Verhalten ein mehrdimensional geartetes Problem (T. Sandholm, 2007). Der Transfer von SAL Reinforcement Learning Ansätzen auf kooperative MAL Anwendungen (vgl. z.B. Kapetanakis & Kudenko, 2005; Wang & Sandholm, 2002) erweist sich als vielversprechend (vgl. Shoham et al., 2007, S. 366). Littman (2001) entwickelt für MAL im Rahmen stochastischer Spiele den *Friend-or-Foe-Q Learning* Algorithmus, der in Spielen mit einem globalen Optimum oder einem Sattelpunkt lernt, das Nash-Gleichgewicht zu spielen. Bei gleichen Gegebenheiten konvergiert auch der *Nash-Q* Algorithmus, sofern strengere Bedingungen bezüglich der Information über das Vorhandensein eines Sattelpunktes oder globalen Optimums erfüllt in sind (vgl. Hu & Wellman, 2003). In stochastischen Spielen, deren Stufenspiele auf eindeutige Gleichgewichte beschränkt sind, erfolgt die Konvergenz ohne die Informationsbedingung (vgl. Littman, 2001).[9] Weiterhin wurde der *Joint Action Learner* Algorithmus für Spiele mit gleichgerichteten Interessen der Agenten derart gestaltet, dass die Agenten zu einem möglicherweise suboptimalen Nash-Gleichgewicht konvergieren (Claus & Boutilier, 1998). Wang und Sandholm (2003) wiederum schlagen einen entsprechend des optimalen Nash-Gleichgewichtes spielenden lernenden Agenten für Spiele mit gleichgerichteten Interessen vor.

Präskriptive Forschungsvorhaben im nichtkooperativen Kontext stellen die Frage, wie ein lernender Agent in konfliktären Interaktionen lernen soll. Ziel ist es, eine möglichst gute Leistung im stochastischen beziehungsweise wiederholten Spielen zu erzielen. Die daraus abgeleitete Aufgabenstellung ist, einen optimalen oder zumindest effektiven, lernenden Agenten zu entwerfen. Die charakterisierende Herausforderung des MAL besteht darin, dass die Umwelt zum Teil durch andere, diese ebenfalls bewohnende Agenten bestimmt wird, die möglicherweise ebenfalls eigene Lernprozesse durchlaufen (vgl. Shoham et al., 2007). Dabei gestaltet sich das nichtkooperative Umfeld aufgrund der divergenten Anreizstrukturen der einzelnen Agenten als wesentlich komplexer als der kooperative Fall, wobei Nichtnullsummenspiele wie das wiederholte Prisoner's Dilemma (vgl. T. W. Sandholm & Crites, 1996) eine größere Herausforderung als rein konfliktäre Konstantsummenspiele (vgl. Littman, 1994; Ortega & Legg, 2018) darstellen (vgl. T. Sandholm, 2007). Dies spiegelt sich insbesondere in den MAL Lösungsansätzen des Reinforcement Learning wider. Während *Minimax-Q Learning* in Nullsummenspielen gut abschneidet (vgl. Littman, 1994, 1996), erzielen der Nash-Q Learning Algorithmus von Hu und Wellman (1998) und der Friend-or-Foe-Q Learning Algorithmus von Littman (2001)

[9] Es sei angemerkt, dass ein Spiele mit Sattelpunkt in die nichtkooperative Domäne fallen, sodass die beiden Ansätze Nash-Q-Learning und Friend-or-Foe Q-Learning sich nicht ausschließlich auf kooperative Forschungsvorhaben beschränkten.

in Nichtnullsummenspiele weniger zufriedenstellende Ergebnisse (vgl. Shoham et al., 2007). Gleiches gilt für den *CE-Q Learning* Ansatz als Generalisierung des Nash-Q Learning und Friend-or-Foe-Q Learning Algorithmus auf Basis korrelierter Gleichgewichte von Greenwald und Hall (2003) (vgl. Shoham et al., 2007). Vielversprechendere Ansätze artikulieren konkrete Erfolgskriterien für die Leistungsbewertung der MAL Algorithmen in Bezug auf die zuvor beschriebene Herausforderung der Payoffmaximierung vor dem Hintergrund der Ergebnisabhängigkeit von der gegnerischen Strategie. Die Kriterienlandschaft sowie die daraus resultierende Lösungslandschaft gestaltet sich dabei heterogen. Der *(IT-)US-L** Algorithmus etwa beruft sich ausschließlich auf experimentelle Leistungsmessung gegen nicht-lernende auf die Historie bedingende Algorithmen (Carmel & Markovitch, 1996, 1998). Bowling und Veloso (2001) konzentrieren sich im Rahmen ihres *Rational Learning* Algorithmus auf die Erreichung von Gleichgewichten. Bedauernsminimierende Lösungen hingegen bewerten die Leistung eines MAL Algorithmus im Sinne einer unteren Schranke formal als *hoch genug*, wenn die verwendete Strategie retrospektiv kein Bedauern erzeugt (vgl. Fudenberg & Levine, 1995, 1998). Powers und Shoham (2005b) wiederum streben mit ihrer *Meta Strategy* Lösung außerdem eine formal optimale Leistung gegen eine vordefinierte Klasse an Gegnern an.

Zusammenfassend gestaltet sich die Forschung zu MAL im Kontext der Spieltheorie als ein Golem diverser Vorhaben mit verschiedensten Motivationen. Von besonderem Interesse ist aufgrund der nichttrivialen Problemstellung und der Debatte zur adäquaten Messung der Algorithmusleistung insbesondere der Bereich präskriptiver nichtkooperativer Anliegen. Im folgenden Kapitel wird daher der aktuelle Forschungsstand zur Leistungsmessung von präskriptiven MAL Algorithmen in nichtkooperativen Anwendung erarbeitet und korrespondierende Lösungskonzepte vorgestellt.

2.2 Erfolgsmessung präskriptiver MAL Algorithmen im Kontext nichtkooperativer Interaktion

Aufgrund der logischen Relevanz der Leistungsbewertung von nichtkooperativ präskriptiven Lösungen stellt das folgende Kapitel relevante Bewertungsansätze vor. Analog zu Abbildung 2.2 gliedern sich diese in *Bedauernsminimierung*, normative *Gleichgewichtsüberlegungen* sowie in eine *integrierte Betrachtung*.

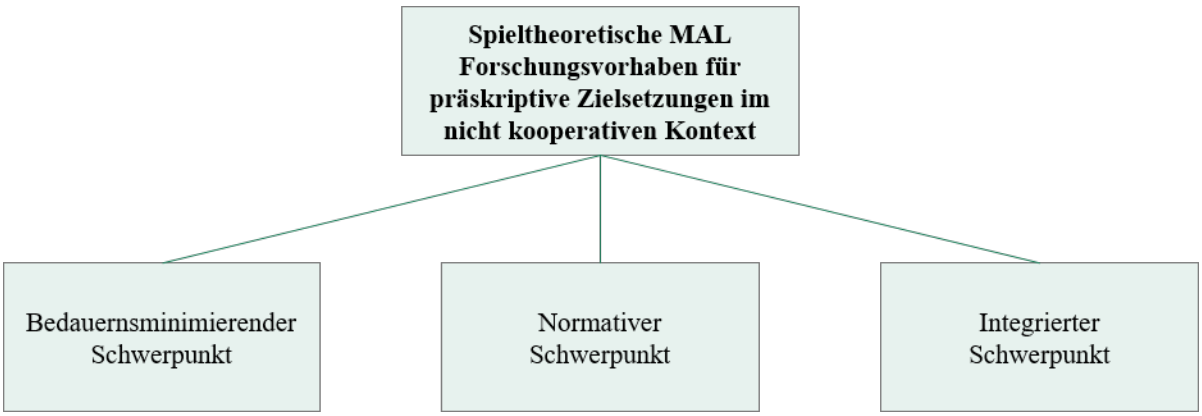


Abbildung 2.2: Erfolgsmessung im Kontext nichtkooperativer präskriptiver MAL Forschung als Anknüpfungspunkte zu Abbildung 2.1. Quelle: Eigene Darstellung.

2.2.1 Lernerfolg im Sinne von Bedauernsminimierung

Die *Erreichung von Auszahlungen über einem gewissen Schwellenwert* ist ein wiederkehrendes Thema spieltheoretischer Forschung, welches aus verschiedenen Blickwinkeln behandelt wurde (vgl. Foster & Vohra, 1999). Die Ausarbeitungen von Fudenberg und Levine (1995, 1998) spiegelt die Literatur zu diesem Ansatz wieder, indem sie folgende Anforderung an Lernregeln stellen:

> **Universelle Konsistenz:** Der Lernalgorithmus soll zumindest die Auszahlung der besten Antwort auf die empirische Verteilung der Aktionen des Gegners in der Spielhistorie, unabhängig von der tatsächlichen Gegnerstrategie, erreichen.

Die Anforderung universeller Konsistenz verlangt, dass der Lernalgorithmus fast sicher mindestens die Auszahlung erwirtschaftet, die er bekommen hätte, wenn er die Häufigkeit aber nicht die Reihenfolge der gegnerischen Aktionen im Vorhinein gewusst hätte. Da die beste Antwort auf die tatsächliche Häufigkeitsverteilung des Gegners mindestens dem Minimax-Payoff entspricht, impliziert das Kriterium der universellen Konsistenz die Gewährleistung der weiteren Kriterien *Sicherheit* und *Konsistenz* (Fudenberg & Levine, 1995, 1998):

> **Sicherheit:** Die durchschnittliche Auszahlung des Lernalgorithmus soll unabhängig vom Gegnerverhalten fast sicher mindestens dem Minimax-Payoff entsprechen.[10]

> **Konsistenz:** Der Lernalgorithmus soll zumindest so gut wie die beste Antwort auf einen Gegner abschneiden, dessen Spielverhalten von unabhängigen Zügen aus einer fixen Verteilung gegeben ist, die der empirisch durchschnittlichen Aktionshäufigkeiten entspricht.

[10] Der Minimax-Payoff ist für Nullsummenspiele identisch mit dem Maximin-Payoff. Der Maximin-Payoff ist der maximale Payoff, der gegen jeden Spieler unabhängig von dessen Verhalten garantiert werden kann.

Im Grenzfall $t \to \infty$ wird auch die *No-Regret* Bedingung durch das Kriterium der universellen Konsistenz sichergestellt (vgl. Shoham et al., 2007). *Regret* oder *Bedauern* ist die Differenz zwischen der rückblickend maximal möglichen Auszahlung einer *einfachen gemischten* Strategie[11] gegeben der Spielhistorie und der tatsächlich durch den Spieler erzielten Auszahlung. Ein Agent erfüllt die No-Regret Bedingung genau dann, wenn das durchschnittliche Bedauern kleiner oder gleich Null gegen alle Gegnerstrategien ist. Dementsprechend handelt es sich bei der Bedingung um eine strenge Anforderung, bei welcher der erwartete Payoff mindestens so groß ist wie der ex-post mögliche erwartete durchschnittliche Payoff einer beliebigen statischen Strategie. Anders gesagt, wird eine Leistung erwartet, die mindestens so gut ist, wie die jeder beliebigen statischen Strategie (vgl. Bowling, 2005).

Eine große Zahl von Algorithmen im Bereich KI und Spieltheorie erfüllen Anforderungen der universellen Konsistenz, beziehungsweise des bedauernsfreien Lernens (vgl. Shoham et al., 2007). Im Anschluss an die Herleitung des Kriteriums der universellen Konsistenz zeigen beispielsweise Fudenberg und Levine (1995), dass *Exponential Fictitious Play*, eine Modifikation des Fictitious Play Ansatzes von Brown (1951), dieses erfüllt. Die Literatur für Lernalgorithmen im Kontext wiederholter Spiele mit multiplen Agenten liefert zahlreiche Beispiele für bedauernsfreie Lösungskonzepte (vgl. z.B. Auer et al., 2002; Bowling, 2005; Freund & Schapire, 1999; Greenwald & Jafari, 2003; Greenwald et al., 2001; Hart & Mas-Colell, 2000; Littlestone & Warmuth, 1994; Zinkevich, 2003), wobei sich das Kriterium im Rahmen stochastischer Spiele nicht erreichen lässt (vgl. Mannor & Shimkin, 2003).

Die Stärke von Lernalgorithmen, deren Bedauern gegen Null strebt, liegt in dem Versuch der Berücksichtigung der Kosten des Lernprozesses (vgl. Powers & Shoham, 2005b). Dies ist insbesondere bei Spielen mit echten Auszahlungen im Sinne von zum Beispiel Geld relevant. Dementsprechend wird der Lernprozess selbst als strategische Aufgabe verstanden, dessen Kosten es zu minimieren gilt, sodass derartige Algorithmen in Nullsummenspielen fast optimal gegen Minimax-Gegner oder jeden stationären Gegner spielen (vgl. T. Sandholm, 2007).

Gleichwohl weisen No-Regret Algorithmen diverse Schwächen auf. Erstens ist das Bedauern trotz einer kontinuierlichen Minimierungsbestrebung und der No-Regret Garantie für $t \to \infty$ nicht sofort Null, sondern nähert sich diesem erst im Durchschnitt über Zeit für den Grenzfall an (vgl. Powers & Shoham, 2005a; T. Sandholm, 2007). Dementsprechend fordern Banerjee und Peng (2003), dass die aus dem Bestreben der Bedauernsminimierung resultierenden Garantien nicht nur im Grenzfall gelten und entwickeln einen Lernalgorithmus, der bedauernsfreie

[11] Als *einfache gemischte Strategie* wird im Rahmen dieser Arbeit eine Strategie definiert, welche die Aktionswahl in jeder Runde des Spiels unabhängig der Historie auf Basis einer fixen Verteilung trifft. Powers und Shoham (2005b) verwenden bezeichnen derartige Strategien als *stationär*, wobei sich der Stationaritätsbegriff stattdessen üblicherweise weiter gefasst ist und sich auf Strategien bezieht, deren Entscheidungsregel für die Aktionswahl in jeder Runde des Spiels gleich ist. Die weiteren Ausführungen dieser Arbeit schließen sich letzterer Definition an.

Auszahlungen mit kleinen zeitlichen polynomialen Schranken garantiert. Zweitens kann Null durchschnittliches Bedauern nicht garantieren, dass die Agenten gelernt haben, ein Gleichgewicht zu spielen. Ein Lernergebnis kann insbesondere nicht garantiert werden, da No-Regret Lerner sich nicht über die Payoffs der anderen Spieler bewusst sind (vgl. T. Sandholm, 2007). Drittens kann die Lösung selbst bei einem Bedauern von Null weit entfernt von paretoeffizient sein (vgl. Powers & Shoham, 2005a; T. Sandholm, 2007). Bereits Fudenberg und Levine (1995) weisen darauf hin, dass No-Regret Ansätze selbst einfachste Muster im Spiel des Gegners nicht nutzen können. Entsprechende Kritik versuchen Fudenberg und Levine (1998) durch eine No-Regret Lösung, die sich an einfache Muster im gegnerischen Verhalten anpasst, zu adressieren. Dennoch kann auch dieser Ansatz eine mögliche Abhängigkeit des gegnerischen Zugverhaltens von dem Verhalten des eigenen Agenten nicht berücksichtigen. No-Regret Algorithmen garantieren zwar eine Mindestauszahlung gegen jeden möglichen Gegner, vernachlässigen dabei jedoch die Möglichkeit, dass das Zugverhalten des Gegners vom eigenen Zugverhalten abhängt (vgl. Powers & Shoham, 2005a).

Tabelle 2.2: Normalform des Prisoner's Dilemma Spiels mit den Auszahlungen des Stufenspiels für den Zeilenspieler gefolgt von denen des Spaltenspielers. Quelle: Eigene Darstellung.

	a_1^2	a_2^2
a_1^1	3/3	0/5
a_2^1	5/0	1/1

Die Unfähigkeit, bedingende Gegner adäquat zu berücksichtigen kann die auszahlungsbezogene Leistung von MAL Algorithmen bisweilen stark beeinflussen, wobei der Effekt insbesondere in Spielen mit wenigen Spielern schwerer wiegt, da hier der Einfluss des Verhaltens des einzelnen Spielers am größten ist. Der Sachverhalt wird am Beispiel des wiederholten Prisoner's Dilemmas (siehe Tabelle 2.2) deutlich, für welches im Rahmen umfangreicher Strategieturniere durch Axelrod (1984), Tit-for-Tat als eine effektiver Strategiealgorithmus hervorging, welcher Kooperation zulässt, sich dabei jedoch nicht ausnutzen lässt. Tit-for-Tat kooperiert initial mit a_1^1 und imitiert nachfolgend die Aktion des Gegners in der letzten Runde. Jedweder Ansatz, der wie beispielsweise Fictitious Play (Brown, 1951) ausschließlich von stationären Gegnern ausgeht, wird im wiederholten Prisoner's Dilemma stets *Abweichen*, also a_2^1, spielen, da es sich hierbei um die einzige beste Antwort zu jedem stationären Gegner handelt und stets ein Ergebnis ohne Bedauern erzeugt (vgl. Shoham et al., 2007). Gegen Tit-for-Tat würde dies unter Vernachlässigung der ersten Runde in einer durchschnittlichen Auszahlung von 1 resultieren. Eine Strategie, die stets *Kooperieren*, also a_1^i, spielt, würde jedoch einen durchschnittlichen Payoff von 3 erwirken (vgl. Powers & Shoham, 2005b). Das Beispiel zeigt, dass No-Regret Ansätze in einem reichhaltigeren Strategieraum nicht zielführend sind (vgl. Powers & Shoham, 2005b). Es kann

sich dabei folglich nur um eine Mindestanforderung an die Leistung von MAL Algorithmen handeln, die jedoch keine Aussage über die Nutzung von Chancen aus interaktivem Spielverhalten trifft.

2.2.2 Lernerfolg im Sinne von Stabilität

Als Alternative zu bedauernsminimierenden Kriterien schlagen Bowling und Veloso (2002) die *Erreichung von Gleichgewichtszuständen* als in Folge von *Rationalität*[12] und *Konvergenz* zur Bewertung von MAL Algorithmen mit Fokus auf stationäre Gegner vor. Kriterien lassen sich wie folgt charakterisieren:

> **Rationalität:** Wenn die Gegner gegen eine stationäre Strategie konvergieren, soll der Lernalgorithmus gegen eine stationäre beste Antwort-Strategie konvergieren.

Dadurch wird sichergestellt, dass der Lernalgorithmus eine beste Antwort lernt, die gegen stationäre Gegner in jedem Fall existiert (vgl. Bowling & Veloso, 2002). Insbesondere wird so gewährleistet, dass der MAL Algorithmus lernt, einen irrationalen stationären Gegner maximal auszunutzen (vgl. T. Sandholm, 2007).

> **Konvergenz:** Der Lernalgorithmus soll gegen bestimmte Lernalgorithmen[13] gegen eine stationäre Strategie konvergieren.

Wenn alle Spieler den Kriterien gemäß *rational* agieren und ihr Spiel *konvergiert*, dann interagieren sie im Nash-Gleichgewicht des Stufenspiels. Jeder Spieler konvergiert gegen eine stationäre Strategie und spielt eine beste Antwort, wenn die Gegner gegen eine stationäre Strategie konvergieren (vgl. Bowling & Veloso, 2002; Powers & Shoham, 2005b). Folglich implizieren die beiden vorangegangenen Anforderungen das Kriterium der Nash-Konvergenz im Selbstspiel:

> **Nash-Konvergenz:** Der Lernalgorithmus soll im wiederholten Selbstspiel gegen das Nash-Gleichgewicht des Stufenspiels konvergieren.

[12] Der Rationalitätsbegriff von Bowling und Veloso (2002) weicht von der klassischen spieltheoretischen Definition ab.

[13] Die Konvergenzeigenschaft kann unter Einhaltung des Rationalitätskriteriums nicht gegen jeden beliebigen Agenten erfüllt werden, sondern muss auf eine Klasse von Lernalgorithmen bedingen (vgl. Powers & Shoham, 2005a). Die MAL Literatur zu No-Regret Kriterien fokussiert sich dabei auf das *Selbstspiel*, bei dem die Konvergenz des Algorithmus im Spiel gegen einen identischen Lernalgorithmus untersucht wird (vgl. Bowling & Veloso, 2002; Conitzer & Sandholm, 2007).

Damit adressieren Bowling und Veloso (2002) einen zentralen Kritikpunkt an der klassischen spieltheoretischen Gleichgewichtstheorie, dass das Spiel einer Gleichgewichtsstrategie gegen einen Spieler, der selbst von einer Gleichgewichtsstrategie abweicht, nicht unbedingt optimal ist (vgl. T. Sandholm, 2007). In diesem Sinne wird die Nash-Konvergenz nur gegen einen stationären Gleichgewichtsspieler gefordert wird, für den abweichenden Fall jedoch nicht zwingend gegeben ist.

Der Reinforcement Learning Algorithmus von Bowling und Veloso (2002) *WoLF-IGA* als Weiterentwicklung des auf dem Gradientenverfahren basierenden *IGA Algorithmus* (Singh et al., 2000) erfüllt das Kriterium der Nash-Konvergenz im Selbstspiel nachweislich für wiederholte 2x2 Spiele. Dabei verwendet WoLF-IGA eine geringere Lernrate, wenn es gewinnt und eine höhere Lernrate, wenn es verliert entsprechend dem Namen des Algorithmus *Win or Learn Fast* (WoLF). Hierdurch ist gewährleistet, dass der Algorithmus sein Verhalten bei vorteilhaftem Verlauf nicht überschnell ändert, jedoch bei unvorteilhaftem Verlauf schnell in der Lage ist, sich an Gegebenheiten anzupassen. Später schlagen Conitzer und Sandholm (2007) mit *AWESOME* einen auf stationäre Gegner fokussierten MAL Algorithmus vor, der die Kriterien von Bowling und Veloso (2002) für alle wiederholten Spiele erfüllt. Die Vorgehensweise von AWESOME[14] ist, sich im Sinne einer besten Antwort an die Strategie des Gegners anzupassen, wenn dieser stationär zu sein scheint. Andernfalls greift der Algorithmus auf eine vordefinierte Gleichgewichtsstrategie zurück. Dabei ist sich AWESOME insofern seiner selbst bewusst, als es versucht dem Gegner möglichst stationäre Signale durch das eigene Verhalten zu senden. Weitere KI Lösungen, die im Selbstspiel gegen ein Gleichgewicht des Stufenspiels konvergieren ergeben sich im Bereich des Reinforcement Learnings. Beispielsweise konvergiert der *Minimax-Q Learner* (vgl. Littman, 1996) im Grenzfall für jedes Nullsummenspiel, sodass stets ein Nash-Gleichgewicht im Selbstspiel erreicht wird. Weiterhin konvergieren *Friend-or-Foe Q-Learning* (vgl. Littman, 2001) und *Nash Q-Learning* (vgl. Hu & Wellman, 2003) zum Nash-Gleichgewicht für bestimmte Spiele. Auch im Bereich modellbasierter Lösungen können Konvergenzeigenschaften im Selbstspiel erfüllt werden. Ein Beispiel ist Fictitious Play in Nullsummenspielen (vgl. J. Robinson, 1951), 2x2 Generalsummenspielen (vgl. Miyasawa, 1961) und in Spielen, die durch Elimination von strikt dominanten Strategien gelöst werden können (vgl. Nachbar, 1990). Darüber hinaus konvergiert der Rational Learning Ansatz von Kalai und Lehrer (1993) gegen das Nash-Gleichgewicht des wiederholten Spiels, wenn absolute Kontinuität angenommen wird (vgl. Shoham et al., 2007).

Die Verwendung von gleichgewichtsbezogenen Leistungsanforderungen wird in der Literatur kritisch hinterfragt. Erstens fokussieren sich die Gleichgewichtsbetrachtungen der MAL Lernalgorithmen stark auf das Selbstspiel, während die Gleichgewichtseigenschaften zwischen verschiedenen Lösungen wenig Beachtung findet (vgl. Shoham et al., 2007). Zweitens ist der Geg-

[14] AWESOME ist ein Akronym für *Adapt When Everybody is Stationary, Otherwise Move to Equilibrium*.

Tabelle 2.3: Normalform des Chicken Games mit den Auszahlungen des Stufenspiels für den Zeilenspieler gefolgt von denen des Spaltenspielers. Quelle: Eigene Darstellung.

	a_1^2	a_2^2
a_1^1	3/3	1/5
a_2^1	5/1	0/0

ner entgegen der Annahme nicht notwendigerweise stationär (vgl. T. Sandholm, 2007), sodass die Kriterien zu suboptimalen Ergebnissen führen können. Dies wird beispielsweise erneut im wiederholten Prisoner's Dilemma deutlich (siehe Tabelle 2.2), in dem im Selbstspiel durch spielen des Nash-Gleichgewichts des Stufenspiels eine Auszahlung von 1 liefert, während der finite Automat Tit-for-Tat ein Ergebnis von 3 erzielt. Ähnliches gilt im wiederholten Chicken Game (siehe Tabelle 2.3). Hier erzielt ein Algorithmus, der sich im Selbstspiel mit dem Gegner bezüglich der Aktionen alternierend abwechselt ebenfalls einen deutlich besseren Payoff, als eine einfache gemischte Strategie ermöglichen würde. Dementsprechend ist die Annahme gegnerischer Stationarität im besten Fall als eine Sonderlösung zu verstehen (vgl. Powers & Shoham, 2005b). Drittens und direkt daraus resultierend macht der Kriteriensatz keine Aussage über die von den Algorithmen erzielten Auszahlungen, insbesondere gegen nicht-stationäre Gegner (vgl. Powers & Shoham, 2005a), sodass sich die entwickelten Lösungen sich potentiell von nicht-stationären Gegnern ausnutzen lassen (vgl. Chang & Kaelbling, 2002). Shoham et al. (2007) hinterfragen entsprechend, ob die Art und Weise des Spiels im Sinne Erreichung von Gleichgewichten den Stufenspiels oder die Erreichung von bestimmten Auszahlungswerten unabhängig einer Konvergenz zielführend ist. Viertens gelten die aus den Kriterien resultierenden Eigenschaften lediglich im Grenzfall, jedoch nicht unbedingt im endlich wiederholten Spiel (vgl. Powers & Shoham, 2005b). Dementsprechend kommen Rationalität und Konvergenz im Sinne der Kriterien lediglich für das Lernergebnis, nicht für den Gesamtlernprozess in Frage, dessen Kosten bei einer derartigen Betrachtung vernachlässigt werden (vgl. T. Sandholm, 2007). Als Reaktion auf das Defizit ergänzen Bowling (2005) die ursprünglich vorgeschlagenen Kriterien (vgl. Bowling & Veloso, 2002) um die Erfordernis, dass das durchschnittliche Bedauern des Agenten Null sein soll. Ergebnis ist der *GIGA-WoLF* Algorithmus, der im Selbstspiel für 2x2 Spiele gegen das Nash-Gleichgewicht konvergiert und darüber hinaus die No-Regret Bedingungen von Fudenberg und Levine (1995) erfüllt. Fünftens ist der Fokus auf das Nash-Gleichgewicht des Stufenspiels nicht notwendigerweise der vielversprechendste Zugang zu Gleichgewichtsanforderungen (vgl. T. Sandholm, 2007). Alternativ stellen sich zum Beispiel korrelierte Gleichgewichte dar, gegen welche der *Correlated Q-Lerner* (Greenwald & Hall, 2003) und dem der *Regret Matching* Ansatz (Hart & Mas-Colell, 2000) konvergieren. Darüber hinaus bieten sich für wiederholte Spiele insbesondere die Gleichgewichte des wiederholten Spiels an Stelle derer

des Stufenspiels an, da erstere bisweilen höhere Auszahlungsergebnisse erzielen können (vgl. T. W. Sandholm & Crites, 1996).[15]

Zusammenfassend wird die Verwendung von Gleichgewichten des Stufenspiels im Selbstspiel als Formalkriterium für präskriptive nichtkooperative MAL Algorithmen in der relevanten Literatur stark hinterfragt. Gleichwohl werden die beschriebenen Kriterien bisweilen als *Minimalanforderung* für jeden vernünftigen nichtkooperativen präskriptiven MAL Algorithmus verstanden (vgl. Conitzer & Sandholm, 2007; T. Sandholm, 2007). Dennoch ist deren Anwendung nicht immer unproblematisch. Zum einen sind die Kriterien zu eng gefasst, da nicht alle Gegner gegen eine stationäre Strategie konvergieren. Zum anderen sind die Kriterien zu weit gefasst, da eine klare beste Antwort nur unter Einschränkung des gegnerischen Strategieraumes möglich ist (vgl. Shoham et al., 2007).

2.2.3 Lernerfolg im Sinne integrierter Anforderungen

Als Reaktion auf die Defizite der vorangegangen Kriteriensätze schlagen Powers und Shoham (2005b) einen integrierten und weiterentwickelten Katalog zur Leistungsmessung von präskriptiven nichtkooperativen MAL Algorithmen vor. Insbesondere soll so die Missachtung der möglichen Abhängigkeit des Gegnerverhaltens von dem Verhalten des MAL Agenten im No-Regret Milieu ausgeräumt werden, welches seinen Ursprung in Interaktionen mit großen Spielerpopulationen hat. Weiterhin soll die Vernachlässigung von Auszahlungsgarantien gegen nichtstationäre, insbesondere bedingende, Gegner durch Stabilitätskriterien adressiert werden, welche sich beispielsweise im wiederholten Prisoner's Dilemma (siehe Tabelle 2.2) gegen einen Tit-for-Tat-Spieler oder im wiederholten Chicken Game (siehe Tabelle 2.3) gegen einen alternierenden Gegner zeigen.

Es handelt sich bei den Kriterien von Powers und Shoham (2005b) jeweils um klar definierte quantitative Auszahlungsziele. Diese sind jedoch nicht absolut festgeschrieben, sondern können sich je nach Anwendung unter Berücksichtigung der Spielpayoffs und der Gegnerklasse parametrisiert werden und lassen so ein Urteil über den Grad der Zielerreichung zu. Dabei ist gefordert, dass ein MAL Algorithmus jedes der Leistungsziele unter Abzug eines von der Gegnerklasse abhängigen Abschlags $\varepsilon > 0$ nach einer endlichen Explorationsphase von T_0 Runden für alle folgenden Runden $t > T_0$ mit einer Wahrscheinlichkeit von $1 - \delta > 0$ erfüllt. Die Desiderata sind mehrstufig gegliedert.

Das erste Kriterium der *gezielten Optimalität* orientiert sich dabei an der Kritik bezüglich Stabilitätskriterien, dass die Qualität einer Antwort-Strategie durch Annahmen bezüglich der Gegenspieler beschränkt sein müssen. Es gibt somit eine optimistische Leistungsanforderung an das Spiel gegen eine bestimmte frei wählbare Gegnerklasse an.

[15] Hierbei sei angemerkt, dass die Gleichgewichte des wiederholten Spiels vom womöglich unbekannten Zinsfaktor der anderen Agenten abhängen (vgl. T. Sandholm, 2007).

> **Gezielte Optimalität:** Wenn der Gegner ein Mitglied der ausgewählten Menge an Gegnern ist, soll die durchschnittliche Auszahlung des Lernalgorithmus im wiederholten Spiel höchstens ε unter Wert des durchschnittlichen Payoffs der besten Antwort gegen den tatsächlichen Gegner liegen.

Das zweite Kriterium wiederum gibt Mindestanforderungen für das Spiel gegen alle anderen Gegner außerhalb der Zielklasse an.[16] Somit ist eine gewisse Absicherung gegen Spieler außerhalb der angedachten Gegnermenge nach unten hin gegeben und beschränkt somit die Möglichkeit von Orchideenlösungen, die ausgezeichnet gegen eine spezifische Gegnernische spielen, andernfalls jedoch nur schwache Leistung erbringen.

> **Sicherheit:** Gegen jeden Gegner soll die durchschnittliche Auszahlung des Lernalgorithmus im wiederholten Spiel höchstens ε unter dem Maximin-Payoff oder Sicherheitswert des Stufenspiels liegen[17].

Zu guter Letzt definiert das dritte Kriterium der *Kompatibilität* die Anforderungen an MAL Algorithmen im Spiel gegen sich selbst. Die zugrundeliegende Logik ist, dass ein zu bewertender Lernalgorithmus im Spiel mit einer identischen Version seiner selbst das für beide jeweils bestmögliche Individualergebnis zulassen können soll.

> **Kompatibilität:** Der Lernalgorithmus soll im wiederholten Selbstspiel eine durchschnittlichen Auszahlung von höchstens ε unter dem Minimum der Auszahlungen in der Menge aller Nash-Gleichgewichte erzielen, welche nicht von anderen Nash-Gleichgewichten paretodominiert werden.

Die integrierten Kriterien von Powers und Shoham (2005b) erlauben für deutlich präzisere und damit bisweilen strengere Anforderungen an MAL Agenten. Ziel ist eine optimalitätsgetriebene spezifische Expertise bezüglich der Zielgegnerklasse anzustreben, ohne dabei anfällig für Spieler außerhalb dieser Klasse zu sein. Hervorzuheben ist, dass die *gezielte Optimalität* dabei auf der *tatsächlichen* Gegnerstrategie bedingt, wenn dieser zur Zielmenge der Gegner des Lernalgorithmus gehört. Somit wird im Spiel gegen adaptive Gegner im Gegensatz zu den No-Regret Kriterien, die eine fixe und unbedingte Wahrscheinlichkeitsverteilung der Gegneraktionen annehmen, die Interdependenz der Aktionswahl aller Spieler berücksichtigt. Im wiederholten Prisoner's Dilemma (siehe Tabelle 2.2) gegen Tit-for-Tat wird jetzt ein Mindestwert von $3-\varepsilon$ gefordert, während das No-Regret Kriterium sich bereits mit einer Auszahlung von 1 genügt

[16] Ist das Kriterium der gezielten Optimalität für die Zielgegnermenge gegeben, ist für diese auch das Sicherheitskriterium erfüllt.

[17] Der Sicherheitswert des Agenten ist $max_{\pi^1} min_{\pi^2} \mathbf{E}(v^1(\pi^1, \pi^2))$, also der Maximin-Wert des erwarteten Payoffs, wenn beiden Spieler π^i aus der Menge gemischter Strategien wählen.

(siehe Kapitel 2.2.1). Weiterhin stellt die Berücksichtigung von möglichen Reihenfolgeeffekten im gegnerischen Verhalten je nach Gegnerparameter ε potentiell höhere Anforderungen, als durch die auf Stationarität fokussierten Stabilitätskriterien formuliert. Im wiederholten Chicken Game (siehe Tabelle 2.3) gegen einen alternierenden Spieler wird jetzt ebenfalls ein Mindestwert von $3 - \varepsilon$ gefordert, während die Stabilitätskriterien auf einen scheinbar zu 50% randomisierenden Gegner mit der stationären besten Antwort im Sinne von stets a_2^1 reagiert und so lediglich eine durchschnittliche Auszahlung von 2.5 realisiert hätten (siehe Kapitel 2.2.2).

Gleichzeitig wird durch das Kriterium der *Sicherheit* ähnlich zu den No-Regret Kriterien ein durchschnittlicher Payoff, welcher höchstens ε unter dem Maximin-Payoff liegen darf, gewährleistet. Wesentlicher Unterschied zu den No-Regret Kriterien ist hier, dass die Anforderung für Gegner außerhalb der klar spezifizierten Zielgegnermenge relevant ist. Für Zielgegner ist die Erfüllung der Sicherheit durch Erfüllung der gezielten Optimalität bereits garantiert.

Auch die Anforderung an *Kompatibilität* ist der *Nash-Konvergenz* des Selbstspiels aus den Stabilitätskriterien in seiner Spezifität überlegen, da eine beliebige beste Antwort nicht länger ausreichend ist. Stattdessen wird eine beste Antwort verlangt, die gleichzeitig die erhaltene Auszahlung über die Menge der besten Antworten maximiert (vgl. Powers & Shoham, 2005b).

Powers und Shoham (2005b) schlagen mit *MetaStrategy* einen Algorithmus vor, der die selbst gesetzten Kriterien für die Zielmenge einfacher gemischter Gegner erfüllen soll. Dabei handelt es sich um einen Hybrid aus den existierenden Ansätzen Fictitious Play (vgl. Brown, 1951), Bully (vgl. Littman & Stone, 2002) und der Maximin-Strategie. Der Algorithmus geht dabei wie folgt vor:

1. **Explorationsphase:** MetaStrategy initialisiert zu Beginn eine Explorationsphase, in welcher die Spieleigenschaften des Gegners bestimmt werden sollen. Während der Explorationsphase setzt der Algorithmus das Bully Modul und bei nicht zufriedenstellender Leistung auch das Fictitious Play Modul ein.

2. **Restspielverlauf:** Am Ende der Explorationsphase legt MetaStrategy seine präferierte Spielweise für den Restspielverlauf fest. Die Entscheidung gliedert sich in eine Mehrstufige Überlegung:

    - Wenn der Gegner bisher als einfache gemischte Strategie erscheint, legt sich der Algorithmus auf Fictitious Play im Sinne einer besten Antwort auf die Verteilung aller gegnerischen Aktionen fest. Das Fictitious Play Modul stellt dabei sicher, dass gegen vermutlich einfache gemischte Strategie eine beste Antwort abhängig von dessen empirischer Aktionsverteilung gefunden wird. Damit im Selbstspiel ein paretoeffizientes Nash-Gleichgewicht erreicht werden kann, wird Fictitious Play insofern *großzügiger* gestaltet, als dass der Agent im Falle mehrerer gleichwertiger Optionen jene wählt, die den gegnerischen Payoff maximiert.

- Falls der Gegner nicht als einfache gemischte Strategie erscheint und das Bully in der Explorationsphase gute Ergebnisse erzielen konnte, wird das Bully Modul als präferierte Spielweise für den Restspielverlauf definiert. Der verwendete Bully Spieler ist ein Agent, der stets die gemischte Strategie spielt, welche seine Auszahlung unter Annahme einer besten Antwort durch den sich an den Bully anpassenden Gegner maximiert. Auch der Bully Spieler wird *großzügig* gestaltet, sodass er bei eigener Indifferenz den Gegnerpayoff maximiert.
- Falls keine der beiden Bedingungen erfüllt sind, legt sich MetaStrategy auf eine beste Antwort auf eine Partition der letzten Runden fest.
- Solange der durchschnittliche Payoff von MetaStrategy nicht unter den Sicherheitswert des Spiels fällt, führt der Algorithmus seine präferierte Spielweise aus. Andernfalls wechselt er solange zu einer Maximin-Strategie, bis die durchschnittliche Auszahlung wieder über dem Sicherheitswert liegt. Klar ist, dass der Maximin-Spieler unabhängig vom Gegnerverhalten versucht, zumindest das Maximum der Zeilenminima der Auszahlungsmatrix zu realisieren, um sich so nach unten hin abzusichern.

Im empirischen Benchmarkturnier gegen ausgewählten Algorithmen wie Bully, Minimax, Fictitious Play (vgl. Fudenberg & Levine, 1998), IGA (vgl. Singh et al., 2000), GIGA-WoLF (vgl. Bowling & Veloso, 2002) und dem Joint-Action Learner (vgl. Claus & Boutilier, 1998) schneidet MetaStrategy sowohl im Spiel gegen andere Agenten, als auch im Selbstspiel gut ab (vgl. Powers & Shoham, 2005b). Ähnliche positive Turnierergebnisse konnten Vu et al. (2006) unter Erfüllung der gegnerspezifischen Kriterien in Spielen gegen mehrere Gegner erzielen.

2.3 Kritische Würdigung

Wie die vorangegangenen Ausführungen illustrieren, gestaltet sich die Literatur zu MAL Algorithmen heterogen. Archetypisch lassen sich berechnende, deskriptive sowie kooperative und nichtkooperative präskriptive Vorhaben unterscheiden. Als von besonderem Interesse für diese Arbeit wurde aufgrund der divergenten Zielsetzungen der Agenten die nichtkooperativ präskriptive Strömung herausgearbeitet (vgl. T. Sandholm, 2007; Shoham & Powers, 2014a; Shoham et al., 2007). Während jede der Strömungen ihren eigenen Herausforderungen ausgesetzt ist, stellt sich die Erfolgsmessung von nichtkooperativ präskriptiven MAL Algorithmen zu deren Bewertung und Vergleich als zentraler Diskussionspunkt dar. Hierbei sind Schwächen sowohl im Rahmen der *Konzeption*, als auch im Rahmen *formaler* und *experimenteller Validierung* der erarbeiteten MAL Algorithmen vorhanden, welche in potentiell unvorteilhafter Gestaltung der Lösungen resultierten kann.

2.3.1 Konzeption der Interaktionslogik

Menschen agieren häufig im Bereich sophistizierter Rationalität aus(vgl. Camerer et al., 2004; Camerer, 1997; Milgrom & Roberts, 1991), sodass sich ihr Spielverhalten als Funktion der zurückliegenden Aktionen der Spieler beschreiben lässt (vgl. Dal Bo & Frechette, 2013; Müller, 2018, S. 1). Allerdings sind nur wenige Lernmodelle in der Lage, derartige Interaktionen abzubilden (vgl. Erev & Haruvy, 2016, S. 684).

Eine von den Schöpfern von MetaStrategy selbst aufgezeigt Schwäche ist der erneute Fokus auf Gegner mit einfachen gemischten Strategien (vgl. Powers & Shoham, 2005b). Auch MetaStrategy vernachlässigt trotz Erfüllung der selbstgesteckten Leistungskriterien die zentrale Eigenschaft des MAL, dass auch gegnerische Spieler in einem Lernprozess sein können und ihr Verhalten daher potentiell vom Verhalten des Agenten abhängt (vgl. Powers & Shoham, 2005a). *Manipulator*, eine subsequent vorgeschlagene Variation von MetaStrategy, adressiert dies unter Einhaltung der Leistungskriterien, indem sich die Zielgegner der neuen Lösung aus adaptiven Spielern mit beschränktem Gedächtnis zusammensetzen, deren Strategie auf einer begrenzten Partition der kürzlichen Spielhistorie bedingt. Unter anderem ersetzt Manipulator das Spielen der besten Antwort auf die empirische unbedingte Verteilung der Aktionen des Gegners im Fictitious Play Modul durch die Berechnung eines besten Aktionszyklus, der gegen den bedingenden Spieler den höchsten erwarteten Payoff liefert. Im Turniervergleich mit anderen algorithmischen Spielern konnte auch Manipulator zufriedenstellende Ergebnisse erzielen (vgl. Powers & Shoham, 2005a).

Weiterhin präsentieren Carmel und Markovitch (1996) mit ihrem *Unsupervised-L** (US-L*) Algorithmus schon früh einen interaktiven MAL Ansatz, der sich auf adaptive Gegner spezialisiert, welche sich durch Deterministischen Finiten Automaten (DFA) modellieren lassen. Dabei beobachtet US-L* den zu modellierenden Gegner zunächst im Spiel mit anderen Agenten und schätzt auf Basis der Aktionshistorie einen möglichst kompakten DFA für den zu modellierenden Gegner. Offline wird dann für das gegebene Modell eine beste Antwort berechnet, welche der Algorithmus dann im nachgelagerten Spiel gegen den zuvor beobachteten Gegner einsetzt. Limitierender konzeptioneller Faktor für eine Anwendung im realweltliche Kontext ist dabei die Unfähigkeit, gemischte Strategien abbilden zu können. Die Methode wird jedoch in der Literatur kaum aufgegriffen.

Den Ausführungen entsprechend konnten im Rahmen dieser Arbeit in der Literatur nur zwei MAL Algorithmen identifiziert werden, welche prinzipiell zu strategisch reichhaltiger sophistizierter Interaktion fähig sind und sich somit zum Spiel gegen menschliche Agenten eignen.

2.3.2 Formale Validierung

Schwächen hinsichtlich formaler Validierung gliedern sich in das Abhandensein von expliziten formalen Leistungskriterien sowie das Vorhandensein von expliziten aber nur partiellen formalen Leistungskriterien.

Einerseits werden konkrete Erfolgskriterien im Kontext von MAL Forschung bisweilen nicht explizit formuliert, sodass ein Vergleich der Lösungen schwer fällt (vgl. Powers & Shoham, 2005b; Shoham et al., 2007). Die Leistungsbewertung von US-L* erfolgt beispielsweise ausschließlich im Rahmen von Strategieturnieren gegen algorithmische Spieler aus der Menge DFA. Formalen Kriterien kommen dabei nicht zum Einsatz.

Der Verzicht auf Formalkriterien ist jedoch bei weitem nicht omnipräsent. Dennoch sind die verwendeten Zieleigenschaften stellenweise unvollständig in Bezug auf das Forschungsziel, einen Leistungsstarken MAL Algorithmus zu entwickeln. Gleichwohl findet darüber hinaus eine reflektierte Diskussion bezüglich geeigneter formaler Kriteriensets zur Beurteilung von MAL Algorithmen statt. Diese gestalten sich jedoch heterogen und nehmen teilweise weiterhin eine nur partielle Perspektive auf die Gesamtleistung. Als einer der ersten Vorschläge hierzu werden bedauernsminimierende Kriterien, die auf Fudenberg und Levine (1995) zurückgehen und unter anderem für die Vernachlässigung der Abhängigkeit des gegnerischen Spielverhaltens vom Verhalten des Agenten kritisiert. Hierdurch werden realisierbare Lösungen, die eine Auszahlung über dem des bedauernsminimierenden Verhaltens realisieren vollständig ausgeklammert.

Auch nachfolgende stabilitätsbezogene Kriterien wie von Bowling und Veloso (2002) sowie Conitzer und Sandholm (2007) befriedigen unter anderem aufgrund ihres Fokus auf streng stationäre Gegner nicht. Insbesondere wird dabei die Gleichgewichtskonvergenz häufig als unreflektierter Goldstandard herangezogen. Konvergenzeigenschaften sind jedoch kein Selbstzweck (vgl. Shoham & Powers, 2014a), vielmehr rücken die erzielten Auszahlungen bei einer derartigen Betrachtung in der Hintergrund (vgl. Shoham et al., 2007). Beispielsweise kann das Spielverhalten der Agenten von den Stabilitätseigenschaften abweichen oder das resultierende Spielverhalten selbst beim Gleichgewichtsspiel weit von optimalen Auszahlung entfernt sein (vgl. T. Sandholm, 2007).

Die Kriterien und abgeleiteten Algorithmen finden nur in einem jeweils spezifischen Umfeld Anwendung. Im Zuge dessen formulieren Powers und Shoham (2005b) einen MAL Algorithmus sowie ein Set an Kriterien im Sinne von Mindestanforderungen an jeden nichtkooperativen präskriptiven Lernalgorithmus, welche sich sowohl durch Flexibilität in Bezug auf den Anwendungskontext des zu bewertenden Algorithmus, als auch durch hinreichende Präzision in Bezug auf die formulierten Anforderungen auszeichnen. Insofern existiert mit den Kriterien von Powers und Shoham (2005b) eine belastbare Grundlage für die formale Bewertung von nichtkooperativ präskriptiven MAL Algorithmen. Dennoch finden diese in der Literatur jen-

seits der Bewertung von MetaStrategy für einfache gemischte Strategien und Manipulator für bedingende Gegner kaum spezifische Anwendung (vgl. Powers & Shoham, 2005a, 2005b). Formale Kriterien sind dabei stets als Grundvorraussetzungen für erfolgreiche MAL Algorithmen zu verstehen, an welche sich empirische Tests anschließen müssen (vgl. Powers & Shoham, 2005b; Shoham et al., 2007).

2.3.3 Experimentelle Validierung

Folgerichtig werden etwa US-L*, MetaStrategy und Manipulator einer ausführlichen experimentellen Validierung unterzogen (vgl. Carmel & Markovitch, 1996; Powers & Shoham, 2005a, 2005b), die jedoch substantielle Mängel in Bezug auf die Vernachlässigung der *Lernkosten*, dem Vorhandensein von *strukturellen Informationsvorteilen* aufweist. Nachfolgend sollen diese hergeleitet werden.

2.3.3.1 Lernkosten

Lernen im Rahmen von wiederholten Spielen ist mit *Kosten* verbunden, die durch entgangene Auszahlungen und Interdependenzen zwischen Agenten entstehen können. Sowohl Carmel und Markovitch (1996), als auch Powers und Shoham (2005a) entwickeln mit US-L* und Manipulator jeweils einen Algorithmus, der im Gegensatz zu den meisten anderen MAL Lösungen in der Lage ist, reichhaltigere Interaktionen zu erfassen, die entstehen können, wenn das adaptive Verhalten von Spielern auf der kürzlichen Spielhistorie bedingt. Dennoch zeichnen die beiden Ansätze sich durch einen vergleichsweise hohen Informationsbedarf aus. In der experimentellen Untersuchung wird von den Autoren zum Ausgleich eine nicht gewerteten *Explorationsphase* zugestanden, was aus Sicht dieser Arbeit die Aussagekraft der Ergebnisse für die Leistung der Gesamtlösung in Frage stellt.

Im Falle von US-L* erfolgt das Lernen des Gegnermodells offline durch Beobachtung des zukünftigen Gegenspielers in 1.000 Spielen gegen andere Agenten. Auf Basis dieser Beobachtungen wird ein Gegnermodell geschätzt, die beste Antwort errechnet und erst dann im Spiel gegen den zu modellieren Gegner eingesetzt (vgl. Carmel & Markovitch, 1996). Zur Behebung der Schwäche entwickeln Carmel und Markovitch (1998) nachfolgend den Iterativen US-L* (IT-US-L*) Algorithmus, der gegen DFA-Spieler auch zu interaktivem online Lernen fähig ist. Manipulator von Powers und Shoham (2005a) dagegen unterläuft weiterhin eine interaktive aber nicht bewertete Explorationsphase, in welcher er gemischte Strategien mit dem Ziel einsetzt, Beobachtungen zu möglichen Gegnerstrategien zu erzeugen. In Folge verliert die Effizienz des Lernalgorithmus[18] durch Missachten der Parametrisierungskosten an Bedeutung für die Bewertung. Die Verwendung einer Explorationsphase führt die dem MAL zugrundeliegende

[18] Die Effizienz des Lernalgorithmus im Sinne dieser Arbeit ist das Verhältnis seiner Leistung zu der Menge erforderlicher Spieldaten.

Problemstellung ad absurdum, indem für die Bewertung lediglich ein Teilaspekt des Lernergebnisses ohne die Kosten des Lernprozesses selbst verwendet wird. Zwar kann der Manipulator Algorithmus entgegen des üblichen Fokusses auf einfache gemischte Strategien bei anderen MAL Lösungen auch bedingende Gegner berücksichtigen, beschränkt sich aber bei der Leistungsmessung lediglich auf einen eingeschwungen Spielzustand, sodass dessen empirische Validierung gegen andere Lernalgorithmen auf Basis der letzten 20.000 Runden von je 200.000 Runden langen Interaktionen bewertet wird. Der Anteil der nicht gewerteten Explorationsphase liegt bei absolut 180.000 Runden, was 90% der gesamten Interaktion ausmacht (vgl. Powers & Shoham, 2005a). Aus Sicht dieser Arbeit handelt es sich hierbei nicht um eine zielführende Bewertungsmethodik für MAL Lernalgorithmen, welche aufgrund der Verhältnismäßigkeit von Informationsbedarf und Lernkosten *online*, also in Echtzeit, stattfinden sollten um die Leistungsfähigkeit des getesteten Algorithmus über die Gesamtrundenzahl nicht in Frage zu stellen.

2.3.3.2 Informationsasymmetrien

Weiterhin findet die empirische Leistungsbewertung der bedingenden MAL Spieler Manipulator und IT-US-L* auf Basis *heterogener Informationen* statt. Eine generell optimale Strategie oder ein generell optimaler Lernalgorithmus existiert nicht. Der individuelle Erfolg hängt immer von der Population an Gegenspielern ab. Auf Basis dieser Erkenntnis formulieren Powers und Shoham (2005b) ihre formalen Erfolgskriterien abhängig von der Klasse an Zielgegnern. Trotz der Sinnhaftigkeit dieser Perspektive entsteht hieraus im Rahmen der experimentellen Validierung gegen andere Algorithmen ein asymmetrischer Informationsvorteil zugunsten des von den Forschern vorgeschlagenen neuen Lösungen.

Erstens konnten sowohl Powers und Shoham (2005a), als auch Carmel und Markovitch (1998) eine Speziallösung für die ausgewählte und klar zugeschnittene Klasse der Gegnerspieler entwerfen. In der Realität sind derartige Grenzen jedoch nicht immer vorhanden oder identifizierbar und es können Spieler aus verschiedenen Klassen präsent sein. Weiterhin wurden die Gegnerspieler nicht notwendigerweise als eine vergleichbare Nischenlösung konzipiert, weshalb die gezeigte Leistungsdifferenz nicht notwendigerweise auf allgemein leistungsschwache MAL Algorithmen schließen lässt.

Zweitens hatten die Forscher über die Gegnerklasse hinaus auch Zugang zu der vollständigen spezifischen *Struktur* und *Parametrisierung* der Lern- und Entscheidungsmechanismen der anderen Spieler. Dieser Informationsvorteil konnte bei der Entwicklung der eigenen Lösung vorteilhaft im Sinne eines Reverse Engineerings der Verhaltensweisen der anderen Agenten genutzt werden. Ein vergleichbarer Vorteil kommt den anderen Turnierteilnehmern jedoch nicht zugute, sodass für die Turnierteilnehmer unterschiedliche Wettbewerbsvoraussetzungen gegeben sind.

Trotz vielfältigen wissenschaftlichen Ansätzen zur Erklärung ihres Aktionsverhaltens ist die Kenntnis einer expliziten Lern- und Entscheidungslogik aufgrund der Komplexität menschli-

chen Handelns ausgeschlossen. Da sich menschliche Spieler in ausführlichen empirischen Auswertungen als näherungsweise auf die Aktionshistorie bedingende Akteure modellieren lassen (vgl. Müller, 2018), bietet sich die Validierung gegen Menschen insbesondere bei bedingenden MAL Algorithmen an. Gerade durch die zunehmende Relevanz von Mensch-Maschine-Interaktionen gestaltet sich eine derartige Betrachtung als interessant. Weiterhin ist der Mensch als natürlicher Agent, der in nichtkooperativen Umgebungen mit mehreren Spielern lernen und interagieren kann ein offensichtlicher Benchmark für synthetische MAL Agenten. Dieser Aspekt gewinnt durch das zugrundeliegende Forschungsziel von Vorhaben zu nichtkooperativ präskriptiven MAL, möglichst effektive Strategien für interessante Umgebungen zu identifizieren an Bedeutung. Dabei sind *effektive Strategien* als solche zu verstehen, die in ihrer Umgebung eine möglichst hohe Auszahlung erwirtschaften können, wobei eine Haupteigenschaft der Umgebung die ihr innewohnenden Gegner sind. Die Auswahl der Gegner selbst sollte hierbei als interessant und relevant motiviert sein (vgl. Shoham & Powers, 2014a; Shoham et al., 2007). Überraschenderweise findet eine derartige Bewertung weder für die genannten bedingenden adaptiven Lösungen statt, noch ist sie überhaupt zentraler Bestandteil der Bewertung von präskriptiven MAL Algorithmen. Eine derartige Betrachtung wird womöglich vermieden, da sie weitaus weniger trivial ist und Einflussfaktoren auf spezifische Spielverläufe weitaus weniger transparent sind. Sich dem beschriebenen Trend widersetzend loben S.34, Albrecht und Stone (2018, vgl.) die Durchführung von Experimenten mit MAL Agenten im Spiel gegen menschliche Gegner als forschungsrelevantes Teilgebiet aus. Gerade die Eigenschaft, dass das Verhalten des zu modellierenden Agenten sich im Zeitverlauf ändern kann, wird hier als besonders präsent vermutet. Derartige temporale Effekte werden von den beschriebenen adaptiven Methoden nicht erfasst.

2.3.4 Zusammenfassung der Forschungslücken

Zusammenfassend lässt sich feststellen, dass präskriptive MAL Lösungen selten für reichhaltige Interaktion ausgelegt werden. Weiterhin besteht ein Mangel an Forschung, welche sowohl formal, als auch experimentell solide validiert wird. Insbesondere sind formale Kriterien bisweilen implizit oder partiell geartet. Sofern formal belastbare Kriterien explizit vorliegen, konnten im Modus der experimentellen Validierung Schwachstellen identifiziert werden:

1. **Konzeption:** Im Kontext nichtkooperativ präskriptiver MAL Algorithmen werden nur wenige Lösungen für bedingende Agenten angeboten, die eine reichhaltige, auf die Aktionshistorie bedingende Interaktion zulassen, obwohl es sich dabei um eine der charakterisierenden Eigenschaften des MAL handelt. Ausnahmen hierzu bilden Manipulator (Powers & Shoham, 2005a) und IT-US-L* (Carmel & Markovitch, 1998), wobei letzterer nicht in der Lage ist gemischte Strategien abzubilden.

2. **Formale Validierung:** Eine Validierung anhand von Formalkriterien nimmt häufig nur eine partielle Perspektive auf die Leistungsbewertung. IT-US-L* beispielsweise verzichtet vollständig auf eine derartige Betrachtung, während sich Manipulator mit einer integrierten Formaluntersuchung klar vom Literaturkorpus abgrenzt.

3. **Experimentelle Validierung:** Sowohl die präskriptive MAL Forschung im Allgemeinen, als auch die bedingende Konzepte IT-US-L* und Manipulator im Speziellen weisen regelmäßig empirisch-methodische Schwachpunkte auf:

    - Vernachlässigung von Lernkosten beziehungsweise der Effizienz der Lernalgorithmen, welche sich direkt auf die Auszahlungsleistung der Algorithmen auswirken und somit ein unvollständiges Ergebnis liefern
    - Vorhandensein von Informationsasymmetrien bezüglich der Gegnerspieler, welche das Experimentdesign in Frage stellen
    - Fokus auf rein synthetische Algorithmusgegner im Rahmen experimenteller Validierungen

2.4 Forschungsgegenstand und Forschungsfragen

Den Grundlagenteil abschließend soll hier auf Basis der identifizierten Forschungslücke der Forschungsgegenstand der Arbeit festgelegt werden. Darauffolgend werden Forschungsfragen und eine Vorgehensskizze zu deren Beantwortung definiert.

2.4.1 Beschreibung des Forschungsgegenstandes

Unter Einbezug der zuvor aufgezeigten Schwächen der bestehenden Literatur befasst sich diese Dissertation mit der Entwicklung eines präskriptiven Lernalgorithmus im Kontext nichtkooperativer Interaktion mit mehreren lernenden Agenten. Eine klare Abgrenzung des Anwendungskontextes Zielsetzung von MAL Forschungsvorhaben ist Voraussetzung für wertstiftende Ergebnisse (vgl. T. Sandholm, 2007; Shoham & Powers, 2014a; Shoham et al., 2007). In diesem Sinne liegt der Fokus auf der Entwicklung eines den eigenen absoluten Payoff maximierenden Agenten für wiederholte 2x2 Spiele mit bekanntem Aktionsraum und bekannter Auszahlungsmatrix, bei der die Aktionen der Spieler beobachtet werden können. Dabei soll einer stringenten empirischen wie auch formalen Validierung Rechnung getragen werden.

Ein erfolgreicher präskriptiver MAL Algorithmus soll in einer gegebenen Landschaft eine hohe Auszahlung erzielen, wobei eine Haupteigenschaft der Landschaft die Klasse möglicher Gegner ist. Diese Gegnerklasse sollte von natürlichem Interesse sein (Shoham & Powers, 2014a; Shoham et al., 2007), da andernfalls die Relevanz des Interaktionsergebnisses geringer ausfällt. Reichhaltige strategische Interaktionen zwischen adaptiven Spielern erfassen zu können,

die auf eine Partition der Spielhistorie bedingen ist ein sowohl inhaltlich als auch methodisch interessantes, wie auch aktuelles Forschungsobjekt (vgl. Carmel & Markovitch, 1998; Müller, 2018; Powers & Shoham, 2005a). Der zu entwickelnde MAL Agent soll daher in der Lage sein, derartig bedingtes strategisches Verhalten zu erkennen und zu modellieren. Für die Umsetzung eines bedingenden adaptiven Agenten haben sich modellbasierte Ansätze wie Manipulator und IT-US-L* als belastbar herausgestellt. Dennoch überzeugen die Lösungen wie in Kapitel 2.3 skizziert nicht vollständig. Im Rahmen dieser Arbeit soll daher ein neuer Ansatz erarbeitet werden. Modellbasierte MAL Agenten setzten sich aus zwei Modulen zusammen; erstens einem Gegnermodell, das die Frage beantwortet, wie das Verhalten des anderen Spielers zu beschreiben ist und zweitens einer Antwortlogik die für ein gegebenes Gegnermodell adressiert, wie der Agent optimaler Weise darauf reagieren soll (vgl. Shoham et al., 2007). Für die performante Modellierung von bedingenden adaptiven Gegnern, insbesondere Menschen, konnte Müller (2018) im Rahmen einer umfassend angelegten experimentellen Untersuchung ihr Konzept der bedingten Markovstrategien empfehlen. Diese beschreiben dabei Häufigkeitsschätzer für Übergangswahrscheinlichkeiten von spezifischen Zugfolgen hin zu den Aktionen des Gegners. Insbesondere zeigt die Untersuchung die Übertragung des deskriptiven Markovmodells auf einen präskriptiven Computerbot, der reichhaltige Interaktionen mit Gegnern sophistizierter Rationalität abbilden kann, als vielversprechendes Forschungsvorhaben auf (Müller, 2018, S. 148). Für den modellierenden Aspekt dieser Arbeit soll daher auf das vielversprechende Konzept der Markovstrategien zurückgegriffen werden. Neben seiner schlanken und transparenten Modellierungslogik zeichnet sich die Lösung durch seine Flexibilität über sämtliche 2x2 Spiele, seine potente Vorhersagekraft bereits auf Basis weniger Runden und durch seine Fähigkeit aus, gemischte Strategien abbilden zu können. Letzteres ist im Kontext von Spielen gegen menschliche Akteure aufgrund von Fehlern und explorativem Verhalten von Bedeutung. Die Methode der Markovstrategien aus dem Bereich der deskriptiven MAL Forschung bietet sich aufgrund ihrer vorteilhaften Eigenschaften für einen Transfer in den präskriptiven Kontext im Sinne eines interaktiven MAL Agenten als vielversprechende Ausgangslage an und wird daher von Müller (2018, S. 148) als explizites Feld zukünftiger Forschung ausgewiesen.

Eine experimentelle Untersuchung der Effektivität des MAL Agenten unabdingbar. KI Literatur beruft sich häufig auf formelle Kriterien als Bewertungsmethode für Lernalgorithmen. Diese sind wichtig, jedoch nicht hinreichend für eine praktische Validität der Algorithmen (S.34 Albrecht & Stone, 2018, vgl.). Obwohl formale Kriterien essentiell für die Bewertung sind, hat sich im Bereich der Informatik gezeigt, dass viele Algorithmen in der Praxis scheitern, obwohl sie Formalkriterien erfüllen und umgekehrt (vgl. Shoham, 2008; Shoham et al., 2007).

Der zu entwickelnde MAL Agent soll möglichst effizient lernen, sodass er bereits nach wenigen Spielrunden hinreichend belastbare Gegnermodelle produziert, auf Basis derer schnell produktive Antwort-Strategien bestimmt werden können. Damit Aspekte wie Lernkosten und

strategische Implikationen explorativen Verhaltens Einfluss im Sinne einer ganzheitlichen Bewertung finden, wird auf eine Explorationsphase verzichtet und der gesamte Spielverlauf als Auswertungsgrundlage herangezogen. Klar ist, dass der Effizienzfokus die Herausforderung verstärkt wird, mit einem Gegner ab der ersten Runde auszahlungswirksam zu interagieren. Somit muss zu das gleichzeitige online Lernen des gegnerischen Verhalten mit der unmittelbaren Payoffmaximierung balanciert werden. Gleichzeitig soll der Aspekt der Lerneffizienz dadurch hervorgehoben werden, dass eine im Literaturvergleich (vgl. Carmel & Markovitch, 1996, 1998; Powers & Shoham, 2005a, 2005b) kurze Rundenzahl verwendet wird (siehe Kapitel 4.1.3).

Für die experimentelle Validierung bieten sich menschliche Akteure als natürlicher Nullpunkt für die Bewertung eines lernenden und bedingenden adaptiv interaktiven Agenten an: Erstens handelt es sich um einen von der Forschung zu präskriptiven Agenten vernachlässigten Aspekt der experimentellen Validierung. Zweitens verhindert das Spiel gegen menschliche Gegner Informationsvorteile bezüglich des Lern- und Entscheidungsverhaltens seitens der Experimentleiter. Drittens handelt es sich insofern um eine strengere Leistungsanforderung, als Menschen keine klar definierte einer einheitlichen Logik folgenden Klasse von Gegnern handelt und dem Algorithmus somit ein gewisses Maß an Generalität abverlangt wird.

Ergänzend dazu soll auch dem üblichen Vorgehen zu experimenteller Validierung anhand algorithmischer Gegner (vgl. Carmel & Markovitch, 1996, 1998; Nudelman et al., 2004; Powers & Shoham, 2005a, 2005b) gefolgt werden. Dafür bietet sich das Strategieturnier von Axelrod (1980) durch seinen wohldiskutierten und wegweisenden Charakter als ergänzende Untersuchung an. Neben einer empirischen Untersuchung erfolgt auch eine Diskussion der Leistungsfähigkeit in Hinblick auf formale Leistungskriterien, da diese ein abstrahierendes Verständnis über Stärken und Schwächen der Lösung fördern können.

In Summe soll der Betrachtungsgegenstand helfen, die identifizierten Lücken in Hinblick auf Konzeption, formale Validierung und experimentelle Validierung von nichtkooperativ präskriptiven MAL Algorithmen (siehe Kapitel 2.3.4) zu schließen. Dafür werden die nachfolgend definierten Forschungsfragen zugrunde gelegt.

2.4.2 Ableitung der Forschungsfragen

Auf Basis des vorangehend skizzierten Forschungsgegenstandes wird die Entwicklung einer spielunspezifischen Methodik zur adaptiven Interaktion mit menschlichen Gegnern in unendlichen wiederholten 2x2 Spielen als Zielsetzung der Arbeit festgehalten. Insbesondere sollen hierbei die vielversprechenden Erkenntnisse von Müller (2018) zur Vorhersage individuellen menschlichen Spielverhaltens mit Markovstrategien Einfluss finden. Der Sachverhalt wird mit Hilfe folgender primärer Forschungsfragen adressiert:

1. **Entwicklung und experimentelle Validierung des Markovagenten im Spiel mit Menschen:** Wie kann ein adaptiv-interaktiver, auf die Spielhistorie bedingender Markovagent, welcher wesentliche Aspekte individuellen Gegnerverhaltens effektiv berücksichtigt, konzipiert werden und wie schneidet dieser im Prisoner's Dilemma in einem ersten experimentellen Vergleich mit menschlichen Spielern im Rahmen einer Prestudy ab?

2. **Weiterführende experimentelle Validierung des Markovagenten im Spiel mit Menschen:** Wie schneidet der Markovagent bei weiterführenden Untersuchungen zum Prisoner's Dilemma, Chicken Game und Hero Game als ausgewählte nichtkooperative wiederholte 2x2 Spiele im experimentellen Vergleich mit menschlichen Spielern ab?

Zur Beantwortung der Forschungsfragen wird eine Vorgehensskizze festgelegt. Zunächst findet die konzeptionelle Entwicklung des Markovagenten statt. Im Anschluss wird die empirische Leistungsfähigkeit relativ zu menschlichen Spielern im Prisoner's Dilemma mit Menschen im Rahmen einer ersten Prestudy sichergestellt. Dem Zwischenfazit schließt sich eine umfassende experimentelle Validierung anhand drei ausgewählter wiederholter 2x2 Spiele an. Dabei wird der Markovagent im experimentellen Vergleich mit menschlichen Gegnern übergreifend im Rahmen einer weiterführenden Validierung über die Spiele *Prisoner's Dilemma*, *Chicken Game* und *Hero Game* untersucht.

Im Rahmen der Diskussion der Hauptergebnisse sollen schließlich sekundäre Forschungsfragen aufgegriffen werden, um die die Hauptergebnisse anhand flankierender Fragestellungen anzureichern und ein ganzheitliches Leistungsbild zu ermöglichen.

3. **Formale Validierung:** Wie schneidet der entwickelte Markovagent unter formalkritischen Gesichtspunkten ab?

4. **Experimentelle Validierung des Markovagenten im Spiel mit algorithmischen Gegnern:** Wie schneidet der entwickelte Markovagent im Turnier mit algorithmischen Gegnern ab?

Das nächste Kapitel befasst sich unmittelbar mit der Entwicklung des angestrebten Markovagenten.

3 Entwicklung eines adaptiven Markovagenten für wiederholte Spiele

Im Folgenden soll zunächst die konzeptionelle Funktionsweise eines Markovagenten anhand Kapitel 3.1 formalisiert erarbeitet und darauffolgend in Kapitel 3.3 dessen Anwendung auf die Implementierung als Programm im Sinne eines interaktiven Agenten dokumentiert werden. Die Modellierung und Vorhersage des Gegnerverhaltens im Rahmen dieser Arbeit bauen auf dem von Müller (2018) vorgeschlagenen Modell zur individuellen Verhaltensvorhersage anhand geschätzter Markovstrategien auf. Wesentliche Herausforderung dabei ist, dass nicht die Vorhersage des nächsten Zuges, sondern der gesamten Strategie des Gegners im Zentrum steht, auf Basis derer der Agent seine Antwort bestimmt. Das Problem zerfällt demnach in die Teilprobleme:

- **Gegnermodellierung**: Wie wird der Gegner sich verhalten?
- **Antwortbestimmung:** Wie soll sich der Agent auf Basis der Vorhersage verhalten?

Ist der Gegner ebenfalls ein lernender Agent, beispielsweise ein Mensch, kommt erschwerend hinzu, dass das Vorhersageergebnis direkten Einfluss auf die Aktionswahl des Agenten hat. Dies kann sich im Falle eines lernenden gegnerischen Agenten unmittelbar auf die Trajektorie des Spiels auswirken.

3.1 Konzeptionelle Grundlagen zu Markovinteraktionen

Zunächst findet eine kurze Einordnung der Anknüpfungspunkte dieser Arbeit in bestehende Forschung zu Markovinteraktionen statt. Dazu nimmt dieses Kapitel eine Einordnung in bestehende Konzepte vor, stellt grundlegende Gestaltungsentscheidungen vor und formuliert Annahmen für das weitere Vorgehen.

Die rein deskriptive Methode *IP-EMS* (Individual Prediction based on Estimation of Markov Strategies) zur Verhaltensvorhersage individuellen Spielerverhaltens in wiederholten Spielen von Müller (2018) ist eine Weiterentwicklung der bedingten Wahrscheinlichkeiten zur Vorhersage von Zugverhalten auf Basis vorangegangener Runden eines Spiels. Diese geht auf Breitmoser (2015) zurück. In diesem Kontext sind Markovstrategien auf vorige Spielzustände bedingende Schätzer begrenzter Komplexität. Ziel dieser gedächtnisbasierten Schätzer ist es, die

Wahrscheinlichkeit des Gegners eine bestimmt Aktion gegeben eines Spielzustandes auf Basis seines vorangegangenen Verhaltens in dem selben Spielzustand zu geben (S. 33 Müller, 2018).

Markovstrategien greifen auf das Konzept der Markovketten zurück. Letztere modellieren Systeme zufälliger Veränderung mit beschränktem Gedächtnis anhand stochastischer Prozesse bestehend aus *Zuständen* und *Übergangswahrscheinlichkeiten* zur Vorhersage zukünftiger Entwicklungen. Markovketten verfügen über folgende Gestaltungsparameter (vgl. Häggström, 2002):

- **Auflösung:** Diskrete oder kontinuierliche Darstellung
- **Ordnung:** Gedächtnistiefe der berücksichtigten vorangegangenen Perioden
- **Zustandsraum:** Endliche oder unendliche Zustandszahl

Zur Abbildung endlicher wiederholter Normalformspiele eignen sich zeitdiskrete endliche Markovketten, da die Auflösung dann, sofern nicht anderweitig limitiert, durch die Runden eines Spiels festgelegt wird und die Anzahl möglicher Zustände endlich ist. Die Ordnung der verwendeten Markovketten ist ein gestalterischer Freiheitsgrad, auf den in Kapitel 3.3.1 eingegangen wird (Müller, 2018, S. 34).

Jenseits der rein deskriptiven Vorhersage von Spielerverhalten lässt sich mit Hilfe einer derartigen Markovlogik ein interaktiver Agent entwickeln, welcher selbst an wiederholten Spielen teilnimmt. Dazu muss auf Basis der geschätzten Übergangswahrscheinlichkeiten als robuster Prädiktor zukünftigen gegnerischen Verhaltens eine beste Antwort abgeleitet werden, die es dem Markovagent erlaubt, eine effektive Auszahlungsleistung zu erzielen (vgl. Müller, 2018, S. 148). Für einen derartigen interaktiven Markovagent wird im Rahmen der Arbeit auf den empirischen Erkenntnisse von Müller (2018) aufbauend angenommen, dass Markovstrategien menschliches Spielverhalten performant beschreiben können. Insbesondere deuten die Experimentalergebnisse von Müller (2018) darauf hin, dass sich menschliche Markovstrategien spezifischen Strategieclustern zuweisen lassen, was auf eine über den Spielverlauf konstante Strategie vermuten lässt. Auf Basis dieser Erkenntnisse lässt sich folgende Annahme über menschliches Spielverhalten im Rahmen dieser Arbeit treffen, deren deskriptive Gültigkeit voraussichtlich Einfluss auf das Ergebnis haben wird:

> *Die wahre Strategie des menschlichen Gegenspielers lässt sich durch eine Markovstrategie mit stationären Übergangswahrscheinlichkeiten hinreichend genau beschreiben.*

Gegeben dieser Annahme gilt, dass jede auszahlungsmaximierende Antwort-Strategie auf eine gegebene Markovstrategie ebenfalls in der Menge der Markovstrategien zu finden ist. Grund

ist, dass jedwedes mögliche Gegnerverhalten sich in dessen Übergangswahrscheinlichkeiten im Sinne spielzustandsbedingter Aktionswahrscheinlichkeiten niederschlägt. Für jede Kombination gegnerischer Übergangswahrscheinlichkeiten muss es mindestens eine Kombination an Übergangswahrscheinlichkeiten des eigenen Spielers geben, die eine in Bezug auf die erwarteten Payoffs eine optimale zukünftige Trajektorie des Spielverlaufs für den eigenen Spieler induziert (vgl. Papadimitriou & Tsitsiklis, 1987; vgl. Carmel & Markovitch, 1998, S. 314-315):

> *Die beste Antwort-Strategie im Sinne einer Maximierung der erwarteten eigenen durchschnittlichen Auszahlung für eine gegebene gegnerische Markovstrategie ist ebenfalls in der Menge der Markovstrategien zu finden.*[19]

Für die Umsetzung eines Markovagenten ist neben den soeben präsentierten Punkten eine robuste Interaktionslogik erforderlich. Diese soll, um die anschließende algorithmische Umsetzung vorzubereiten, zunächst anhand einer möglichst generellen Formallogik erarbeitet werden.

3.2 Formalisierung adaptiver Agenten

Für die formale Darstellung des adaptiven Markovagenten für Zweipersonenspiele, soll zunächst in Kapitel 3.2.1 anhand einer Logik für allgemeine modellbasierte Lernalgorithmen eine Grundlage geschaffen werden, auf Basis derer im darauffolgenden Kapitel 3.2.2 die Erarbeitung eines Formalmodells für adaptive Markovagenten stattfindet.

3.2.1 Lernalgorithmus für modellbasierte deterministische Agenten

Die sich hier anschließende Formallogik für einen unspezifischen deterministischen modellbasierten lernenden Agenten orientiert sich maßgeblich an den Ausführungen von Carmel und Markovitch (1998, S. 310-313).

3.2.1.1 Grundlegendes zu deterministische Agenten

Als Grundlage für einen deterministischen modellbasierten lernenden Agenten werden zunächst Aspekte charakterisiert, die einen allgemeinen deterministischen Agenten und dessen Aktionskontext definieren.

Stufenspiel Ein Zweipersonenspiel ist definiert als das Tupel $G = (A^1, A^2, u^1, u^2)$. Das Tupel besteht aus der den endlichen Aktionsmengen A^i der beiden Spieler $i, j \in \{1,2\}$ mit $i \neq j$ sowie deren Auszahlungsfunktion $u^i : A^1 \times A^2 \to \mathbb{R}$. Die Auszahlungsfunktion gibt mit r^i die für einen gemeinsamen simultanen Zug (a^1, a^2) für Spieler i realisierte Auszahlung an.

[19] Im Extremfall kann für jeden denkbaren Spielverlauf eine deterministische optimale Strategie durch eine Markovstrategie mit einer Ordnung gleich der Spiellänge deterministisch abgebildet werden.

Wiederholtes Spiel Analog dazu charakterisiert $G^{\#} = (S^1, S^2, U^1, U^2)$ ein wiederholtes Spiel über T Runden auf Basis des Stufenspiels G, sodass die Spieler in jeder Runde $t \leq T$ ihre Aktionen $(a_t^1, a_t^2) \in A^1 \times A^2$ wählen gemäß ihrer s^i aus der Menge ihrer Strategien S^i. Die Auszahlungsfunktion des wiederholten Spiels $U^i : S^1 \times S^2 \to \mathbb{R}$ legt den Nutzen einer durch die Strategiekombination (s^1, s^2) charakterisierten wiederholten Interaktion fest.

Spielhistorie Die von den beiden Spielern in jeder Runde t des wiederholten Spiels $G^{\#}$ simultan ausgeführten Aktionen (a_t^1, a_t^2) definieren die Spielhistorie $h(t)$ als eine endliche Folge von gemeinsamen Zügen ab der ersten Runde bis zur vorigen Runde $t-1$. $\emptyset$ bezeichnet die leere Historie für $h(1)$.

$$h(t) = \begin{cases} \emptyset & \text{falls } t = 1 \\ ((a_{t_1}^1, a_{t_1}^2), ..., (a_{t-1}^1, a_{t-1}^2)) & \text{sonst} \end{cases} \tag{3.1}$$

Deterministische Strategie Eine Strategie s^i von Spieler i legt für eine gegebene Spielhistorie $h(t) \in H(G^{\#})$ dessen nachfolgendes Verhalten fest. Ist die Strategie deterministisch gilt $s^i : H(G^{\#}) \to A^i$, sodass für jede Spielhistorie eine eindeutige Aktion $a_t^i \in A^i$ für den folgenden Zug definiert werden kann.

Deterministischer Pfad Ein bekanntes deterministisches Strategiepaar (s^1, s^2) definiert einen *Pfad* im Sinne eines bereits ex ante identifizierbaren Spielverlaufs $\hat{h}$ für $G^{\#}$, sodass eine Beziehung im Sinne von $s^i : \hat{H}(G^{\#}) \to A^i$ besteht. Auf Basis der Vorhersage über die vorige Runde kann so eine Vorhersage über das Verhalten in der nächsten Runde gemacht werden. Es handelt sich dabei um eine unendliche Folge von gemeinsamen Zügen.

$$\hat{h}_{(s^1,s^2)}(t) = \begin{cases} \emptyset & \text{falls } t = 0 \\ \hat{h}_{(s^1,s^2)}(t-1) \parallel (s^1(\hat{h}_{(s^1,s^2)}(t-1)), s^2(\hat{h}_{(s^1,s^2)}(t-1))) & \text{sonst} \end{cases} \tag{3.2}$$

Dabei kennzeichnet $\parallel$ den Verkettungsoperator zweier Tupel. Der Pfad ist im Gegensatz zur *rückwärts* gerichteten Historie auch für noch folgende Spielrunden definiert. Es gilt $\hat{h}_{(s^1,s^2)}(t) = h(t)$, sofern es sich bei s^i um deterministische Strategien handelt.

3.2.1.2 Beschreibung des Lernalgorithmus

Das folgende Kapitel erarbeitet relevante Begrifflichkeiten für einen modellbasierten Lernalgorithmus in fortgeführter Anlehnung an Carmel und Markovitch (1998, S. 310-313). Maßgeblich ist dabei die folgende methodenübergreifende, für modellbasierte MAL Algorithmen typische Schrittfolge (vgl. Shoham & Powers, 2014b):

1. **Initialisierung:** Der Lernalgorithmus startet in Runde $t = 0$ mit einem beliebigen Gegnermodell $\hat{s}_0^j$.

2. **Berechnung der besten Antwort:** Für das Gegnermodell berechnet der MAL Agent die beste Antwort-Strategie $s_t^i = s^{*i}(\hat{s}_t^j, U^i)$ auf Basis seiner Nutzenfunktion. Mit Hilfe der besten Antwort-Strategie bestimmt der Lernalgorithmus seine nächste Aktion $a_t^i = s_t^i(h(t))$.

3. **Aktualisierung des Gegnermodells:** Mit Hilfe der bis dato gesammelten Stichproben des Gegnerverhaltens aktualisiert der Agent das Gegnermodell gemäß $\hat{s}_t^j = L^i(D^j(h(t)))$.

4. **Wiederholung:** Solange das Spiel nicht beendet ist, gehe zu Schritt 2.

Beste Antwort-Strategie Für eine gegebene Auszahlungsfunktion U^i ist $s^{*i}(s^j, U^i)$ genau dann als die beste Antwort-Strategie auf eine gegebene gegnerische Strategie s^j definiert, wenn $\forall s \in S^i : U^i(s^{*i}(s^j, U^i), s^j) \geq U(s, s^j)$. Die Nutzenfunktion kann zur Bestimmung der besten Antwort-Strategie im endlich wiederholten Spiel als Durchschnitt der Auszahlungen gemessen werden, welche der simulierte Pfad des Strategiepaars induziert (vgl. Shoham & Powers, 2014a, S. 4; Shoham & Leyton-Brown, 2008):[20]

$$U^i(s^1, s^2) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} u^i(s^1(\hat{h}_{(s^1,s^2)}(t)), s^2(\hat{h}_{(s^1,s^2)}(t))) \tag{3.3}$$

Folglich ist die dem Begriff der besten Antwort-Strategie zugrundeliegende Zielfunktion, die eigene Nutzenfunktion des wiederholten Spiels im Sinne des Durchschnitts der Auszahlungen zu maximieren. Klar ist, dass es für eine gegebene gegnerische Strategie mehrere beste Antwort-Strategien geben kann. Zum Beispiel sind für Tit-for-Tat sowohl Always Cooperate, als auch Grim Trigger, als auch Tit-for-Tat selbst beste Antworten des wiederholten Spiels mit unbekanntem Ende.

Gegnermodell Als zentrale Eigenschaft des modellbasierten MAL unterhalten lernende Agenten eine Hypothese über das zukünftige Verhalten des Gegners. Dieses Gegnermodell wird über die als *allgemein bekannt* angenommenen und somit im Spielverlauf beobachtbaren Handlungsoptionen A^i der Spieler gebildet. Das beobachtbare Spielerverhalten ist im Sinne der *Revealed Preferences* ein Spiegel der Präferenzen (Richter, 1966, vgl.). Aufgrund unterschiedlicher Annahmen über die anderen Spieler können unterschiedliche Spieler trotz einer geteilten Spielhistorie zu unterschiedlichen Vorhersagen über den weiteren Spielverlauf kommen. Die

[20] Als alternative Nutzenfunktion ist insbesondere für unendlich wiederholte Spiele die diskontierte Summe aller Auszahlungen denkbar (vgl. Carmel & Markovitch, 1998, S. 311-312). Im Rahmen dieser Arbeit wurde hierauf aufgrund der Sensitivität der besten Antwort hinsichtlich des Zinsfaktors verzichtet.

Vorhersage von Spieler i ist demnach formal durch $\hat{h}_{(s^i,\hat{s}^j)}$ gegeben, wobei s^i die eigene Strategie bezeichnet und $\hat{s}^j$ als Gegnermodell die aktuelle Hypothese über das gegnerische Verhalten widerspiegelt.

Lernalgorithmus Ein modellbasierter lernender Agent besteht folglich aus den zwei Modulen, erstens dem auf Basis der Historie gebildeten *Gegnermodells* und zweitens der dafür berechneten *Antwort-Strategie* (vgl. Shoham & Powers, 2014b, S. 1). Das Gegnermodell eines Lernalgorithmus L kann somit formal auf Basis einer Menge an Beobachtungspaaren $(h(t), a_t^j)$ gebildet werden, wobei $h(t)$ die Historie des wiederholten Spiels zu Zeitpunkt t und a_t^j die durchgeführte gegnerische Aktion zu Zeitpunkt t darstellen. Die verfügbare Stichprobe D^j des Lernalgorithmus ist die Kette der bis zum aktuellen Zeitpunkt realisierten Beobachtungspaare des gegnerischen Verhaltens. Die Stichprobe ist demnach $D^j(h(t)) = ((h(k), a_k^j) | 1 \leq k \leq t)$. Auf Basis dieser Stichprobe bestimmt der Lernalgorithmus das aktuelle Gegnermodell $\hat{s}^j = L^i(D^j(h(t)))$. Die Antwort-Strategie des lernenden Agenten hängt sowohl vom Gegnermodell $\hat{s}^j$, als auch von der Nutzenfunktion U^i des lernenden Agenten ab. Es handelt sich unter Berücksichtigung der Überlegungen zu Gegnermodell und bester Antwort somit um eine Strategie gemäß:

$$\begin{aligned} s^i_{U^i,L^i}(h(t)) &= s^{*i}(L^i(D^j(h(t)), U^i)(h(t)) \\ &= s^{*i}(\hat{s}^j, U^i)(h(t)) \end{aligned} \tag{3.4}$$

Zusammenfassend ergibt sich auf Basis von Gleichung 3.4 ein sich unter Einbezug des Spielverlaufs strategisch adaptiv verhaltender modellbasierter Agent. Die grundlegende Funktionsweise eines solchen Spielers ist in Abbildung 3.1 aufgearbeitet. Im nächsten Schritt soll der bis hierhin allgemeine deterministische Lernalgorithmus auf den Kontext der Markovstrategien (vgl. Müller, 2018) übertragen werden.

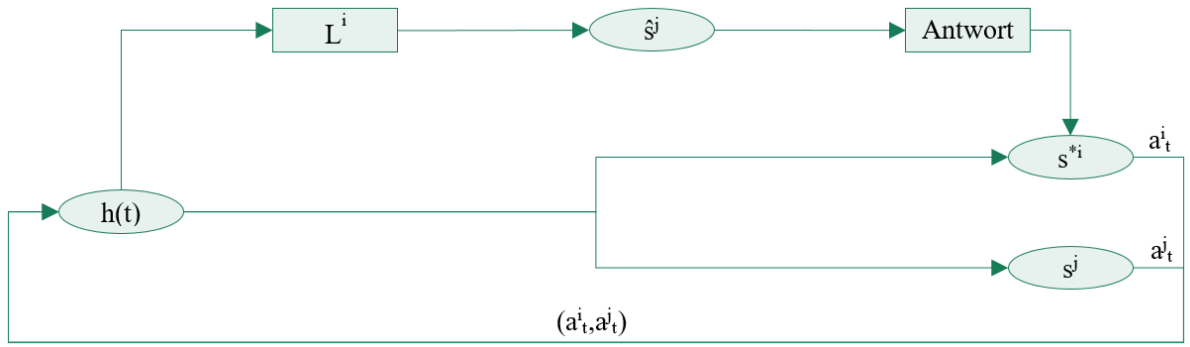


Abbildung 3.1: Interaktionslogik eines allgemeinen modellbasierten deterministischen Lernalgorithmus. Quelle: In Anlehnung an Carmel und Markovitch (1998, S. 312).

3.2.2 Lernalgorithmus für stochastisch bedingende Markovagenten

Der allgemeine modellbasierte deterministische Agent von Carmel und Markovitch (1998, S. 310-313) soll nun unter Einbezug der Ausführungen von Müller (2018) hin zu einem stochastischem, auf der Spielhistorie bedingenden Markovagenten entwickelt werden.

3.2.2.1 Grundlegendes zu stochastischen Markovagenten

Zunächst werden grundlegende Begrifflichkeiten und Konzepte zu Markovstrategien als Grundlage für die darauffolgende Gestaltung eines lernenden Markovagenten formalisiert.

Markovstrategie Im Gegensatz zur deterministischen Charakterisierung des Strategiebegriffs im vorigen Kapitel ist es für den Transfer auf den Kontext der Markovstrategien als bedingte Übergangswahrscheinlichkeiten erforderlich, die getroffene Definition auf stochastische Strategien zu erweitern. Im Falle einer stochastischen Markovstrategie gilt demnach $s^i_{Markov}: H(G^{\#}) \rightarrow \mathbf{P}[A^i]$, sodass für jede Spielhistorie eine Eintrittswahrscheinlichkeit für jede Aktion $a^i_t \in A^i$ für den folgenden Zug definiert werden kann. Eine Markovstrategie $s^i_{Markov} = (M^i, \sigma^i)$ ist ein Tupel bestehend dabei aus einer *Übergangsmatrix* M^i und einer *Initialisierungslogik* σ^i. Die Initialisierungslogik gibt das Spielerverhalten für jene anfängliche Runden an, in denen noch nicht ausreichend Beobachtungen vorhanden sind, um den Markovzustandsraum zu initialisieren. Aus Gründen übersichtlicher Notation wird nachfolgend $s^i = s^i_{Markov}$ verwendet.

Übergangsmatrix Die Markov-Übergangsmatrix $M^i_o : Z^i_o) \rightarrow \mathbf{P}[A^i]$ für Spieler i gibt die bedingten Übergangswahrscheinlichkeiten zwischen dessen Markovzuständen Z^i und dessen darauffolgender Handlung $a^i_t \in A^i$ an, wobei die Struktur des Markovzustandsraumes und damit die Struktur der Übergangsmatrix von der Ordnung o^i der verwendeten Markovkette abhängt. Es handelt sich demnach um eine Matrix folgender Struktur:

$$\begin{aligned} M^i &= (m^i_{z^i,a^i}) \\ \text{mit } m^i_{z^i,a^i} &= \mathbf{P}[a^i_t|z_t] \end{aligned} \tag{3.5}$$

Es handelt sich folglich um die bedingte Wahrscheinlichkeit der nächsten Handlung a^i_t auf Basis des aktuellen Markovzustandes z^i_t. Klar ist, dass sich die Wahrscheinlichkeiten für alle Aktionen $a^i \in A^i$ je Zustand zu einer Wahrscheinlichkeit von 100% addieren müssen, sodass gilt $\sum_{a^i \in A^i} m^i_{z^i,a^i} \overset{!}{=} 1 \forall z^i \in Z^i$. Auf Basis dieser Nebenbedingung wird die Übergangswahrscheinlichkeit m^i für 2x2 Spiele nachfolgend vereinfacht mit $m^i_{z^i,a^i_2} = \mathbf{P}[a^i_{2,t}|z_t]$ als die Wahrscheinlichkeit für die zweite Aktion $a^i_{2,t}$ des Spielers i auf Basis des Zustandes z^i_t angegeben. Die Wahrscheinlichkeit für $a^i_{1,t}$ ergibt sich logischerweise unmittelbar aus $\mathbf{P}[a^i_{1,t}] = 1 - \mathbf{P}[a^i_{2,t}]$.

Markovzustand Der Markovzustand z_t^i eines Spielers gibt an, in welcher Entscheidungssituation sich der Markovagent befindet. Es handelt sich bei dem Markovzustand um eine Partition des kürzlichen Endes der Historie $z^i = h(t) \setminus h(t-o^i)$, wobei o^i als *Markovordnung* im Sinne einer Gedächtnistiefe angibt, wie viele der vorangegangenen Perioden für die Interaktionslogik des Agenten mit einbezogen werden (siehe Kapitel 3.1).

Markovordnung Die Ordnung oder *Gedächtnistiefe* o^i eines Markovagenten i legt fest, über wie viele der vergangen Runden sich das Gedächtnis der Markovkette erstreckt. Die möglichen Zustände eines Markovagenten mit Ordnung o^i ist die Menge aller möglichen Spielverläufe des wiederholten Spiels $G^{\#}$ nach $t = o^i$ Runden, also $Z^i_{o^i} = H(G^{\#}, t = o^i)$. Die Gedächtnistiefe wird technisch als Markovkette der Ordnung o^i abgebildet. Für eine spezifische Runde t ist der Zustand für Markovagent i somit gegeben durch die letzten o^i Einträge der Historie $h(t)$:

$$
\begin{aligned}
z^i_{o^i,t}(h(t)) &= h(t) \setminus h(t-o^i) \\
&= \begin{cases} \emptyset & \text{falls } o^i = 0 \vee t \leq o^i \\ ((a^i_{t-o^i}, a^j_{t-o^i}), \ldots, (a^i_{t-1}, a^j_{t-1})) & \text{sonst} \end{cases}
\end{aligned} \tag{3.6}
$$

Während Müller (2018) die Gedächtnistiefe auf die ausgeführten Handlungen *aller* Spieler in den o^i letzten Runden bezieht, soll im Rahmen dieser Arbeit eine höher auflösende Definition der Markovordnung verwendet werden. Konkret sind die Markovketten in dieser Arbeit mit $O^i = (o^i_i, o^i_j)$ nach den spezifischen Spielern differenziert. o^i_i beziffert dabei die Anzahl der vorangegangenen Züge von Spieler i, welche in den Zustand von Spieler i Einfluss finden, während o^i_j analog für die Anzahl der für Spieler i relevanten vorangegangenen Züge von Spieler j angibt. Aus qualitativer Sicht entspricht die Beispielparametrisierung $O^i = (1,1)$ der Berücksichtigung des letzten Zuges beider Spieler für die bedingte Wahrscheinlichkeit des nächsten Zuges von Spieler i. Jedoch kann anders als bei Müller (2018) ebenfalls auf ein Modell der Art $O^i = (0,1)$ zurückgegriffen werden. In diesem Beispiel bedingt die Markovstrategie von Spieler i lediglich auf dem letzten Zug von Spieler j. Formal gilt:

$$z^i_{O^i,t}(h(t)) = h(t) \setminus h(t-o^i_i, t-o^i_j)$$

$$= \begin{cases} \emptyset & \text{falls } t \leq o^i_{max} \vee o^i_1 = o^i_2 = 0 \\ ((\emptyset, a^j_{t-o^i_2}), ..., (\emptyset, a^j_{t-1})) & \text{falls } t > o^i_{max} \wedge o^i_1 < o^i_2 \wedge o^i_1 = 0 \\ ((a^i_{t-o^i_1}, \emptyset), ..., (a^i_{t-1}, \emptyset)) & \text{falls } t > o^i_{max} \wedge o^i_2 < o^i_1 \wedge o^i_2 = 0 \\ ((\emptyset, a^j_{t-o^i_2}), ..., (a^i_{t-1}, a^j_{t-1})) & \text{falls } t > o^i_{max} \wedge o^i_1 < o^i_2 \wedge o^i_1, o^i_2 \neq 0 \\ ((a^i_{t-o^i_1}, \emptyset), ..., (a^i_{t-1}, a^j_{t-1})) & \text{falls } t > o^i_{max} \wedge o^i_2 < o^i_1 \wedge o^i_1, o^i_2 \neq 0 \\ ((a^i_{t-o^i_1}, a^j_{t-o^i_2}), ..., (a^i_{t-1}, a^j_{t-1})) & \text{falls } t > o^i_{max} \wedge o^i_2 = o^i_1 \wedge o^i_1, o^i_2 \neq 0 \end{cases} \tag{3.7}$$

Je Markovzustand kann der Agent eine spezifische Wahrscheinlichkeit für dessen nachfolgende Aktion besitzen. Die Anzahl und Struktur der Markovzustände wird durch die Markovordnung festgelegt, welche folglich die Interaktionslogik, welche durch den Markovagenten abgebildet werden kann maßgeblich determiniert.

Initialisierungslogik Für einen Markovspieler mit Gedächtnistiefe O^i kann in den ersten $o^i_{max} = max(o^i_1, o^i_2)$ Runden eines Spiels keine Handlungsentscheidung auf Basis der Übergangsmatrix getroffen werden. Definitionsgemäß muss eine Markovstrategie der Ordnung O^i auf eine bestehende Historie im Sinne einer Markovkette aus Aktionen der Spieler mit Länge o^i_{max} zurückgreifen können. Andernfalls ist der Zustandsraum des Markovspielers unterbestimmt. Für die anfänglichen o^i_{max} Runden ist daher eine *Initialisierungslogik* σ^i erforderlich. Nur so kann eine Markovstrategie dem Strategiebegriff im Sinne von Entscheidungsplänen für *jede* Situation eines wiederholten Spiels gerecht werden (S. 33 Müller, 2018). Die Initialisierungslogik ist ein Strategietupel und kann eine beliebige in den Runden $t = 1...o^i_{max}$ operationalisierbare Spiellogik enthalten. Beispieltypen für Initialisierungslogiken sind:

- **Deterministisch:** Eine im Vorhinein festgelegte deterministische Aktionsfolge
- **Stochastisch:** Eine im Vorhinein festgelegte fixe Verteilung, aus denen Aktionen gezogen werden
- **Sub-Markovstrategie:** Eine Markovstrategie geringerer Ordnung, wobei anzumerken ist, dass der Sonderfall einer Sub-Markovstrategie mit Ordnung $o^i = 0$ einer Random Strategie mit Wahrscheinlichkeit $m^i = \mathbf{P}[a^i_{2,t}|\emptyset]$ entspricht

Beispiel Markovstrategie Um das bis hierhin erarbeitete Formalmodell eines Markovagenten zu veranschaulichen, soll eine Beispielstrategie modelliert werden. Exemplarisch lässt sich Tit-for-Tat im wiederholten Prisoner's Dilemma als Markovstrategie darstellen:

- **Ordnung:** Die Tit-for-Tat zugrundeliegende Logik schreibt ein Zugverhalten vor, welches ausschließlich vom letzten Zug des anderen Spielers abhängt. Dementsprechend ist die Ordnung des Tit-for-Tat-Spielers $i = 1$ durch $O^1 = (0,1)$ gegeben. Es werden $o_1^1 = 0$ eigene und $o_2^1 = 1$ gegnerische letzte Aktionen bei der Entscheidungsfindung einbezogen.

- **Zustandsraum:** Der mögliche Zustandsraum entspricht demnach den beiden Aktionen des Gegnerspielers $j = 2$ mit $Z^1 = \{(\emptyset, a_{1,t-1}^2), (\emptyset, a_{2,t-1}^2)\}$.

- **Übergangsmatrix:** Auf Basis des beschriebenen Zustandsraumes ergibt sich demnach eine Übergangsmatrix mit Struktur $M_{TFT}^1 = (\mathbf{P}[a_{2,t}^1|(\emptyset, a_{1,t-1}^2)], \mathbf{P}[a_{2,t}^1|(\emptyset, a_{2,t-1}^2)])^T$. Inhaltlich imitiert Tit-for-Tat stets den letzten Zug des anderen Spielers. Somit sind die bedingten Übergangswahrscheinlichkeiten $M_{TFT}^1 = (0\%, 100\%)^T$.

- **Initialisierungslogik:** Da in der ersten Spielrunde noch kein gegnerischer Zug vorliegt, der von Tit-for-Tat imitiert werden kann, eröffnet die Strategie stets mit einer Kooperation. Die Initialisierungslogik ist also mit $\sigma^1 = (a_{1,t=1}^1)$ deterministischer Art.

- **Markovstrategie:** In Summe ergibt sich für Tit-for-Tat eine Markovstrategie $s_{TFT}^1 = ((0\%, 100\%)^T, (0\%))$ mit Zustandsraum $Z^1 = \{(\emptyset, a_{1,t-1}^2), (\emptyset, a_{2,t-1}^2)\}$.

Es sei angemerkt, dass es sich bei Tit-for-Tat um eine deterministische Strategie handelt, da $m_{TFT}^i \in i \in 0,1$. Eine Tit-for-Tat-Strategie mit zehnprozentigem Rauschen, also einer Wahrscheinlichkeit von 10% die von der reinen Tit-for-Tat-Strategie abweichende Handlung auszuführen wird durch die Markovstrategie $s_{TFTr}^1 = ((10\%, 90\%)^T, (10\%))$ mit Zustandsraum $Z^1 = \{(\emptyset, a_{1,t-1}^2), (\emptyset, a_{2,t-1}^2)\}$ beschrieben.

Stochastischer Pfad Ein bekanntes Paar stochastischer Markovstrategien (s^1, s^2) definiert einen *Pfad* im Sinne eines bereits ex ante identifizierbaren *probabilistisch erwarteten* Spielverlaufs $\hat{h}$ für $G^{\#}$. Im Falle stochastischer Strategien bestehen die Elemente von $\hat{h}$ für 2x2 Spiele statt aus konkreten Aktionen aus den Eintrittswahrscheinlichkeiten für die Aktion a_2^i je Spieler, sodass gilt $s^i : \hat{H}(G^{\#}) \rightarrow \mathbf{P}[A^i]$. Beispielhaft gestaltet sich $\hat{h}_{(s^1,s^2)}(t) = ((\mathbf{P}[a_{2,0}^1], \mathbf{P}[a_{2,0}^2]), (\mathbf{P}[a_{2,1}^1], \mathbf{P}[a_{2,1}^2]), ..., (\mathbf{P}[a_{2,t-1}^1], \mathbf{P}[a_{2,t-1}^2]))$. Der stochastische Pfad kann mit Hilfe der Parameter der Markovstrategie bestückt werden, sodass $\mathbf{P}[a_{2,t}^i] = (\mathbf{P}[a_{2,t}^i|z_t])_{z \in Z} = (m_{z_t}^i)_{z \in Z}$.

3.2.2.2 Beschreibung des Lernalgorithmus

Auf Basis der vorangehend festgelegten Begrifflichkeiten können nun die Aspekte *besten Markovantwort*, des *Markov Gegnermodells* und des übergreifenden *Lernalgorithmus* festgelegt werden.

Beste Markovantwort-Strategie Die Suche nach einer besten Markovantwort-Strategie gliedert sich in ein zweistufiges Problem; erstens der Frage nach der Ordnung der besten Antwort und zweitens in die Frage nach der Parametrisierung der besten Antwort.

Zur Bestimmung der *Antwort-Ordnung* einer besten Markovantwort-Strategie kann sowohl eine obere, als auch eine untere Schranke gefunden werden. Für eine *obere Schranke* kann auf die Ausführungen von Papadimitriou und Tsitsiklis (1987) zur effizienten Lösung des beste Antwort Problems für jeden Markovprozess zurückgegriffen werden. Demnach ist die Ordnung O^{*i} einer besten Markovantwort-Strategie $M^{*i}_{O^{*i}}(M^{j}_{O^{j}}, U^{i})$ auf eine gegebenes Markovgegnermodell $M^{j}_{O^{j}}$, welches sich im Zustand z^{j} befindet, nach oben beschränkt durch $O^{*i} \leq O^{j}$, wenn die durchschnittliche Auszahlung als Nutzenfunktion für U^{i} verwendet wird. Grundsätzlich kann auch in der Menge der Markovstrategien mit höherer Ordnung als der des Gegnermodells O^{j} nach einer besten Markovantwort gesucht werden. Diese kann zwar individuelle Aktionsmuster erzeugen,[21] niemals jedoch einen höheren Nutzen realisieren als spezifische alternative Lösungen mit $O^{*i} \leq O^{j}$; aus Antworten höherer Ordnung kann also kein Sophistizitätsvorteil gezogen werden (Press & Dyson, 2012, vgl.[S. 10412-10413).[22] Eine Antwort-Strategie mit $O^{*i} > O^{j}$ wäre für das gegebene Gegnermodell überkomplex und liefert keinen Mehrwert. Für eine *untere Schranke* kann die Suche nach einer besten Markovantwort-Strategie auf Strategien mit $O^{*i} \geq O^{j}$ beschränkt werden, da Markovstrategien kleinerer Ordnung stets eine echte Teilmenge der Markovstrategien größerer Ordnung sind, sodass gilt $\{M^{i}_{O^{i}_{1}}\} \subset \{M^{i}_{O^{i}_{2}}\}$ falls $O^{i}_{2} > O^{i}_{2}$. Eine Suche in der Strategiemenge mit $O^{*i} \geq O^{j}$ beinhaltet folglich stets alle möglichen Antwort-Strategien aus der Strategiemenge $O^{*i} < O^{j}$ (siehe Beispiel in Gleichung A.2, S. 157), weshalb letztere aus Effizienzgründen nicht weiter betrachtet werden muss. Zusammenfassend lässt sich somit die Ordnung der besten Markovantwort-Strategie unter Berücksichtigung der angeführten Schranken als identisch zu der Ordnung des angenommenen Gegnermodells festlegen. Dies deckt sich mit den Erkenntnissen von Axelrod (1984) zur Relevanz von Klarheit bezüglich der eigenen Aktionslogik für den Gegner. Es gilt:

$$O^{*i} = O^{j} \tag{3.8}$$

In Hinblick auf die Parametrisierung der *Antwort-Strategie* kann die Nutzenfunktion aus Gleichung 3.3 im Falle bedingter stochastischer Strategien anhand des antizipierten Pfades $\hat{h}$ des Strategiepaars keine deterministische Aussage mehr treffen. Stattdessen muss zur Bestimmung der besten Antwort-Strategie im wiederholten Spiel als der Durchschnitt der *erwarteten* Auszahlungen für den Pfad des Strategiepaares bestimmt werden:

[21] Sei beispielsweise die gegnerische Strategie $s^{j} = \{M^{j}_{(0,1)} = (0,1)^{T},\ \sigma^{j} = 0\}$. Dann vermag die Antwort-Strategie $s^{i} = \{M^{i}_{(1,1)} = (1,0,0,1)^{T},\ \sigma^{i} = 0\}$ einen Spielverlauf zu erzeugen, der durch Antwort-Strategien der Ordnung $O^{i} = O^{j} = (0,1)$ nicht abzubilden ist.

[22] Siehe auch Carmel und Markovitch (1998, S. 314-315) für eine Beweisführung zur besten Antwort für DFAs.

$$U^i_{\mathbf{E},\infty}(s^1,s^2) = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbf{E}[u^i(s^1(\hat{h}_{(s^1,s^2)}(t)),s^2(\hat{h}_{(s^1,s^2)}(t)))] \tag{3.9}$$

Die beste Markovantwort $s^{*i}(s^j,U^i)$ auf die Gegnerstrategie s^j wird durch die Auswahlfunktion B^i als $B^i : (s^j,U^i_{\mathbf{E},\infty}) \to S^i$ bestimmt. Die beste Antwort muss unter Berücksichtigung der durchschnittlichen erwarteten Auszahlung die Bedingung $\forall s \in S^i : U^i_{\mathbf{E},\infty}(s^{*i}(s^j,U^i),s^j) \geq U_{\mathbf{E},\infty}(s,s^j)$ unter der vereinfachenden Zusatzbedingung $O^{*i} = O^j$ erfüllen. Die Suche nach der besten Markovantwort $s^{*i} = (M^{*i},\sigma^{*i})$ zergliedert sich in dessen Komponenten M^{*i} und σ^{*i}.

Markov Gegnermodell Die Schätzung des Gegnermodells erfolgt unter der Annahme, dass die generische Strategie hinreichend gut durch eine stationäre Markovstrategie beschrieben werden kann (siehe S. 34)[23]. Darauf aufbauend gestaltet sich eine zweistufige Lernaufgabe für den Algorithmus, zu deren Beantwortung lediglich die bis zur aktuellen Runde t beobachtbaren Spielhistorie $h(t)$ zur Verfügung steht:

1. **Struktur:** Der Lernalgorithmus muss die vermutete Gedächtnistiefe des Gegners $\hat{O}^j$ aus der Menge möglicher plausibler Ordnungen Ω^j bestimmen, welche den Zustandsraum $Z^j_{\hat{O}^j}$ der gegnerischen Interaktionslogik determiniert.
2. **Parametrisierung:** Der Lernalgorithmus muss die Initialisierungslogik $\hat{\sigma}^j$ sowie eine plausible Übergangsmatrix $\hat{M}^j_{\hat{O}^j}$ der vermuteten gegnerischen Markovstrategie $\hat{s}^j$ bestimmen, welche die Aktionswahrscheinlichkeiten der antizipierten gegnerischen Strategie $\hat{s}^j_{\hat{O}^j}$ determinieren.

Lernalgorithmus Das Gegnermodell eines Markov Lernalgorithmus L_{Markov} kann somit formal auf Basis einer Menge an Beobachtungspaaren $(z^j_{O^j,t},a^j_t)$ gebildet werden, wobei $z^j_{O^j,t}$ der Markovzustand des Gegners in jeder Runde t des wiederholten Spiels und a^j_t die durchgeführte gegnerische Aktion aus diesem Zustand heraus darstellen. Die verfügbare Stichprobe $D^j_{O^j}$ beschreibt die Kette der bis zum aktuellen Zeitpunkt realisierten Beobachtungspaare des aus Zustand und Aktion des Gegners. Die Stichprobe ist für einen Markovgegener neben der beobachteten Historie $h(t)$ von der dessen Zustände z^j bestimmenden Gedächtnistiefe O^j des Gegners abhängig:

$$D^j_{O^j}(h(t)) = \begin{cases} \emptyset & \text{falls } o^j_{max} \leq t \\ ((z^j_{O^j,t},a^j_k)|o^j_{max} < k \leq t) & \text{sonst} \end{cases} \tag{3.10}$$

[23] Es gelten die Annahmen zur Informationsverfügbarkeit, welche im Rahmen der Ausführungen zum Gegnermodell im Falle eines Lernalgorithmus für deterministische Agenten getroffen wurden (siehe Seite 36).

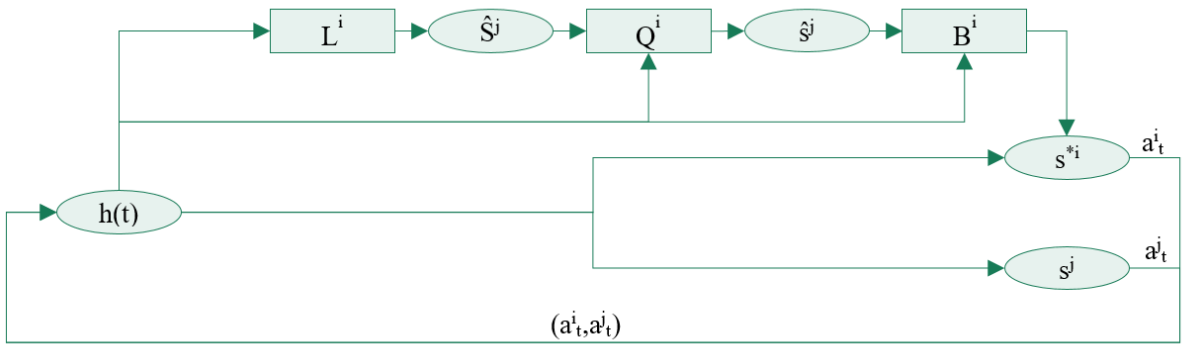


Abbildung 3.2: Interaktionslogik des Markovagenten. Quelle: Eigene Darstellung in Anlehnung an Carmel und Markovitch (1998, S. 312).

Auf Basis dieser Stichprobe bestimmt der Lernalgorithmus in jeder Runde t je möglicher Gedächtnistiefe $O^j \in \Omega^j$ eine Menge möglicher Gegnermodelle $\hat{S}^j_{\Omega^j} = \{\hat{s}^j_{\hat{O}^j} = L^i_{Markov}(D^j_{O^j}(h(t)))_{O^j \in \Omega^j}$, aus denen eine Selektionsfunktion Q^i_{Markov} glaubwürdigste Gegnermodell $\hat{s}^{*j}_{\hat{O}^j} = Q^i_{Markov}(\hat{S}^j_{\Omega^j})$ auswählt.

Die Antwort-Strategie s^i des lernenden Agenten hängt sowohl vom Gegnermodell $\hat{s}^{*j}_{\hat{O}^j}$, als auch von der Nutzenfunktion $U^i_{\mathbf{E},\infty}$ des lernenden Agenten ab. Es handelt sich unter Berücksichtigung der Überlegungen zu Gegnermodell und bester Antwort somit um eine Strategie gemäß:

$$\begin{aligned} s^i_{U^i,L^i}(h(t)) &= B^i(Q^i_{Markov}(L^i_{Markov}(D^j_{O^j}(h(t))), U^i_{\mathbf{E},\infty})(h(t)) \\ &= B^i(\hat{s}^{*j}, U^i_{\mathbf{E},\infty})(h(t)) \\ &\text{mit } O := O_{s^{*i}} = O_{\hat{s}^{*j}} \end{aligned} \tag{3.11}$$

Zusammenfassend ergibt sich auf Basis von Gleichung 3.11 ein sich unter Einbezug des Spielverlaufs strategisch adaptiv verhaltender Markovagent. Die grundlegende Funktionsweise eines solchen Spielers ist in Abbildung 3.2 aufgearbeitet. Der dargestellte Markovagent ist in dem Sinne subjektiv rational, dass er anhand subjektiver Überzeugungen über die gegnerische Strategie eine optimale Antwort-Strategie für sich selbst ermittelt (vgl. Kalai & Lehrer, 1993). Im folgenden Kapitel wird der beschriebene lernende Markovagent als Computerprogramm implementiert werden.

3.3 Umsetzung des Markovagenten als lernende Interaktionslogik

Der nachfolgende Teil dokumentiert die Umsetzung des Markovagenten. Ziel ist, das Konzept von Müller (2018) zur individuellen Vorhersage von Spielerverhalten mit Markovstrategien hin zu einer interaktionsfähigen und performanten Spiellogik zu erweitern. Das Kapitel gliedert

sich dabei in zwei Teile. Zunächst findet eine Eingrenzung der Betrachteten Markovstrategien unter Einbezug des Anwendungskontextes statt. Das anschließende Kapitel konzentriert sich dann auf die Ausgestaltung der konkreten Interaktionslogik. Diese besteht aus zwei Teilschritten. Erstens, der Antizipation des Gegnerverhaltens auf Basis eines anhand der Spielhistorie geschätzten Gegnermodells und zweitens, der darauf aufbauenden Ableitung einer leistungsstarken Antwort-Strategie. Der Markovagent wurde in der Programmiersprache Python umgesetzt. Als Grundgerüst wurden Elemente der Axelrod Softwarebibliothek verwendet (Vince Knight et al., 2020). Diese wurde tiefgreifend angepasst und erweitert, um den lernenden Markovagenten, die Interaktionsumgebung mit Menschen, den Experimentsetup und die Turniersituation mit Algorithmen abzubilden. Der iterative Entwicklungsprozess fußt neben konzeptionellen Überlegungen und maßgeblich auf den Erkenntnissen von funktionalen Pre-Tests im Rahmen empirischer Erhebungen (Versionierung siehe Tabelle 4.5).

3.3.1 Eingrenzung der Modellstruktur

Zunächst soll unter Berücksichtigung empirischer Ergebnisse und konzeptioneller Überlegungen mögliche Ordnungen Ω^i der Markovstrategien von AgentM festgelegt werden. Im Anschluss daran findet die Herleitung der daraus resultierenden Markovzustände Z_O^i statt.

3.3.1.1 Auswahl von passenden Markovordnungen

Im Folgenden soll die Gedächtnistiefe zur Modellierung des Verhaltens menschlicher Spieler durch den Markovagent erarbeitet werden. Ziel ist hierbei, Modelle zu bilden, anhand derer sich empirisches Gegnerverhalten als Grundlage für einen performanten MAL Agenten effektiv und effizient abbilden lässt. Wie in Kapitel 3.2.2 dargelegt, wird die Struktur des Markovzustandsraumes für ein gegebenes Spiel durch die verwendete Gedächtnistiefe bestimmt. Die Gedächtnistiefe legt fest, auf welcher Partition der Handlungshistorie Spielverhalten bedingt.

Hinsichtlich der *Effizienz der Vorhersage* ist eine möglichst kleine Gedächtnistiefe vorteilhaft. Grund ist, dass der Markovagent schon in frühen Runden auf Basis der Vorhersagen seines Gegnermodells mit dem anderen Spieler interagiert. Infolgedessen ist ein robuster und effizienter Schätzer erforderlich, der verhindert, dass Fehleinschätzungen den weiteren Spielverlauf durch die darauf folgende Aktionswahl des Agenten negativ beeinflussen. Letzteres ist aufgrund der Interdependenz von Strategien und der Pfadabhängigkeit von Spielverläufen von besonderer Bedeutung. Die Mächtigkeit des Markovzustandsraumes wächst für 2x2 Spiele exponentiell mit $|Z_O| = 2^{\sum_{o \in O} o}$. Folglich ist der Zustandsraum für ein Markovmodell mit $O = (1,1)$, welches auf das Ergebnis der letzten Runde bedingt von Mächtigkeit $|Z_{(1,1)}| = 4$ und für ein Markovmodell mit $O = (1,1)$, welches auf das Ergebnis der letzten beiden Runde bedingt bereits von Mächtigkeit $|Z_{(2,2)}| = 16$. Ein großer Zustandsraum ist aufgrund der hohen Anzahl der auf Basis

der Spielhistorie zu schätzenden Gegnermodellparameter (z.B. 16) unvorteilhaft (vgl. Powers & Shoham, 2005a, S. 818). Erstens sind eine tendenziell längere Spielhistorie erforderlich um für die einzelnen Parameter $m \in M_{Z_O}$ ein vergleichbar große und damit belastbare Stichprobe zu generieren. Zweitens nimmt die Gefahr fehlender Schätzwerte mit zunehmender Gedächtnistiefe aufgrund der Mächtigkeit des Markovzustandsraumes zu (vgl. Müller, 2018, S. 36-37). Dabei ist nicht gewährleistet, dass eine spezifische Spielhistorie überhaupt die Beobachtung mindestens einer Aktion je Markovzustand zulässt, ohne die einzelne Parameter der zu schätzenden Übergangsmatrix mit $m = \emptyset$ einen Fehlwert aufweisen würden. Sowohl unzureichend belastbare Schätzwerte, als auch Fehlwerte beeinträchtigen in Markovmodellen großer Ordnung Effizienz und Robustheit der Schätzer relativ zu schlankeren Modellen (vgl. analog für DFA in Carmel & Markovitch, 1998, S. 327).

Hinsichtlich der *Effektivität der Vorhersage* wird eine, das menschliche Gegnerverhalten möglichst gut beschreibende, Gedächtnistiefe angestrebt. Unter Berücksichtigung limitierender Kognitionskapazität menschlicher Agenten scheinen auch hier vergleichsweise kleine Modelle vorteilhaft (vgl. Simon, 1990). Die empirischen Ergebnisse der *Strategiemethode* zeigen, dass ein Großteil der Probanden lediglich auf die letzte Runde des Spiels bedingen, wenn sie gebeten werden, ihre Strategie als Programm zu artikulieren (vgl. Dal Bo & Frechette, 2013, 2018, 2019; Romero & Rosokha, 2018). Dementsprechend wurden keine Modelle mit $o \geq 2$ verwendet. Weiterhin empfiehlt Müller (2018, S. 148) auf Basis seiner ausführlichen und robusten experimentellen Ergebnisse zur individuellen Vorhersage menschlichen Spielverhaltens explizit die Verwendung einer Gedächtnistiefe $O = (1,1)$ für einen interaktiven Markovagenten. Insbesondere konnte mit Hilfe des einrundigen Markovgedächtnisses eine konsistent bessere Vorhersageleistung als mit Modellen größerer Ordnung erzielt werden (Müller, 2018, S.144-145).

Zusammenfassend scheint eine Beschränkung der bedingenden Partition der Spielhistorie sinnvoll, da andernfalls Overfitting durch eine Modellierung jenseits der tatsächlich genutzten kognitiven Kapazitäten menschlicher Spieler droht (vgl. Powers & Shoham, 2005a). Idealisierungen wie unendliche kognitive Kapazitäten und unbeschränkte gegenseitige Rekursion der Agentenmodelle stellen eine womöglich ineffektive Grundlage für anwendungsorientierte informatisch-spieltheoretische Modelle dar (vgl. Shoham, 2008). Empirische Ergebnisse zur individuellen Verhaltensvorhersage legen eine Gedächtnistiefe von $O = (1,1)$ nahe. Im Rahmen einer Gegnermodellierung im interaktiven Kontext kann es jedoch im Sinne der obigen Ausführungen zu Interdependenz und Pfadabhängigkeit vorteilhaft sein, anhand eines weniger detaillierten Modells die Effektivität der Vorhersage zugunsten einer effizienteren Modellbildung zu substituieren. Da diesbezüglich noch keine belastbare Aussage getroffen werden kann, werden für den weiteren Verlauf der Arbeit drei Varianten des adaptiven Markovagenten entwickelt und empirisch geprüft:

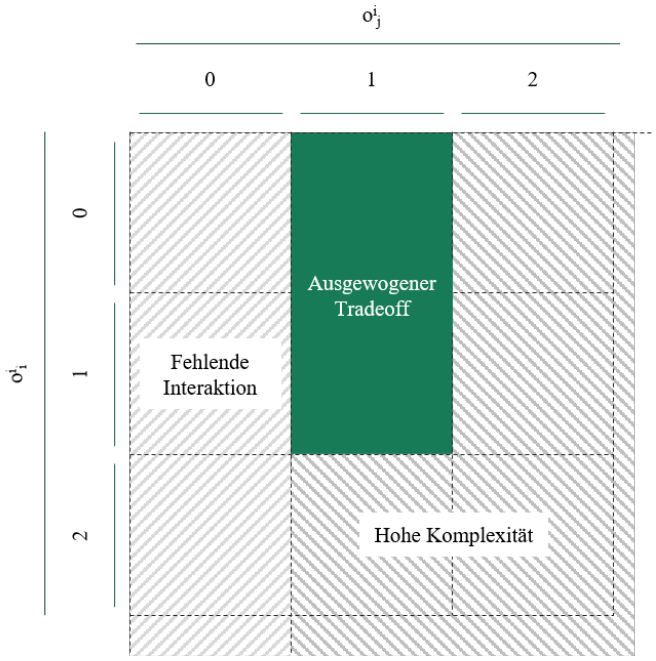


Abbildung 3.3: Parametrisierung des Markov Agenten (Spieler $i = 1$) anhand zielführender Gedächtnistiefen. Quelle: Eigene Darstellung.

1. **AgentM11:** Dieser Markovagent führt ein Gegnermodell, welches auf die Aktionen beider Spieler in der letzten Runde bedingt, sodass für dessen Gedächtnistiefe $O^j_{AgentM11} \in \Omega^j_{AgentM11} = \{(1,1)\}$ gilt.

2. **AgentM01:** Dieser Markovagent führt ein Gegnermodell, welches auf die Aktion des Gegenspielers in der letzten Runde bedingt, sodass für dessen Gedächtnistiefe $O^j_{AgentM01} \in \Omega^j_{AgentM01} = \{(0,1)\}$ gilt.

3. **AgentMx1:** Dieser Markovagent führt zwei Gegnermodelle; erstens eines, welches auf die Aktion des Gegenspielers in der letzten Runde bedingt und zweitens eines, welches auf die Aktionen beider Spieler in der letzten Runde bedingt, sodass für dessen Gedächtnistiefe in Summe $O^j_{AgentMx1} \in \Omega^j_{AgentMx1} = \{(0,1),(1,1)\}$ gilt. AgentMx1 wählt in jeder Runde auf Basis der Selektionsfunktion Q^i_{Markov} empirisch für den Spielverlauf glaubwürdigere Gegnermodell aus.

Die Auswahl wird wird in Abbildung 3.3 zusammengefasst.[24]

Tabelle 3.1: Zustandsräume für Markovagenten mit Ordnung $O^i \in \{(0,1),(1,1)\}$. Quelle: Eigene Darstellung.

Ordnung O^i	Markov-Zustände $z^i \in Z^i_{O^i}$
$(0,1)$	$(\emptyset, a^j_1)$
	$(\emptyset, a^j_2)$
$(1,1)$	(a^i_1, a^j_1)
	(a^i_1, a^j_2)
	(a^i_2, a^j_1)
	(a^i_2, a^j_2)

3.3.1.2 Bestimmung des Markovzustandsraumes

Auf Basis der definierten Gedächtnistiefen kann nun der Markovzustandsraum für die Markovagenten bestimmt werden. Tabelle 3.1 stellt diese dar. Im konkreten Anwendungsfall einer Markovkette erster Ordnung $O^i = (1,1)$ für 2x2 Spiele ergibt sich ein Zustandsraum von $Z^i_{(1,1)} = \{(a^i_1, a^j_1), (a^i_1, a^j_2), (a^i_2, a^j_1), (a^i_2, a^j_2)\}$ mit den Zustandselementen $z^i_{(1,1)} \in Z^i_{(1,1)}$. Jeder der $z^i_{(1,1)}$ Zustände korrespondiert dabei mit einem Zustand der 2x2 Spielmatrix, bestehend aus den Aktionspaaren der beiden Spiele. Der erste Buchstabe von $z^i_{(1,1)}$ entspricht hierbei der vorangegangenen Handlung des Spielers i, wohingegen der zweite Buchstabe die vorangegangene Handlung des Spielers j abbildet. Analog gilt für einen Spieler, der nur auf die gegnerische letzte Aktion bedingt mit $O^i = (0,1)$ der Zustandsraum $Z^i_{(0,1)} = \{(\emptyset, a^j_1), (\emptyset, a^j_2)\}$ mit den Zustandselementen $z^i_{(0,1)} \in Z^i_{(0,1)}$.

Mit Hilfe der Zustände können nun die Übergangsmatrizen für die ausgewählten Markovketten spezifiziert werden. Diese lautet analog zur vereinfachten Darstellung[25] von Gleichung 3.5 für $O^i = (0,1)$ und $O^i = (1,1)$:

$$M^i_{(0,1)} = (m^i_{z^i_{(0,1)}, a^i_2}) = \begin{pmatrix} m^i_{(\emptyset, a^j_1), a^i_2} \\ m^i_{(\emptyset, a^j_2), a^i_2} \end{pmatrix} = \begin{pmatrix} m^i_{(\emptyset, a^j_1)} \\ m^i_{(\emptyset, a^j_2)} \end{pmatrix} \tag{3.12}$$

[24] Modelle mit $o^i_j = 0$ wurden von der Betrachtung ausgeschlossen, da sie per Definition nicht in der Lage sind, gegnerisches Verhalten abzubilden. In derartigen Markovmodellen bedingen die zukünftigen Aktionen eines Spielers ausschließlich auf dessen eigenen vergangenen Aktionen.

[25] In der vereinfachten Darstellung wird mit lediglich einer Spalte die Wahrscheinlichkeit für das Durchführen von Aktion a^i_2 dargestellt. Die für 2x2 Spiele redundante Wahrscheinlichkeit für die a^i_1 ergibt sich aus $\mathbf{P}[a^i_1] = 1 - \mathbf{P}[a^i_2]$.

$$M^i_{(1,1)} = (m^i_{z^i_{(1,1)},a^i_2}) = \begin{pmatrix} m^i_{(a^i_1,a^j_1),a^i_2} \\ m^i_{(a^i_1,a^j_2),a^i_2} \\ m^i_{(a^i_2,a^j_1),a^i_2} \\ m^i_{(a^i_2,a^j_2),a^i_2} \end{pmatrix} = \begin{pmatrix} m^i_{(a^i_1,a^j_1)} \\ m^i_{(a^i_1,a^j_2)} \\ m^i_{(a^i_2,a^j_1)} \\ m^i_{(a^i_2,a^j_2)} \end{pmatrix} \tag{3.13}$$

Die Matrix M^i legt dabei auf Basis der aktuellen Markovzustände $z^i_{O,t} \in Z_O$ (Zeilen) die Wahrscheinlichkeiten für die nächsten Handlungen $a^i_t \in A^i$ (Spalten) von Spieler i in Runde t fest. Formal gilt $m^i_{z^i_O,a^i_2} = \mathbf{P}[a^i_2|z^i_O]$.

Tabelle 3.2: Darstellung der Pavlov-Strategie in der Normalformspielmatrix mit Werten für $\mathbf{P}[a^i_2|z^i]$ als Wahrscheinlichkeiten in jedem Zustand $z^i_{(1,1)}$ die Aktion a^i_2 zu spielen. Quelle: Eigene Darstellung.

	a^2_1	a^2_2
a^1_1	0%	100%
a^1_2	100%	0%

Zusammenfassend wurden unter Berücksichtigung technischer Limitationen sowie empirischer Forschungsergebnisse zur Strategiewahl menschlicher Spieler für den Markovagenten die Gedächtnistiefen $O^i = (0,1)$ und $O^i = (1,1)$ ausgewählt. Beispielhaft lässt sich auf Basis dieser Struktur die Pavlov-Strategie als $s^i_{Pavlov} = s^i_{(1,1)} = ((0\%, 100\%, 100\%, 0\%)^T, (0\%))$ Markovstrategie und Tit-for-Tat als $s^i_{(0,1)}$ Markovstrategie (siehe Kapitel 3.2.2.1) darstellen. Tabelle 3.2 veranschaulicht den Zusammenhang für die Pavlov-Strategie.

3.3.2 Interaktionslogik des Markovagenten

Das folgende Kapitel widmet sich der Entwicklung der Interaktionslogik des Markovagenten. Das Kapitel gliedert sich analog zum sich rundenweise wiederholenden Ablauf eines Markovagenten in Abbildung 3.2. Demnach gliedert sich das Kapitel anhand der folgenden Punkte:

- **Lernen der Gegnermodelle:** 3.3.2.1 beschäftigt sich zunächst mit der Schätzung möglicher Gegnermodelle durch den Lernalgorithmus L^i auf Basis der Spielhistorie.
- **Auswahl des Gegnermodell:** Kapitel 3.3.2.2 legt im Anschluss den Fokus auf die Auswahl eines möglichst vorhersagestarken Schätzmodells anhand der Gegnermodell-Selektionsfunktion Q^i.
- **Auswahl der Antwort:** Kapitel 3.3.2.3 erarbeitet abschließend die Markovantwort-Selektionsfunktion B^i durch AgentM für ein gegebenes Gegnermodell.

3.3.2.1 Aktualisierung der Gegnermodelle

Zentraler Bestandteil des Opponent Modelings im Rahmen einer markovbasierten Interaktion ist die Bestimmung der spielerindividuellen Markovstrategie $\hat{s}^j$ als Modell der tatsächlichen gegnerischen Strategie s^j im laufenden Spiel. Dieser Prozess, welcher durch die Lernalgorithmus L^i beschrieben wird, soll in den folgenden Kapiteln hergeleitet werden.

Bestimmung der Initialisierungslogik Die Schätzung der gegnerischen Initialisierungslogik ist zur Vorhersage des Gegnerverhaltens nicht notwendig, da jede Interaktion als neu aufgefasst wird und keine Transferlernleistung bezüglich eines spezifischen Spielers über mehrere Spiele hinweg stattfindet. Infolgedessen ist es für L^i ausreichend, die Übergangsmatrix M^j des Gegners zu lernen.[26]

Bestimmung der bedingten relativen Aktionswahrscheinlichkeiten Die iterative Schätzung der gegnerischen Übergangsmatrix $\hat{M}^j$ dient als Prädiktor des gegnerischen strategischen Spielverhaltens und richtet sich nach den Ausführungen zur markovbasierten Strategievorhersage von Müller (2018, S. 33-41). Die allgemeinen erarbeitete Formaldarstellung eines Markovagenten in Kapitel 3.2.2 werden als bekannt vorausgesetzt.

Die Schätzung der Übergangsmatrix gliedert sich trivialerweise in die Schätzung ihrer Matrixelemente $\hat{m}_z^j$ je Zustand. Dafür steht lediglich die rundenaktuelle Spielhistorie mit $h(t) = ((a_1^i, a_1^j), ..., (a_{t-1}^i, a_{t-1}^j))$ (siehe Gleichung 3.1) zur Verfügung. Aus ihr generiert AgentM für jede zulässige Gedächtnistiefe $O^j \in \Omega^j$ eine Reihe an Beobachtungen gemäß der Stichprobenfunktion $D^j(h(t), O^j)$ (siehe Gleichung 3.10). Diese erzeugt für jede Spielrunde $o_{max}^j < k \leq t$ ein Tupel $(z_{O^j}^j(k), a_k^j)$ bestehend aus dem jeweiligen Markovzustand *des Gegners* und dessen Aktionswahl aus dem Zustand heraus. Klar ist, dass für die ersten $k : 1 \leq k \leq o_{max}^j$ Runden mit $D_{O^j}^j(h(t)_{t \leq o_{max}^j}) = \emptyset$ keine Beobachtung erzeugt werden kann, da der Markovzustand aufgrund fehlender Historie für ein Modell der Ordnung O^j noch nicht gebildet werden kann (siehe Gleichung 3.6). Auf Basis der aktuellen Stichprobe berechnet der Lernalgorithmus von AgentM $\hat{M}_{O^j}^j = L_{Markov}^i(D_{O^j}^j)$. Für jeden Zustand $z_{O^j}^j$ wird dafür anhand einer Funktion $C_{(z,a_2^j)}(D^j)$ gezählt, wie oft der Gegner in der Stichprobe D^j aus Zustand z heraus Aktion a_2^j spielte. Ergänzend zählt Funktion $C_{(z,\emptyset)}(D^j)$, wie oft Zustand z unabhängig von der Aktionswahl in der Stichprobe vorkommt, sodass anhand $F_{z_{O^j}^j}(D^j)$ die empirische relative bedingte Häufigkeit des Gegners, Aktion a_2^j in Zustand $z_{O^j}^j$ zu spielen als vorläufiger Schätzer für $m_{z_{O^j}^j}^j$ ermittelt werden kann:

[26] Gleichwohl ist es der Vollständigkeit halber möglich die ersten o_{max}^j Handlungen des Gegners als dessen deterministische Initialisierungslogik zu erfassen, sodass $\hat{\sigma}_{O^j}^j = (a_{t=1...o_{max}^j}^j)$.

$$F_z(D^j) = \begin{cases} \emptyset & \text{falls } C_{(z,\emptyset)}(D^j) = 0 \\ \frac{C_{(z,a_2^j)}(D^j)}{C_{(z,\emptyset)}(,D^j)} & \text{sonst} \end{cases} \tag{3.14}$$

$$\begin{aligned} C_{(z,a_2^j)}(D^j) &= \sum_{\substack{(z_t^j,a_t^j)\in D^j: \\ z_t^j=z \wedge a_t^j=a_{t,2}^j}} 1 \\ C_{(z,\emptyset)}(D^j) &= \sum_{\substack{(z_t^j,a_t^j)\in D^j: \\ z_t^j=z}} 1 \end{aligned} \tag{3.15}$$

Die Verwendung der empirischen Häufigkeit der Aktionswahl des Gegners aus einem Zustand heraus integriert die auf einer Partition der Spielhistorie bedingenden Markovlogik mit experimentellen Ergebnissen, über die Verwendung gemischter Strategien durch Menschen (vgl. Müller, 2018; Romero & Rosokha, 2019). Nach dem schwachen Gesetz großer Zahlen kann die relative bedingte Häufigkeit $F_{z_{Oj}^j}$ als unvoreingenommener Schätzer für die wahre Wahrscheinlichkeit $\mathbf{P}[a_t^j = a_{t,2}^j | z_t^j = z]$ verwendet werden, da erstere der letzteren im Grenzfall asymptotisch annähert (vgl. Stewart, 2009, S. 183-184):

$$\lim_{t\to\infty}[|F_z(D^j) - \mathbf{P}[a_t^j = a_{t,2}^j | z_t^j = z]| > \varepsilon] = 0 \tag{3.16}$$

Die Berechnung der relativen Häufigkeit der gegnerischen Aktionswahl bedingend auf dem Zustand des Spielers erfolgt rundenweise iterativ. Diese kann ein robuster Prädiktor für zukünftiges Gegnerverhalten sein und ist eine geeignete Basis für einen interaktiven Markovagenten (vgl. Müller, 2018, S. 148).

Bestimmung der Übergangsmatrix unter Verwendung eines Priors Für Markovzustände z_{Oj}^j, in denen noch kein Gegnerverhalten beobachtet wurde, kann die relative bedingte Häufigkeit $F_{z_{Oj}^j}$ nicht berechnet werden. Infolgedessen kann es insbesondere in frühen Runden des Spiels mit $F_{z_{Oj}^j} = \emptyset$ zu Fehlwerten kommen. Die empirischen Ergebnisse von Müller (2018) legen zudem nahe, dass selbst im späteren Spielverlauf regelmäßig fehlende Werte enthalten sein können.[27] Eine vollständige Vorhersage über zukünftiges Gegnerverhalten auf Basis der relativen bedingte Häufigkeit allein ist demnach nur dann möglich, wenn mindestens eine Beobachtung je Zustand vorliegt (siehe Spalte *Kein Prior* in Tabelle 3.3). Die vollständige Vorhersage eines Pfades ist jedoch für die Bewertung möglicher Antwort-Strategien unbedingt notwendig (siehe Gleichung 3.9). Daher ist die Verwendung eines Priors für die Antizipation

[27] Die Anzahl der möglichen Fehlwerte steigt mit der Mächtigkeit des Markovzustandsraumes. Daraus folgt, dass eine technische Abwägung zwischen Detaillierungsgrad des Zustandsraumes (hohe Auflösung der Interaktionslogik) und Vorhersagekraft (geringe Beobachtungszahl je Zustand) stattfinden muss (siehe Kapitel 3.3.1).

Tabelle 3.3: Veranschaulichung der Aktualisierung verschiedener Gegnermodelle mit $O = (1,1)$ auf Basis unterschiedlicher Prior $\hat{M}_0^j$ und Priorgewichte γ_0 anhand eines exemplarischen Spielverlaufs. Quelle: Eigene Darstellung.

				Ohne Prior		Indifferenter Prior mit $\gamma_0 = 0$		Empirischer Prior mit $\gamma_0 = 1$	
t	a^i	a^j	$z^j_{O,t}$	$\hat{M}^j_{(1,1)}$	$\hat{m}^j_{z^j}$	$\hat{M}^j_{(1,1)}$	$\hat{m}^j_{z^j}$	$\hat{M}^j_{(1,1)}$	$\hat{m}^j_{z^j}$
1	a^i_2	a^j_2	–	$(-,-,-,-)$	–	$(0.5,0.5,0.5,0.5)$	–	$(0.12,0.72,0.63,0.88)$	–
2	a^i_2	a^j_1	$a^j_2 a^i_2$	$(-,-,-,-)$	–	$(0.5,0.5,0.5,0.5)$	0.5	$(0.12,0.72,0.63,0.88)$	0.88
3	a^i_2	a^j_2	$a^j_1 a^i_2$	$(-,-,-,0.0)$	–	$(0.5,0.5,0.5,0.0)$	0.5	$(0.12,0.72,0.63,0.44)$	0.72
4	a^i_1	a^j_2	$a^j_2 a^i_2$	$(-,1.0,-,0.0)$	0.0	$(0.5,1.0,0.5,0.0)$	0.0	$(0.12,0.86,0.63,0.44)$	0.44
5	a^i_2	a^j_1	$a^j_2 a^i_1$	$(-,1.0,-,0.5)$	–	$(0.5,1.0,0.5,0.5)$	0.5	$(0.12,0.86,0.63,0.63)$	0.63
6	a^i_2	a^j_1	$a^j_1 a^i_2$	$(-,1.0,0.0,0.5)$	1.0	$(0.5,1.0,0.0,0.5)$	1.0	$(0.12,0.86,0.32,0.63)$	0.86
7	a^i_2	a^j_2	$a^j_1 a^i_2$	$(-,0.5,0.0,0.5)$	0.5	$(0.5,0.5,0.0,0.5)$	0.5	$(0.12,0.57,0.32,0.63)$	0.57
8	a^i_1	a^j_2	$a^j_2 a^i_2$	$(-,0.67,0.0,0.5)$	0.5	$(0.5,0.67,0.0,0.5)$	0.5	$(0.12,0.68,0.32,0.63)$	0.63

möglicher Interaktionspfade erforderlich. Für die Integration eines Priors mit den empirischen relativen Häufigkeitswerten über beobachtete Gegneraktionen kann auf die Aktualisierungsregel von Fictitious Play zurückgegriffen werden (vgl. Fudenberg & Levine, 1995, S. 1069).[28] Der ex ante zu definierende Prior $\hat{M}_0^j = (\hat{m}_z^j)_{z \in O^j}$ für die gegnerischen Übergangswahrscheinlichkeiten wird dafür dem Gewichtungsfaktor γ_0 belegt, während die Anzahl $\gamma_F = C_{(z_{O^j}^j, \emptyset)}(D_{O^j}^j(h(t)))$ der bisherigen Beobachtungen von Zustand $z_{O^j}^j$ in der Stichprobe als Gewichtungsfaktor für die bedingten relativen Häufigkeiten herangezogen wird. Somit lässt sich die folgende Aktualisierungsregel für die Schätzer $\hat{m}^j$ von AgentM festlegen, welche die geschätzte Übergangsmatrix des Gegners $\hat{M}_{O^j}^j$ vollständig auf Basis der in der aktuellen Runde vorliegenden Stichprobe $D_{O^j}^j$ parametrisiert:

$$\begin{aligned} \hat{m}_{z_{Oj}^j}^j &= \frac{\gamma_0}{\gamma_0 + \gamma_F} \hat{m}_{z_{Oj},0}^j + \frac{\gamma_F}{\gamma_0 + \gamma_F} F_{z_{Oj}^j}(D_{O^j}^j) \\ \gamma_F &= C_{(z_{O^j}^j, \emptyset)}(D_{O^j}^j) \end{aligned} \tag{3.17}$$

Es schließt sich somit (1) die Frage nach einem passenden Priorwert $\hat{M}_0^j$ und (2) die Frage nach einem passenden Priorgewicht γ_0 an, welche nachfolgend diskutiert werden. Die Thematik wurde im Rahmen einer Prestudy unter Experimentalbedingungen für das effektive Prototyping des AgentM untersucht (siehe Kapitel B.2.3). Dabei legen die Ergebnisse einen *Priorwert* auf Basis empirischer Spielverlaufsdaten und ein *Priorgewicht* von $\gamma_0 = 1$ als mögliche effektive Parametrisierung nahe. Eine Ausführung zu den der Gestaltungsentscheidung zugrunde liegenden Überlegungen schließt sich an.

Hinsichtlich des *Priorwertes* $\hat{M}_0^j$ ist eine Möglichkeit zur Adressierung von fehlenden Werten die Verwendung des Indifferenzprinzips von LaPlace (1812). Dabei wird in Markovzuständen, für welche noch keine gegnerischen Zuginformationen beobachtet werden konnte, ein Prior von 50% verwertet (vgl. Müller, 2018, S. 40-41). Hierdurch kann ein vollständig definierter Schätzer für die Übergangsmatrix des Gegenspielers erstellt werden (siehe Spalte *Indifferenter Prior* in Tabelle 3.3). Die Problematik des indifferenten Priors wird besonders in Runde 2 deutlich, da hier stets von einem vollständig randomisierenden Gegner ausgegangen wird (siehe Zeile 2 in Spalte *Indifferenter Prior* in Tabelle 3.3). Mit einem solchen Spieler kann keine strategische Interaktion zustande kommen,[29] weshalb der Markovagent sich in seiner Antwort lediglich myopisch optimieren kann. Die Wahl der myopischen Antwort in Runde 2 kann dazu führen, dass der gesamte nachfolgende Spielverlauf beeinflusst wird. Insbesondere im wiederholten Pr-

[28] Der charakterisierende Unterschied von Fictitious Play ist, dass die Wahrscheinlichkeiten nicht auf einem Zustand des Spiels bedingen, sondern ohne Aktionskontext über das ganze Spiel erhoben werden.

[29] Ein vollständig randomisierender Gegner entspricht einer Strategie mit $O^j = (0,0)$, der von der bisherigen Spielhistorie völlig unabhängig ist. Insbesondere kann einem Spieler dieser Ordnung keine höhere Interaktionslogik aufinstruiert werden (vgl. 10412-10413 Press & Dyson, 2012).

isoner's Dilemma würde AgentM stets mit Abweichung antworten. Dies ist insbesondere vor dem Hintergrund der Pfadabhängigkeit von Spielverläufen problematisch.

In der Pfadabhängigkeit, beziehungsweise Interferenz von Vorhersage und realisiertem Spielverlauf liegt neben dem präskriptiven Untersuchungsgegenstand ein wesentlicher Unterschied zu der Arbeit von Müller (2018). Bei dem rein deskriptiven Ansatz von Müller (2018) ist die Verwendung eines indifferenten Priors nicht abträglich, da dieser keinen Einfluss auf den Spielverlauf hat. Im Kontext eines interagierenden Agenten ist der Schätzer jedoch bereits in der ersten Runde unmittelbar für den Gesamterfolg relevant. Grund ist, dass die Strategiewahl des Markovagenten nicht nur auf dem Schätzer des aktuellen Markovzustandes, sondern auf der *gesamten* geschätzten Markov-Übergangsmatrix beruht. Dementsprechend können Übergangsmatrizen mit indifferenten Prior-Wahrscheinlichkeiten, die je nach Spielverlauf bis in späte Runden bestehen können, das strategische Verhalten des Bots signifikant beeinflussen. Klar ist, dass infolgedessen auch das Verhalten des Gegners beeinflusst wird. Da tendenziell mehr Informationen gesammelt werden können, nimmt die beschriebene Problematik mit fortschreitendem Spiel ab. Gleichwohl verbleibt der Effekt bis zur Exploration jeden Markovzustandes. Im Beispiel in Tabelle 3.3 konnte etwa für $\hat{m}^i_{a_1^i a_1^j}$ bis zuletzt keine Beobachtung gewonnen werden, sodass für diesen Zustand noch immer von randomisierenden Gegnerverhalten ausgegangen wird. Der Schätzer konvergiert so nur langsam zu einem der tatsächlichen Strategie entsprechenden Modell.

Um stattdessen bereits zu Beginn des Spiels möglichst aussagekräftige Prior für das Gegnermodell heranziehen zu können, wird der Priorwert auf Basis empirischer Spielverlaufsdaten berechnet. Als Basis werden die umfassenden Daten menschlicher Interaktionsverläufe in diversen Spieltypen von Müller (2018) verwendet. Mit ihnen kann je ausgewähltem Spieltyp (siehe Kapitel 4.1.1) die durchschnittliche Wahrscheinlichkeit für eine Aktion a_2^j des Gegenspielers ermittelt werden. Diese sind in Tabelle 3.4 dargestellt. Für die Berechnung wurden die Zuginformationen aller Runden berücksichtigt, da Priorwerte wie im Beispiel in Tabelle 3.3 bis in späte Spielrunden bestand haben können. Mit Hilfe von Priorwerten, die auf empirischen menschlichen Interaktionen je Spieltyp basieren, wird eine belastbarere Vorhersage über erwartetes menschliches Verhalten trotz dem Vorhandensein von Fehlinformationen erwartet.

Tabelle 3.4: Übersicht über die verwendeten empirischen Priors in Abhängigkeit des Spieltyps nach Bruns (2015), D. Robinson und Goforth (2006) (siehe Kapitel 4.1.1) und der Markovordnung. Quelle: Eigene Darstellung.

	$\hat{M}_0^j$	
Spieltyp	$O^j = (0,1)$	$O^j = (1,1)$
Inferior	$(0.37, 0.83)$	$(0.12, 0.72, 0.63, 0.88)$
Unfair	$(0.20, 0.54)$	$(0.13, 0.50, 0.49, 0.59)$
Biased	$(0.73, 0.41)$	$(0.65, 0.34, 0.85, 0.53)$

Hinsichtlich des *Priorgewichts* γ_0 wurden vier Gestaltungsfälle unterschieden:

- **Sofortige vollständige Aktualisierung:** Bei einem $\gamma_0 = 0$ wird der Priorwert mit der ersten Beobachtung des zugehörigen Zustandes *vollständig* durch die neue Information ersetzt.
- **Überproportionale graduelle Aktualisierung:** Bei einem $0 < \gamma_0 < 1$ wird der Priorwert mit jeder zusätzlichen Beobachtung des zugehörigen Zustandes *überproportional* durch die neue Information ergänzt.
- **Proportionale graduelle Aktualisierung:** Bei einem $\gamma_0 = 1$ wird der Priorwert mit jeder zusätzlichen Beobachtung des zugehörigen Zustandes *proportional* durch die neue Information ergänzt.
- **Unterproportionale graduelle Aktualisierung:** Bei einem $\gamma_0 > 1$ wird der Priorwert mit jeder zusätzlichen Beobachtung des zugehörigen Zustandes *unterproportional* durch die neue Information ergänzt.

Eine unterproportionale graduelle Aktualisierung wurde ausgeschlossen, da hierdurch spielverlaufsspezifische und damit relevantere Informationen im Gegensatz zum unspezifischen Prior nicht in angemessenem Maße genutzt werden. Das sofortige und vollständige Priorupdate wurde im Rahmen der experimentellen Prestudy untersucht und verworfen. Grund ist, dass der Priorwert beim Erhalt der ersten Information für einen Zustand vollständig verworfen und je nach Gegneraktion augenblicklich durch einen 0% oder 100% Wert ersetzt wird, obwohl der Gegenspieler womöglich eine gemischte Strategie verfolgt (vgl. Müller, 2018; Romero & Rosokha, 2019). Infolgedessen besteht ein Risiko, dass vom Markovagent eine für den jeweiligen Zustand deterministische Entscheidungslogik des Gegners angenommen wird. Wie im Rahmen der Prestudy ersichtlich, kann dies regelmäßig zu übersteuerten Antwort-Strategien im frühen Spielverlauf führen, welche die nachfolgenden beeinträchtigen können. Stattdessen wird erwartet, dass die proportionale graduelle Aktualisierungsregel in Bezug auf empirische Ergebnisse

der Prestudy I sowohl weniger überreaktive Antwort-Strategien finden, als auch, nicht unabhängig davon, bessere Auszahlungsleistungen erzielen (siehe Kapitel C und B.2.3).[30] Zusammenfassend wird der Priorwert mit einem Priorgewicht von $\gamma_0 = 1$ also wie eine vollwertige Zuginformation behandelt, die mit allen beobachteten Aktionen des Gegners für den spezifischen Zustand gemittelt wird (siehe beispielsweise in Tabelle 3.3).

Beschränkung des Aktionsspeichers In der bis hierhin stattgefundenen Beschreibung besitzt AgentM ein unbegrenztes Gedächtnis über die Spielhistorie. Hieraus entsteht die Gefahr, dass der Markovspieler nicht in angemessener Geschwindigkeit auf einen Verhaltenswechsel im Sinne grundlegend anderer Übergangswahrscheinlichkeiten für einzelne Zustände durch den Gegner reagieren kann, da er veraltete Zuginformationen in der Erstellung des Gegnermodells übergewichtet (vgl Carmel & Markovitch, 1998, S. 331). Young (2004) bezeichnet die Eigenschaft von MAL Algorithmen mit unbeschränktem Gedächtnis, unter Umständen zu langsam auf neue Informationen zu reagieren, als *Trägheit* und legt die Verwendung eines beschränkten Gedächtnisses nahe. Im Falle des Markovagenten ist diese Überlegung insbesondere deshalb naheliegend, da die Annahme eines unbeschränkten Gedächtnisses für die Berechnung der Übergangswahrscheinlichkeiten aufgrund begrenzter kognitiver Kapazitäten (vgl. Simon, 1990) der zu modellierenden Menschen unangemessen sein könnte. Die Annahme näherungsweise stationärer Übergangswahrscheinlichkeiten legt ein unbegrenztes Gedächtnis nahe, doch bei unzureichender Spiellänge kann die empirische Herausforderung bestehen, dass das Modell nicht genug Informationen sammeln konnte um sich einzuschwingen. Beispielsweise kann ein Gegner zu Anfang eines Spiels versuchen den Markovagenten durch exploratives oder sogar exploitatives Verhalten zu testen. Erst im Anschluss an eine solche Kennenlernphase legt sich der Beispielgegner auf ein langfristiger ausgerichtetes Spielverhalten fest. Bei Berücksichtigung der gesamten Spielhistorie für die Erstellung des Gegnermodells würden die anfänglichen, aber nachfolgend nicht weiter relevanten Zuginformationen des Gegenspielers das Gegnermodell und somit das Interaktionsverhalten des Agenten verzerren.[31]

In diesem Spannungsfeld aus theoretischer Überlegung und nicht klar zu bestimmender pragmatischer Abwägungen wurde für den Markovagent ein begrenztes Gedächtnis eingeführt. Dieses soll als Kompromisslösung verhindern, dass etwaiges Artefaktwissen den Spielverlauf verzerrt, wird ein Aktionsspeicherlimit τ^j eingeführt, das festlegt, die wie viele der letzten Züge von

[30] Die unterproportionale graduelle Aktualisierung wurde aufgrund ihrer weniger intuitiven Parameterwahl nicht weiter untersucht, jedoch der Vollständigkeit halber mit aufgezählt.

[31] Der Sachverhalt stellt beispielhaft die Möglichkeit dar, dass die Strategie des Gegners die Mächtigkeit der modellierten Markovstrategie über den Gesamtspielverlauf betrachtet übersteigt, aber in Segmenten des Spielverlaufs zutreffend ist.

Gegner j je Zustand z_{O^j} für die Schätzung des Gegnermodells herangezogen werden.[32] Die Beschränkung gilt je Zustand und nicht global, da sonst Informationen aus einem Zustand verloren gehen können, der länger nicht mehr besucht wurde, ohne, dass es neue substituierende Informationen für ihn gäbe. Mit Hilfe von τ^j wird anhand von $D^j_{O^j}$ (siehe Gleichung 3.10) die beschränkte Stichprobe $D^j_{O^j,\tau}$ bestimmt, welche je Zustand z_{O^j} die maximal τ^j letzten Zuginformationen enthält:

$$\begin{aligned} D^j_{O^j,\tau}(h(t)) &= ((z_k,a_k) \in D^j_{O^j}(h(t)) : z_k = z \wedge t_{0,z} \leq t \forall z \in Z_{O^j}) \\ t_{0,z}(D^j_{O^j}(h(t))) &= min(k : |((z_k,a_k) \in D^j_{O^j}(h(t)) : z_k = z \wedge o^j_{max} < k \leq t)| \leq \tau) \end{aligned} \tag{3.18}$$

Das Aktionsspeicherlimit moderiert somit die Aktualität der Aktionsinformationen des Gegners zur Berechnung der einzelnen Wahrscheinlichkeitsschätzer je Zustand. Klar ist, dass der empirische Prior für jeden spezifischen Markovzustand, bei erreichen vom τ^j Beobachtungen für ein Modell mit $\gamma_0 = 1$ nicht mehr in Berechnung des Zustandsschätzers einfließt:

$$\begin{aligned} \hat{m}^j_{z^j_{O^j}} &= \begin{cases} \frac{\gamma_0}{\gamma_0+\gamma_F}\hat{m}^j_{z^{O^j},0} + \frac{\gamma_F}{\gamma_0+\gamma_F}F_{z^j_{O^j}}(D^j_{O^j,\tau}) & \text{falls } \gamma_F < \tau^j \\ F_{z^j_{O^j}}(D^j_{O^j,\tau}) & \text{sonst} \end{cases} \\ \gamma_F &= C_{(z^j_{O^j},\emptyset)}(D^j_{O^j,\tau}) \end{aligned} \tag{3.19}$$

Im Rahmen der experimentellen Prestudies im Prototypingprozess für AgentM wurden $\tau^j = \infty$, $\tau^j = 5$ und $\tau^j = 10$ untersucht (siehe Kapitel C und B.2.2). Ein unbeschränktes Gedächtnis mit $\tau^j = \infty$ zeigt dabei die eingangs beschriebene Trägheitsproblematik, was besonders bei entgegen der Annahme fehlender Stationarität der gegnerischen Strategie Probleme aufwirft. Ein zu kleines Gedächtnis wiederum kann wie im Fall von $\tau^j = 5$ dazu führen, dass das Gegnermodell zu sensibel auf Fehleingaben reagiert. Weiterhin wird damit die maximale Auflösung des Schätzers $\hat{m}^j_{z^j_{O^j}}$ auf eine Schrittweite $\frac{1}{\tau^j} = \frac{1}{5}$ limitiert, wodurch eine nur sehr grobe Vorhersage möglich ist. Als vielversprechendste Parametrisierung der Prestudies hat sich daher der Aktionsspeicher $\tau^j = 10$ herausgestellt, welcher eine ausreichend auflösende Schrittweite von $\frac{1}{10}$ zulässt, ohne dabei zu träge auf sich verändernde Verhaltensweisen zu reagieren. In Summe ergibt sich eine Gegnermodellierung, welche sich im Gegensatz vieler MAL Ansätze flexibel und effizient auf sich veränderndes Gegnerverhalten einstellen kann, was insbesondere im Spiel gegen Menschen von Bedeutung ist (vgl. Albrecht & Stone, 2018, S. 34).

[32] Ein Aktionsspeicherlimit je Zustand bewirkt, dass Daten für seltener erreichten Zustände nicht von Informationen aus häufiger besuchten Zuständen verdrängt werden können. Der Markovagent kann somit neben der *Aktualität* auch die *Relevanz* von Zuginformationen berücksichtigen.

3.3.2.2 Auswahl eines möglichst passenden Gegnermodells

Auf Basis der Erkenntnisse von Müller (2018) wurden in Kapitel 3.3.1 unter Berücksichtigung praktischer Überlegungen Markovmodelle mit den Gedächtnistiefen $\Omega^j_{AgentM01} = \{(0,1)\}$, $\Omega^j_{AgentM11} = \{(0,1),(1,1)\}$ und $\Omega^j_{AgentMx1} = \{(1,1)\}$ für AgentM definiert. Infolgedessen führt AgentMx1 zu jeder Zeit zwei mögliche Gegnermodelle; auf Basis einer Ordnung von $O^j = (0,1)$ sowie auf Basis einer Ordnung von $O^j = (1,1)$. Die Verwendung strukturell verschiedener Markovstrategien als Menge möglicher Gegnermodelle, welche durch Ω^j charakterisiert werden trägt Erfordernis, im Kontext lernender Agenten die Möglichkeit sich verändernden gegnerischen Verhaltens jenseits üblicher Stationaritätsannahmen zu berücksichtigen (vgl. Albrecht & Stone, 2018, S. 34). Die Auswahl des passenden Gegnermodells für AgentMx1 in jeder Runde ist Thema dieses Kapitels. Dafür wird die *Selektionsfunktion* Q^i definiert, die aus der Menge der möglichen Strategieschätzer $\hat{S}^j_{\Omega^j}$ auf Basis eines Qualitätsmaßes das vorhersagestärkste Gegnermodell $\hat{s}^j_{O^j \in \Omega^j}$ auswählt.

Berechnung der Vorhersagequalität Die Vorhersage der gegnerischen Aktion für den aktuellen Zustand $z^j_{O^j,t}$ ist mit $\hat{m}^j_{z^j_{O^j,t},t} = \mathbf{P}[a^j_{2,t}|z^j_{O^j,t}]$ ein Wahrscheinlichkeitsmaß, während die tatsächliche Aktionswahl des Gegners mit $a^j_t \in \{a^j_1, a^j_2\} \equiv \{0\%, 100\%\}$ binär ausfällt. Feltovich (2000) präsentiert für eine derartige Konstellation drei mögliche Kennzahlen für die Bewertung der Vorhersagegenauigkeit; den *Root Mean Squared Error (RMSE)*, den *Log-Likelihood Schätzer* $ln(L)$ und die *Proportion of Inaccuracy (POI)*.

Die *Log-Likelihood* Methode gibt die Summe der logarithmierten Vorhersagewahrscheinlichkeiten an und stellt somit ein Qualitätsmaß für die Vorhersagequalität dar, welches angibt, wie wahrscheinlich das Vorhersagemodell die Aktionshistorie reproduzieren könnte (vgl. Müller, 2018). Gemäß Feltovich (2000) gilt:

$$ln(L) = \sum_{k=1}^{t} [(1 - a^j_k) * ln(1 - \hat{m}^j_{z^j_{O^j,k},k}) + a^j_k * ln(\hat{m}^j_{z^j_{O^j,k},k})] \tag{3.20}$$

Die Log-Likelihood Methode eignet sich entgegen den Bedingungen dieses Anwendungskontextes, wenn mindestens eines der Modelle korrekt ist und das Ziel der Analyse die Auswahl des besten Modells ist (vgl. Feltovich, 2000). Stattdessen handelt es sich bei allen zu bewertenden Gegnermodellen lediglich um Näherungen. Weiterhin würde das Qualitätsmaß aufgrund der regelmäßigen Vorhersage von Nullwerten ($ln(1) = 0$) nicht definiert sein (vgl. Müller, 2018). Zusätzlich weist die Methode eine starke Sensitivität bezüglich geringer Wahrscheinlichkeitswerte auf, sodass eine Fehlvorhersage in einer Runde die Anpassungsgüte des Gesamtmodells unangemessen verzerren könnte (vgl. Müller, 2018; Selten, 1998). Zusammenfassend gestaltet sich die Log-Likelihood Methode insbesondere durch Suche nach einer guten Näherung statt

der Auswahl eines korrekten Modell als situativ ungeeignet (vgl. Erev & Roth, 1998; Feltovich, 2000).

Die *POI-Methode* trifft keine Aussage über das Ausmaß des Vorhersagefehlers; stattdessen findet lediglich eine direktionale Bewertung der Vorhersage statt. Konkret wird der Wahrscheinlichkeitsschätzer $\hat{m}^j$ zur deterministischen Klassifizierung der nächsten gegnerischen Aktion herangezogen und im Abgleich mit dem Realergebnis als richtig, falsch oder richtungslos bewertet (vgl. Feltovich, 2000):

$$POI = \frac{1}{t}\sum_{k=1}^{t}\begin{cases} 0, & \text{falls } |\hat{m}^j_{z^j_{O^j,k},k} - a^j_k < 0.5| \\ 0.5, & \text{falls } |\hat{m}^j_{z^j_{O^j,k},k} - a^j_k = 0.5| \\ 1, & \text{falls } |\hat{m}^j_{z^j_{O^j,k},k} - a^j_k > 0.5| \end{cases} \tag{3.21}$$

Die Bewertung des Fehlerausmaßes ist jedoch bei der Auswahl des besten Gegnermodells erwünscht, insbesondere da mit der POI-Schätzer nicht zwischen zwei Modellen unterschieden werden kann, die in die gleiche Güteklasse fallen.

Die *Root Mean Squared Error (RMSE)* hingegen entspricht der Wurzel des arithmetischen Mittels der quadrierten Abweichung zwischen vorhergesagter Wahrscheinlichkeit $\hat{m}^j$ und tatsächlicher Aktion a^j (vgl. Feltovich, 2000). Die Kennzahl bewertet demnach den Abstand zwischen Vorhersage und Realisierung. Durch das Quadrieren der Abweichung werden größere Abweichungen gegenüber kleineren Abweichungen überproportional gewichtet.

$$RMSE = \sqrt{\frac{1}{t}\sum_{k=1}^{t}(\hat{m}^j_{z^j_{O^j,k},k} - a^j_k)^2} \tag{3.22}$$

Analog zu Müller (2018) ist der RMSE aufgrund der Berücksichtigung des Ausmaßes der Abweichung und der positiven Eigenschaften des RMSE als Äquivalent der Quadratic Scoring Rule (vgl. Selten, 1998) der POI-Methode vorzuziehen. Auch Young (2007, S. 430) sehen ähnlich dazu in der Erfüllung von $\frac{1}{t}\sum_{k\leq t}|\hat{m}^j_{z^j_{O^j,k},k} - a^j_k|^2 \to 0$ für $t \to \infty$ die Bedingung für das Stattfinden eines konvergenten Lernprozesses im schwachen Sinne.

Gleichwohl gibt es Kritikpunkte am RMSE, welche aufgrund des Quadrierens sensitiv auf Ausreißer reagieren kann (vgl. Chai & Draxler, 2014). Weiterhin werden Interpretationsschwierigkeiten angeführt, die auf das Verletzen der Dreiecksungleichung zurückgehen, sodass Fehlerwerte nicht ohne Weiteres inhaltlich mit anderen quadrierten Fehlerwerten aus der Menge der quadrierten Fehler verglichen werden können. Verursachend liegt diesem Sachbestand die Tatsache zugrunde, dass quadrierte Fehler nicht nur von der mittleren absoluten Abweichung, dem

Mean Absolute Error (MAE), sondern auch von der Variabilität innerhalb der Menge der Fehler abhängt (vgl. Willmott et al., 2009):

$$MAE = \frac{1}{t}\sum_{k=1}^{t} |\hat{m}^j_{z^j_{O^j},k} - a^j_k| \tag{3.23}$$

Die MAE-Methode gibt somit die durchschnittliche Diskrepanz zwischen Vorhersage und tatsächlicher Realisierung an und beantwortet somit die Frage um wie viel Prozent der Vorhersage-Schätzer im Schnitt vom tatsächlichen Ergebnis abweicht. Sie ist demnach leicht zu berechnen und einfach zu interpretieren (vgl. Willmott et al., 2009), wenngleich der RMSE aufgrund der stärkeren Gewichtung von größeren Fehlern die Leistung verschiedener Modelle durch schnellere Konvergenz besser herausarbeitet (vgl. Chai & Draxler, 2014). Unter Berücksichtigung dieser Aspekte wird die Minimierung des MAE als Zielgröße der Selektionsfunktion Q^i von AgentMx1 verwendet. Der MAE bietet sich als anwendungsgerechte, intuitiv verständliche und vergleichbare Alternative zu den eingangs beschriebenen Methoden an. Der MAE weist dabei Parallelen zur Bedingung für konvergente Lernprozesse im engeren Sinne auf $|\hat{m}^j_{z^j_{O^j,t},t} - a^j_t| \to 0$ für $t \to \infty$ (vgl. Young, 2007, S. 430).

Auswahl anhand der Vorhersagequalität Zur Auswahl eines Gegnermodells berechnet AgentMx1 in jeder Runde für jedes der möglichen Gegnermodelle $O^j \in \Omega^j$ iterativ die mittlere absolute Abweichung *MAE* zwischen Vorhersage $\hat{M}^j_{O^j}$ Modelle und tatsächlicher Beobachtung als Leistungskennzahl für deren Vorhersagequalität. Für AgentM01 und AgentM11 (siehe Kapitel 3.3.1) ist aufgrund der Beschränkung auf eine einzige Markovordnung keine derartige Auswahl erforderlich. Auf dieser Basis wird für AgentMx1 in jeder Runde das aktuell glaubwürdigste Gegnermodell $\hat{s}^j_{O^j}$ anhand des geringsten Fehlers aus der Menge möglicher Modelle $\hat{S}^j_{\Omega^j}$ ausgewählt:

$$\begin{aligned} \hat{s}^j_{O^j} &= Q^i(\hat{S}^j_{\Omega^j}) \\ &= \arg\min_{\hat{s}^j_{O^j}} MAE[\hat{S}^j_{\Omega^j}] \end{aligned} \tag{3.24}$$

Für den Sonderfall, dass zwei $O^j \in \Omega^j$ gleich gut abschneiden, wählt AgentMx1 stets das Modell mit dem kleinsten Markovzustandsraum. Der Ansatz folgt der Heuristik *Ockhams Rasiermesser*, auch als *Prinzip der Parsimonie* bekannt, die stets die einfachste plausible Erklärung für einen gegebenen Sachverhalt heranzieht. Carmel und Markovitch (1998) verfolgen bei der Auswahl passender Gegnermodelle für ihren deterministischen Lernalgorithmus einen identischen Ansatz.[33] Unter der Annahme, dass die Gegnerstrategie als Markovstrategie modelliert

[33] Eine Alternative ist die gewichtete Auswahl der verschiedenen Gegnermodelle anhand einer zum Vorhersagefehler invers proportionalen Wahrscheinlichkeit (vgl. Suryadi & Gmytrasiewicz, 1999). Eine derartige Lösung

werden kann (vgl. Müller, 2018), ist es die Aufgabe von AgentM, das einfachste Gegnermodell zu finden, welches mit der Stichprobe des Gegnerverhaltens konsistent ist (vgl. Carmel & Markovitch, 1996).

Anhand dieser Informationen kann AgentM eine Aussage über die geschätzte gegnerische Strategie $\hat{s}^j_{O^j} = \{\hat{M}^j_{O^j}, \hat{\sigma}^j_{O^j}\}$ treffen. Mit Hilfe dieses ausgewählten Schätzers über das gegnerische Spielverhalten soll im nächsten Schritt eine Antwort-Strategie abgeleitet werden.

3.3.2.3 Auswahl einer möglichst passenden Antwort

Als letzten Schritt wählt der Markovagent anhand der geschätzten gegnerische Strategie eine möglichst passende Antwort-Strategie aus, auf Basis derer die eigene Aktion für die nächste Spielrunde bestimmt werden kann. Die Selektionsfunktion $B^i(\hat{s}^{*j}, U^i_{\mathbf{E},\infty})$ bestimmt die beste Antwort unter Berücksichtigung des ausgewählten Gegnermodells $\hat{s}^{*j}$ sowie der eigenen Nutzenfunktion $U^i_{\mathbf{E},\infty}$.

Berechnung des Nutzens möglicher Antworten Auf Basis einer Gegnerstrategie s^j können mögliche Antwort-Strategien hinsichtlich ihrer Qualifizierung als *beste* Antwort-Strategie anhand der Nutzenfunktion eines allgemeinen lernenden Markovagenten (siehe Gleichung 3.9, S. 44) bewertet werden. Die Nutzenfunktion im allgemeinen adaptiven Markovmodell berechnet dafür die durchschnittliche erwartete Auszahlung einer unendlichen wiederholten Interaktion zwischen der Gegnerstrategie und einer möglichen Antwort-Strategie. Der Prozess gliedert sich somit in die Schritte (1) Berechnung eines simulierten Interaktionsverlaufs und (2) Nutzenbewertung des simulierten Interaktionsverlaufs.

Die *Berechnung eines simulierten Interaktionsverlaufs* wird anhand des stochastischen Pfades $\hat{h}_{(s^i,s^j)}$ (siehe Kapitel 3.2.2.1) charakterisiert, welcher für jede Runde t die Wahrscheinlichkeiten für Aktion a_2 der beiden Spieler mit $(\mathbf{P}[a^i_{2,t}], \mathbf{P}[a^j_{2,t}])$ angibt. Eine alternative Darstellung der simulierten Interaktion zweier Markovagenten ist die Abbildung der Übergangswahrscheinlichkeiten von Zustand $z^i_t \in Z^i$ in der nächsten Runde auf Basis der Strategien beider Spieler in Zustand $z^i_{t+1} \in Z^i$ zu kommen (vgl. Press & Dyson, 2012). Diese Darstellung ist für die Interaktionssimulation effizienter, da sowohl der Interaktionspfad, als auch die Strategien mit dem Markovzustandsraum die identische strukturelle Basis besitzen. Das nachfolgende Vorgehen ist in diesem Sinne abweichend von den auf Carmel und Markovitch (1998) basierenden Ausführungen zur besten Antwort in Kapitel 3.2.

wurde aufgrund der Tendenz, einen häufigen und für den Gegner intransparenten Strategiewechsel hervorzurufen nicht umgesetzt. Insbesondere legen die Befragungen der Prestudy-Teilnehmer sowie die Ergebnisse von Axelrod (1984) nahe, dass transparentes und kausal nachvollziehbares Verhalten grundlegend für das Zustandekommen produktiver Interaktion ist.

Die Übergangswahrscheinlichkeiten eines Strategiepaares (s^i, s^j) werden in einer Übergangsmatrix $\Psi^i : z_t^i \to \mathbf{P}[z_{t+1}^i]$ zusammengefasst, welche die Zustände aus Sicht von Spieler i beschreibt. Die Übergangswahrscheinlichkeit ist dabei das Produkt der Wahrscheinlichkeiten aller Aktionen $a \in z_{t+1}^i$. Für einen $O^i = (1,1)$ Spieler ergibt sich somit:[34]

$$
\begin{aligned}
\Psi^i &= (\mathbf{P}[z_{t+1}^i | z_t^i]) \\
&= (\mathbf{P}[a_{t+1}^i \in z_{t+1}^i | z_t^i] \mathbf{P}[a^j \in z_{t+1}^i | z_t^i]) \\
&= \begin{pmatrix}
(1-m^i_{a_1^i a_1^j})(1-m^j_{a_1^j a_1^i}) & (1-m^i_{a_1^i a_1^j}) m^j_{a_{1,t}^j a_1^i} & m^i_{a_1^i a_1^j}(1-m^j_{a_1^j a_1^i}) & m^i_{a_1^i a_1^j} m^j_{a_1^j a_1^i} \\
(1-m^i_{a_1^i a_2^j})(1-m^j_{a_2^j a_1^i}) & (1-m^i_{a_1^i a_2^j}) m^j_{a_2^j a_1^i} & m^i_{a_1^i a_2^j}(1-m^j_{a_2^j a_1^i}) & m^i_{a_1^i a_2^j} m^j_{a_2^j a_1^i} \\
(1-m^i_{a_2^i a_1^j})(1-m^j_{a_1^j a_2^i}) & (1-m^i_{a_2^i a_1^j}) m^j_{a_1^j a_2^i} & m^i_{a_2^i a_1^j}(1-m^j_{a_1^j a_2^i}) & m^i_{a_2^i a_1^j} m^j_{a_1^j a_2^i} \\
(1-m^i_{a_2^i a_2^j})(1-m^j_{a_2^j a_2^i}) & (1-m^i_{a_2^i a_2^j}) m^j_{a_2^j a_2^i} & m^i_{a_2^i a_2^j}(1-m^j_{a_2^j a_2^i}) & m^i_{a_2^i a_2^j} m^j_{a_2^j a_2^i}
\end{pmatrix}
\end{aligned}
\tag{3.25}
$$

Sei $\theta^i(t)$ der Zustandsvektor für die aktuelle Runde t, der binär angibt, in welchem Zustand sich der Spieler i befindet. Dann können die Wahrscheinlichkeiten aller Zustände $\theta^i(t+1)$ in der nächsten Runde durch Multiplikation der $|Z^i| \times |Z^i| = 4 \times 4$-dimensionalen Übergangsmatrix Ψ^i mit dem $|Z^i| \times 1 = 4 \times 1$-dimensionalen Zustandsvektor $\theta^i(t)$ bestimmt werden:[35]

$$
\begin{aligned}
\theta^i(t)_{t>o_{max}} &= \begin{cases} 0 & \text{falls } z^i \neq z_t^i \\ 1 & \text{falls } z^i = z_t^i \end{cases} \\
\theta^i(k)_{k>t} &= (\mathbf{P}[z_k^i])_{z^i \in Z^i_{O^j}} \\
&= \Psi^i \theta^i(k-1)
\end{aligned}
\tag{3.26}
$$

Für eine über Δt Runden in die Zukunft simulierte wiederholte Interaktion kann mit Hilfe $\theta^i(k)$ mit den Wahrscheinlichkeiten der Zustände z^i für jede Runde $k \in \{t+1, ..., t+\Delta t\}$ bestimmt werden. Die iterative Berechnung ist erforderlich, da nicht von einer stationären Markovinteraktion im Sinne von $\theta^i(t+1) = \theta^i(t)$, ausgegangen werden kann (vgl. Press & Dyson, 2012). Den Spielverlauf zusammenfassend gibt $Theta^i_{Z^i,\Delta t}$ die kumulierten Wahrscheinlichkeiten über die simulierte Interaktion als Interaktionsergebnis an:

[34] Für Strategien der Ordnung $O^i = (0,1)$ ist es technisch erforderlich, diese strukturell auf eine äquivalente Strategie mit einer Ordnung von $(1,1)$ zu transformieren. Eine inhaltliche Veränderung der Strategien findet dabei nicht statt. Gleichung A.1 im Anhang A beschreibt den Prozess. Grund ist, dass das Interaktionsergebnis andernfalls für die sich anschließende Bewertung anhand der Nutzenfunktion unterdefiniert ist, da die Auszahlungen von den Aktionen beider Spieler abhängen, $O^i = (0,1)$-Strategien jedoch nur einen Zustandsraum mit den Aktionen des Gegenspielers beschreiben.

[35] Klar ist, dass die Beziehung erst ab Runde $t \geq o_{max}$ besteht, da zuvor noch kein Markovzustandsraum initialisiert wurde.

$$\Theta^i_{\Delta t}(t) = \sum_{k=t+1}^{t+\Delta t} \theta^i(k) \tag{3.27}$$

Im zweiten Schritt schließt sich die *Nutzenbewertung des simulierten Interaktionsverlaufs* für ein gegebenes Strategiepaar an. Da der Markovzustandsraum für $O = (1,1)$ den Zuständen der Auszahlungsmatrix eines 2x2 Spiels entspricht, kann der Auszahlungsvektor des Stufenspiels $R^i = (r^i(a^i,a^j))_{(a^i,a^j)\in A^i\times A^j}$ für Spieler i herangezogen werden. Analog zur Nutzenfunktion des allgemeinen lernenden Markovagenten (siehe Gleichung 3.9, S. 44) ergibt sich somit die durchschnittliche erwartete Auszahlung für die über Δt simulierte Interaktion des Strategiepaares (s^i, s^j) als:

$$U^i_{\mathbf{E},\Delta t}(s^i,s^j) = \frac{1}{\Delta t} R^i \Theta^i_{\Delta t}(t) \tag{3.28}$$

Anhand dieser Bewertung können alle potentiellen Antwort-Strategien $s^i \in S^i$ für eine gegebene Gegnerstrategie s^j hinsichtlich ihrer Auszahlungsleistung bewertet werden. Im nächsten Schritt findet eine Beschreibung über die Parametrisierung des Simulationsprozesses zur Bestimmung der besten Antwort-Strategie s^{*i} statt.

Parametrisierung der Simulation von Strategiepaarungen Mit Hilfe der Ausführungen des vorigen Ausschnittes kann für ein Gegnermodell M^j, welches sich in Zustand z^j befindet eine beste Antwortmatrix $M^i{*}i$ ermittelt werden. Diese erfolgt durch die numerische Interaktionssimulation zwischen dem Gegnermodell und allen in Betracht gezogenen Antwortmatrizen von AgentM sowie deren Nutzenbewertung. Für die Simulation möglicher Antwort-Strategien ist es erforderlich, (1) *Rundenzahl der Simulation* und (2) die die *Menge möglicher Antwortmatrizen* festzulegen.

Die *Rundenzahl der Simulation* Δt drückt die Weitsicht des lernenden Markovagenten aus. Je größer Δt, desto eher nimmt AgentM vorübergehende Payoffeinbußen in Kauf, um gegeben dem Gegnermodell einen langfristig für sich attraktiveren Spielverlauf herbeizuführen. Dabei sind zwei Extremfälle zu unterscheiden (Müller, 2018, S. 148):

- Für $\Delta t = 1$ handelt es sich um einen vollständig myopischen Agenten, der lediglich den Payoff der nächsten Runde maximiert.
- Für $\Delta t = \infty$ handelt es sich um einen vollständig weitsichtigen Agenten, der anfänglich entgangene Auszahlungen seiner Antwort-Strategie vollständig vernachlässigt.

Beide Extreme gestalten sich als für diesen Anwendungsfall ungeeignet. $\Delta t = 1$ verhindert, dass eine strategische Interaktion zwischen den Spielern zustande kommt und AgentM sein Potential entfalten kann. $\Delta t = \infty$ legt eine unangemessen hohe Konfidenz an das aktuelle geschätzte

Gegnermodell $\hat{M}^j$. Eine derart weitsichtige Optimierung macht dann Sinn, wenn sichergestellt werden kann, dass die simulierte Interaktion so stattfinden wird. Jedoch ist das verwendete Gegnermodell ein *Schätzer* des Gegnerverhaltens, welches unpräzise oder unvollständig sein kann. Infolgedessen ist es erforderlich auch kurzfristigere Payoffeffekte zu berücksichtigen. So können zwei Effekte besichtigt werden. Erstens, der direkte Auszahlungseffekt für die unmittelbar auszuführende eigene Handlung und zweitens, den Einfluss der eigenen Handlung auf das Verhalten der anderen adaptiven Agenten. Die Herausforderung dabei liegt insbesondere in der Abhängigkeit des Ergebnisses vom Verhalten des Gegners, dessen Strategie privat ist (vgl. Carmel & Markovitch, 1996).

Die Auswahl der simulierten Rundenzahl muss außerdem zwei weiteren, davon nicht unabhängigen Überlegungen Rechnung tragen; erstens der Periodizität von Markovinteraktionen und zweitens der Konvergenz der erwarteten Erreichbarkeiten. Tabelle 3.5a veranschaulicht mögliche Periodizitätseffekte. Anhand des Tabellenbeispiels ergibt sich ein vierperiodiges Interaktionsmuster. Die Spanne der Periode ergibt sich aus der Dimensionalität des Markovzustandsraumes. Konkret lässt die operationalisierte Markov-Interaktionslogik stets Perioden der Länge $|Z_O|$ zu - also eine Periodenlänge von 2 für $O = (0,1)$, beziehungsweise von 4 für $O = (1,1)$. Ein vorzeitiges Abschneiden der Periode hätte eine Ungleichgewichtung der Periodenelemente zur Folge[36]. Um derartige Ungleichgewichte zu vermeiden muss $\Delta t mod |Z| \stackrel{!}{=} 0$ gelten. Tabelle 3.5b veranschaulicht die mögliche Konvergenz der erwarteten Erreichbarkeiten bei der Paarung von mindestens einer gemischten Strategie, sodass im Beispiel bereits ab Runde 6 lediglich eine geringe Differenz zu den Werten in der letzten Runde besteht. Folglich lässt sich die Interaktion derartiger Markovstrategien in eine anfängliche konvergierende und eine darauffolgende in Bezug auf Erreichbarkeiten bereits konvergiert Phase gliedern. Der simulierte Spielverlauf sollte also nicht so kurz sein, dass die Anfangsphase übergewichtet wird, jedoch sollte der Spielverlauf auch nicht so lang sein, dass die Berechnung der Erreichbarkeiten unverhältnismäßigen Ressourcenaufwand erzeugt.[37]

Die vorangegangenen Überlegungen synthetisierend bezieht AgentM daher heuristisch die $\Delta t = 24$ nächsten Runden in die simulationsbasierte Suche nach einer effektiven interaktiven Antwort-Strategie mit ein.

Die *Menge möglicher Antwortmatrizen* bestimmt, ob die gewählte Antwort stets eine global beste Antwort-Strategie ist oder ob es sich gegebenenfalls um eine hinreichende Näherung an

[36] Im Beispiel der Tabelle 3.5a würden beim Abschneiden nach 23 Runden abweichend von den Tabellenwerten durchschnittliche erwartete Erreichbarkeiten von 21.74%, 26.09%, 26.09% und 26.09% resultieren.

[37] Alternativ könnte die Durchschnittsbildung für eine Betrachtung des rein eingeschwungenen Teils der Interaktion lediglich anhand der letzten $|Z|$ Runden des Spielverlaufs erfolgen. Jedoch wird die anfängliche Konvergenzphase bewusst zur Berücksichtigung etwaiger Koordinationskosten der ersten Runden einer Strategiepaarung miteinbezogen.

Tabelle 3.5: Exemplarische Erreichbarkeit und Auszahlung für ausgewählte Paarungen im wiederholten Prisoner's Dilemma. Quelle: Eigene Darstellung.

(a) Paarung $M^i_{(1,1)} = (1,0,0,1)^T$ und $M^j_{(1,1)} = (1,0,1,0)^T$ ausgehend von Zustand $z^i_{t=0} = a^i_1 a^j_1$.

	$\theta^i(t)$			
t	$a^i_1 a^j_1$	$a^i_1 a^j_2$	$a^i_2 a^j_1$	$a^i_2 a^j_2$
1	–	–	–	1
2	–	–	1	–
3	–	1	–	–
4	1	–	–	–
5	–	–	–	1
6	–	–	1	–
...	...	...	...	...
23	–	1	–	–
24	1	–	–	–
$\Theta^i_{\Delta t=24}$	6	6	6	6
$\mathbf{P}[z^i]$	25%	25%	25%	25%

(b) Paarung $M^i_{(1,1)} = (0,0,0.3,0.8)^T$ und $M^j_{(1,1)} = (0.2,0.6,0,1)^T$ ausgehend von Zustand $z^i_{t=0} = a^i_1 a^j_2$.

	$\theta^i(t)$			
t	$a^i_1 a^j_1$	$a^i_1 a^j_2$	$a^i_2 a^j_1$	$a^i_2 a^j_2$
1	0.40	0.60	–	–
2	0.56	0.44	–	–
3	0.62	0.35	–	–
4	0.65	0.35	–	–
5	0.66	0.34	–	–
6	0.66	0.34	–	–
...	...	...	...	...
23	0.67	0.33	–	–
24	0.67	0.33	–	–
$\Theta^i_{\Delta t=24}$	15.56	8.44	–	–
$\mathbf{P}[z^i]$	64.8%	35.2%	0.0%	0.00%

dessen Nutzenleistung handelt.[38] Eine Begrenzung der Strategiemenge ist daher naheliegend, insbesondere da dem lernenden Agent nur begrenzte Kapazitäten für die rundenweise anfallende Rechenaufgabe zur Verfügung stehen (vgl. Carmel & Markovitch, 1996, 1998). Ziel ist demnach, neben der Adressierung von Rechenleistungsgrenzen, möglichst nutzenstarke Strategien als Reaktion auf gegnerische Strategien vorzuhalten. Neben der Optimierung des rechnerisch erwarteten Payoffs spielt hier insbesondere die Transparenz der Antwort-Strategien eine Rolle. Strategische Transparenz ist für das menschliche Verstehen strategischer Ursache-Wirkungs-Beziehungen elementar, durch welche Interaktionen tendenziell koordinationseffizienter verlaufen können (vgl. Axelrod, 1984). Bleibt für menschliche Spieler ein Verständnis über die Logik der Strategie des anderen Spielers aus, wird diese als erratisch-zufällig empfunden, sodass beispielsweise Koordinationslösungen nicht implementiert werden können. Grundsätzlich sind Antwort-Strategien in gemischten Strategien denkbar. Die potentielle simulatorisch errechnete Mehrleistung wird jedoch als von den negativen Effekten einer intransparenten Strategiewahrnehmung Seitens des Gegners als überschattet angenommen.Um die Gefahr intransparenten Verhaltens zu reduzieren, beschränken sich die simulierten Antwort-Strategien auf Reinstrate-

[38] Kapitel 3.3.1 erarbeitete bereits, weshalb die beste Antwort-Strategie stets in der Menge der Markovstrategien zu finden ist, die der gleichen Ordnung, wie der des Gegners entspricht.

gien, also Strategien mit $m^{*i} \in \{0,1\}$. Daraus resultieren für $O^i = (0,1)$ insgesamt $2^{|Z_{(0,1)}|} = 4$ mögliche Antwortmatrizen, während $O^i = (1,1)$ insgesamt $2^{|Z_{(1,1)}|} = 16$ Antworten induziert. Es sei angemerkt, dass zwei identische Antwort-Übergangsmatrizen je nach Markovzustand der aktuellen Runde unterschiedliche Aktionen für die nächste Runde bedeuten können.

Bestimmung von Antwort-Strategien Anhand der festgelegten Simulationsparameter kann für jedes mit $\tau = 10$ beobachtbare Gegnermodell M^j und jeden Zustand z eine Antwortmatrix M^{*i} festgelegt werden. Im ersten Schritt wird dazu für das vorliegende Gegnermodell s^j und den aktuellen Zustand anhand von Gleichung 3.28 (siehe S. 64) der erwartete Nutzen $U^i_{\mathbf{E},\Delta t}$ für jede mögliche Antwort-Strategie ermittelt. Im zweiten Schritt werden nur jene möglichen Antworten weiter betrachtet, welche den erwarteten Nutzen von AgentM maximieren. Aus den verbleibenden Antworten wird schließlich die Antwort-Strategie selektiert, die gleichzeitig den erwarteten Nutzen des Gegners maximiert. So soll sichergestellt werden, dass der Gegenspieler einen Anreiz hat der Interaktion wie von der Simulation abgebildet weiterhin zu folgen. Somit ergibt sich die Auswahlfunktion B^i der besten Antwort-Strategien von AgentM als (vgl. Powers & Shoham, 2005a, 2005b):

$$\begin{aligned} S^{*i} &= B^i(s^j, U^i_{\mathbf{E},\Delta t}) \\ &= \underset{s^i_{max} \in S^i_{max}(s^j)}{\arg\max} [U^j_{\mathbf{E},\Delta t}(s^j, s^i_{max})] \\ \text{mit } S^i_{max}(s^j) &= \{x \in S^i : U^i_{\mathbf{E},\Delta t}(x, s^j) \geq \max_{y \in S^i}[U^i_{\mathbf{E},\Delta t}(y, s^j)]\} \end{aligned} \tag{3.29}$$

Bisweilen ist es möglich, dass B^i keine *eindeutige* Antwort-Strategie s^{*i} identifizieren kann, da mit $|S^{*i}| > 1$ für die beschriebenen Auswahlkriterien mehrere Lösungen in Frage kommen. Der Sachverhalt wird in Tabelle 3.6 im wiederholten Prisoner's Dilemma gegen die Grim Trigger-Strategie $s^j_{(1,1)} = \{(0,1,1,1)^T, (0)\}$ verdeutlicht. Tabelle 3.6a zeigt den Spielverlauf für die Antwort-Strategie Tit-for-Tat $s^{*i}_{(1,1)} = \{(0,1,0,1)^T, (0)\}$.Die exemplarische Strategiepaarung erzielt einen durchschnittlichen erwarteten Payoff von 3 für beide Spieler, da beide Spieler kontinuierlich kooperieren. Tabelle 3.6b hingegen zeigt den Spielverlauf für die Antwort-Strategie Always Cooperate $s^{*i}_{(1,1)} = \{(0,0,0,0)^T, (0)\}$. Auch diese exemplarische Strategiepaarung erzielt einen durchschnittlichen erwarteten Payoff von 3 für beide Spieler, da beide Spieler kontinuierlich kooperieren. Tatsächlich ist die erwartete Auszahlung gegen die Grim Trigger-Strategie für jede beliebige Antwort-Strategie $s^{*i}_{(1,1)} = \{(0, m^i_{(a^i_1,a^j_2)}, m^i_{(a^i_2,a^j_1)}, m^i_{(a^i_2,a^j_2)})^T, (0)\}$ identisch, wobei $m^i \in \{0,1\}$ definitionsgemäß einer Reinstrategie genügen muss. Folglich erzielen $2^3 = 8$ mögliche Antwort-Strategien den selben erwarteten Nutzen für beide Spieler. Eine Aussage, welche der im Beispiel 8 verschiedenen Antwort-Strategien zu wählen ist, kann auf Basis des erwarteten Nutzens nicht getroffen werden. Daher findet das Indifferenzprinzip Anwendung

Tabelle 3.6: Exemplarische Erreichbarkeit und Auszahlung für ausgewählte Paarungen im wiederholten Prisoner's Dilemma. Quelle: Eigene Darstellung.

(a) Paarung mit $M^{*i}_{(1,1)} = (0,1,0,1)^T$ und $M^{j}_{(1,1)} = (0,1,1,1)^T$ ausgehend von Zustand $z^i_{t=0} = a^i_1 a^j_1$.

	$\theta^i(t)$			
t	$a^i_1 a^j_1$	$a^i_1 a^j_2$	$a^i_2 a^j_1$	$a^i_2 a^j_2$
1	1	–	–	–
2	1	–	–	–
3	1	–	–	–
4	1	–	–	–
5	–	–	–	–
6	–	–	–	–
...	...	...	...	...
23	1	–	–	–
24	1	–	–	–
$\Theta^i_{\Delta t=24}$	24	–	–	–
$\mathbf{P}[z^i]$	100%	0%	0%	0%
R^i	3	0	5	1

(b) Paarung $M^{*i}_{(1,1)} = (0,0,0,0)^T$ und $M^{j}_{(1,1)} = (0,1,1,1)^T$ ausgehend von Zustand $z^i_{t=0} = a^i_1 a^j_1$.

	$\theta^i(t)$			
t	$a^i_1 a^j_1$	$a^i_1 a^j_2$	$a^i_2 a^j_1$	$a^i_2 a^j_2$
1	1	–	–	–
2	1	–	–	–
3	1	–	–	–
4	1	–	–	–
5	–	–	–	–
6	–	–	–	–
...	...	...	...	...
23	1	–	–	–
24	1	–	–	–
$\Theta^i_{\Delta t=24}$	24	–	–	–
$\mathbf{P}[z^i]$	100%	0%	0%	0%
R^i	3	0	5	1

(vgl. LaPlace, 1812), sodass die Übergangswahrscheinlichkeiten aller ausgewählten Antwort-Strategien S^{*i} gleichgewichtet aggregiert werden:

$$\begin{aligned} s^{*i} &= (M^{*i}, \sigma^{*i}) = ((m^{*i}_z)_{z \in Z_O}, \sigma^{*i}) \\ m^{*i}_z(S^{*i}) &= \frac{1}{|S^{*i}|} \sum_{s \in S^{*i}} m_z(s) \end{aligned} \tag{3.30}$$

Im oben genannten Beispiel ergibt sich somit für die gegnerische Grim Trigger Strategie $s^{*i} = \{(0, 0.5, 0.5, .05), (0)\}$ als Antwort-Strategie des Markovagenten. Es ist anzumerken, dass sich der erwartete Nutzen beider Spieler durch die Linearkombination der Antwort-Strategien S^{*i} nicht ändert (vgl. Press & Dyson, 2012). Gleichwohl wird durch die Anwendung des Indifferenzprinzips deutlich gemacht, dass solche Übergangswahrscheinlichkeiten der Antwort-Strategie für den erwarteten Spielverlauf keine Rolle spielen, für welche $m^i \notin \{0,1\}$ gilt ausschließlich, wenn der Gegner sich entsprechend seinem vorhergesagtem Gegnermodell verhält.

Einmalige Erstellung einer Datenbank bester Antwort-Strategien Aus Gründen der Laufzeitoptimierung und Sicherung der Stabilität des späteren Experimentablaufs wird die Interaktionssimulation nicht live in jeder Runde eines Spiels auf Basis des aktuellen Gegnermodells durchgeführt. Stattdessen wird *einmalig* im Vorhinein eine Strategiedatenbank erzeugt, die für jedes beobachtbare Gegnermodell und jeden Markovzustand bereits die Antwort-Strategie vorweist. Im Spielverlauf muss diese lediglich vom Markovagent abgerufen werden.[39] Die Menge I der im Vorhinein zu simulierenden Interaktionen lässt sich in Abhängigkeit der Mächtigkeit des Markovzustandsraumes und der Menge möglicher Gegner- und Antwort-Strategien berechnen:

$$\begin{aligned} |I| &= |\{s^i\}||\{s^j\}|2^{o^i_{max}+o^j_{max}} \\ |\{M^i\}| &= 2^{|Z|} \text{ für } m^i \in \{0,1\} \\ |\{M^j\}| &= 33^{|Z|} \text{ für } \tau = 10 \end{aligned} \tag{3.31}$$

Da als Antwort-Strategien nur Reinstrategien zugelassen wurden gilt $m^i \in \{0,1\}$. Bezüglich der Gegnermatrix kann mit einem Aktionslimit von $\tau = 10$ gilt hingegen:

$$m^j_\tau \in \{\frac{a}{b}\} \text{ mit } 0 \leq a \in \mathbb{N} \leq b \in \mathbb{N}^* \leq \tau \tag{3.32}$$

Für ein beispielhaftes $\tau = 4$ sind demnach redundanzfrei die Werte $\{0, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, 1\}$ für m^j möglich. Für das gewählte $\tau = 10$ von AgentM sind analog zu Gleichung 3.32 Werte zu ergänzen. In Summe ergeben sich je Markovzustand so 33 eindeutige Werte für m^j. Weiterhin muss jede Interaktion für jede der $2^{o^i_{max}+o^j_{max}}$ Zustandskombinationen der aktuellen Runde simuliert werden. Im Summe ergeben sich so für $O = (0,1)$ eine Anzahl von $17,424$ zu simulierenden Interaktionen, während $O = (1,1)$ erfordert, $75,898,944$ Interaktionen zu simulieren.

Auf Basis der Gesamtheit der gesampelten Interaktionen wird für jedes mit $\tau = 10$ beobachtbare Modell M^j und jeden Zustand eine beste Antwortmatrix M^{*i} in der Strategiedatenbank abgespeichert. Die Datenbank umfasst somit anlog zu Gleichung 3.31 in Summe Antworten auf $|\{s^j\}|2^{o^i_{max}+o^j_{max}}$ Konstellationen. Dies entspricht für $O = (0,1)$ einer Menge von $4,356$ Einträgen und für $O = (1,1)$ einer Menge von $4,743,684$ Einträgen, die aus der zuvor beschriebenen Menge an simulierten Interaktionen gewonnen werden.

Abruf der besten Markovantwort-Strategie aus der Datenbank In jeder Runde findet die Auswahl der aktuellen passenden Antwort-Strategie auf Basis der bisherigen Spielhistorie

[39] Tatsächlich kann es ausreichend sein, lediglich die erwarteten Erreichbarkeiten der Spielzustände anhand Θ^i (siehe Gleichung 3.27, S. 64) im Vorhinein zu berechnen. Die beste Antwort kann für ein gegebenes Spiel auch live jede Runde anhand des erwarteten Nutzens (siehe Gleichung 3.28) für das vorliegende Gegnermodell ermittelt werden.

statt. Dazu wählt AgentM anhand des aktuellen Schätzers $\hat{s}^j$ über die Gegnerstrategie (siehe Kapitel 3.3.2.1 und 3.3.2.2) eine passende Antwort-Strategie aus der simulierten Strategiedatenbank aus.[40] Dabei wird angenommen, dass der aktuelle Schätzer über die Gegnerstrategie der wahren und unveränderlichen Gegnerstrategie entspricht. Demnach erfolgt die Auswahl der Antwort-Strategie auf Basis einer fortgeführten Interaktion der geschätzten Gegnerstrategie und der ausgewählten Antwort-Strategie. Diese Annahme basiert dabei stets auf den aktuellen Informationsstand. Bei Ankunft neuer Zuginformationen wird das Gegnermodell in jeder Runde des Spiels aktualisiert. Infolgedessen kann die Auswahl einer neuen Antwort-Strategie erfolgen (vgl. Carmel & Markovitch, 1998). Zwar wird die Gegnerstrategie als fix angenommen, doch AgentM lernt diese auch dann über Zeit, wenn sich diese im Spielverlauf verändert. Die Auswahl der Antwort-Strategie für ein Gegnermodell $\hat{M}^j$ und den Zuständen z_t^i und z_t^j der beiden Spieler aus der Datenbank erfolgt nach einer stufenweisen Logik:

1. Zunächst werden in einer Vorfilterung alle Einträge ausgeschlossen, deren Ordnung O nicht der Ordnung des Gegnermodells $O(\hat{M}^j)$ entspricht.

2. Weiterhin werden alle Einträge ausgeschlossen, bei denen sich die beiden simulierten Strategien nicht in den aktuellen Markovzuständen der beiden Spieler $z^i = z_t^i$ und $z^j = z_t^j$ befinden. Dadurch ist gewährleistet, dass sich der simulierte Spielverlauf mit dem tatsächlich von der Spielhistorie induzierten Spielverlaufs deckt.

3. Nachfolgend werden aus den verbleibenden Paarungen jene ausgewählt, für deren gegnerische Übergangsmatrix die größtmögliche inhaltliche Übereinstimmung mit dem Schätzer für die gegnerische Übergangsmatrix besteht. Hierfür wird die euklidische Distanz zwischen geschätzter Übergangsmatrix $\hat{M}_Z^j$ und allen verbliebenen simulierten gegnerischen Übergangsmatrizen $\{M_Z^j\}$ berechnet:

$$d[\hat{M}_Z^j, M_Z^j] = \sqrt{\sum_{z \in Z_O} (\hat{m}_z^j -, m_z^j)^2} \tag{3.33}$$

   Es wird nun jene Strategiepaarung der Datenbank ausgewählt, für die der euklidische Abstand am geringsten ist. Da bei der Erstellung der Datenbank alle für $\tau = 10$ beobachtbaren Gegnermodelle für alle beobachtbaren Zustände $z \in Z_o \forall O \in \Omega$ simuliert wurden, ist dieses Ergebnis bei der gegebenen Parametrisierung stets eindeutig.

Ergebnis des Auswahlprozesses ist eine Markovantwort M^{*i}, welche die neue Interaktionslogik von AgentM darstellt. Die Antwort wird jede Runde auf Basis des aktuellen Gegnermodells

[40] Die Antwort-Strategie kann alternativ gemäß den Ausführungen dieses Kapitels 3.3.2.3 in Echtzeit ermittelt werden. In diesem Fall wurde aus Gründen der Prozessstabilität und Laufzeit eine vorgelagerte Berechnung umgesetzt.

neu festgelegt wird. Mit ihr kann anhand des gegenwärtigen Markovzustandes z_t^i die Aktion $a_t^i(z_t^i)$ des Markovagenten ermittelt werden, wenn dieser eine Markovstrategie wählt. Es gibt jedoch Situationen, in denen AgentM keine Markovstrategie spielt, worauf im nächsten Abschnitt eingegangen wird.

Auswahl einer Antwortaktion Es ist nicht immer zielführend im Sinne einer langfristigen Maximierung der Auszahlungsleistung, die beste Markovantwort auf ein Gegnermodell zu spielen. Im Rahmen der Prestudy zur Betaversion des Markovagenten im Prisoner's Dilemma konnte anhand der generierten Spielverläufe bestätigt werden, dass das sofortige Spielen einer besten Markovantwort regelmäßig darin resultiert, dass die beiden Spieler im häufig für beide wenig attraktiven Nash-Gleichgewicht des Stufenspiels, der beidseitigen Abweichung, festsitzen. Selbst, wenn der Gegenspieler versucht, eine Kooperation anzubieten, wird diese dann durch AgentM nicht angenommen, da dieser aufgrund seines Gegnermodells davon überzeugt ist, dass im aktuellen Zustand das Spielen einer Abweichung auch langfristig sinnvoll ist. Ergebnis ist ein destruktiver Zirkelschluss, wenn der Gegner tatsächlich und abweichend vom aktuellen Gegnermodell kooperationsbereit ist. Durch das konsequente Spielen einer Abweichung kommt AgentM nicht mehr in jene Markovzustände, in denen er unangemessene Informationen über den Gegner führt. Infolgedessen ist das Spiel, womöglich ungerechtfertigter Weise, in einer Abweichungslösung festgefahren, was für beide Spieler suboptimal ist. Als Haupttreiber dafür konnte die Maximierung des eigenen erwarteten Payoff durch AgentM auf Basis eines noch nicht ausreichend belastbaren Gegnermodells identifiziert werden. Insbesondere im Spiel gegen Grim Trigger ähnlichen Strategien kann dadurch eine nachteilige Eskalationsspirale ausgelöst werden, die im Vergleich zu einem alternativen kooperativeren Spielverlauf suboptimal für beide Spieler ist. Die Herausforderung bei Grim Trigger Gegners liegt in der Irreversibilität der negativen Konsequenzen des Aufdeckens ihrer Handlungslogik (vgl. Powers & Shoham, 2005b). Eine verwandte Problematik zeichnet sich auch im Spiel gegen Tit-for-Tat ähnliche Strategien ab. Kommt AgentM in frühen Runden des Spiels z.B. wiederholten Prisoner's Dilemmas auf Basis des aktuellen Gegnermodells zu dem Schluss, dass eine Abweichung sinnvoll ist, löst das bei dem Tit-for-Tat-Gegner spätestens im nächsten Zug ebenfalls eine Abweichung aus. Infolgedessen erhärtet sich der Verdacht von AgentM, gegen einen unkooperativen Gegner zu spielen zunehmend. Die Konsequenz ist eine für beide Spieler suboptimale Abweichungsspirale. Die Relevanz der beschriebenen Dynamik wird durch empirische Ergebnisse, die mit Grim Trigger und Tit-for-Tat verwandte Strategiecluster als häufige von Menschen gewählte Aktionslogik identifizieren (Dal Bo & Frechette, 2018, S. 83; Müller, 2018, S. 97-98). Auch für weitere Gegnerstrategien zeigt Axelrod (1984), dass das Ergebnis in wiederholten Spielen durch Kooperationsbereitschaft bei gleichzeitiger Wehrhaftigkeit gegen Ausbeutung zielführend ist.

Die kooperative Ausrichtung der Aktionslogik orientiert sich maßgeblich an den Erkenntnissen der Strategieturniere von Axelrod (1984). Als Kooperativlösung im Sinne von AgentM jene reine paretoeffiziente Strategie verstanden werden, welche die Summe der Auszahlungen der Spieler maximiert. Ziel ist, bei fehlender Evidenz über nichtkooperatives Gegnerverhalten eine für beide Parteien nachteilige Abweichungsspirale zu vermeiden. Fehlende Evidenz subsummiert sowohl das Fehlen von Informationen über das Gegnerverhalten, als auch das Vorhandensein einer rein kooperativen Spielhistorie. In beiden Fällen spielt der Markovagent selbst möglichst kooperativ, um eine Abwärtsspirale durch initiatives Abweichen zu vermeiden. Zwar gibt es Spielerpopulationen, in denen präventiv und initiativ ausbeutendes Verhalten auch langfristig dominant ist, doch liegen dem Agenten in diesem Anwendungsfall keine Informationen oder Pre-Tests diesbezüglich vor. Einzige Informationsquelle ist der aktuelle Spielverlauf. Folgende Fälle werden von AgentM unterschieden:

1. Wenn genau ein Zustand des Stufenspiels kooperativ ist, spielt AgentM diesen so lange, bis der Gegner das erste Mal von der Kooperativlösung abweicht.
2. Wenn mehrere Zustände des Stufenspiels kooperativ sind, spielt AgentM in der ersten Runde den vom Gegenspieler präferierten Kooperationszustand und wechselt dann zum Markov-Modul.
3. Falls kein kooperativer Zustand existiert, spielt AgentM in der ersten Runde, falls vorhanden, das präferierte Nash-Gleichgewicht des Stufenspiels und wechselt dann zum Markov-Modul.
4. Trifft keiner der vorangegangenen Fälle zu, randomisiert AgentM in der ersten Runde und wechselt dann zum Markov-Modul.

Die Logik des ersten Falles wird als *Markov-for-Tat* Logik bezeichnet und ist von den positiven Eigenschaften der Tit-for-Tat-Strategie inspiriert, da AgentM so lange kooperativ spielt, bis der Gegenspieler proaktiv-einseitig davon abweicht. Erst danach greift die zuvor beschriebene Aktionslogik nach der besten Markovantwort auf Basis einer Maximierung des eigenen erwarteten Nutzens. Die zweistufige Logik gestaltet sich wie folgt:

1. Spiele kooperativ, wenn eine der folgenden Punkte zutrifft:
    a) Es handelt sich um die erste Runde des Spiels. Eine Entscheidung auf Basis des Gegnerverhaltens ist noch nicht möglich, da der Markovzustandsraum noch nicht initialisiert werden konnte.
    b) Der Gegner hat in den vorigen Runden durchgehend ebenfalls kooperativ gespielt.

2. Andernfalls wähle die Aktion für den aktuellen Zustand auf Basis der besten Antwortübergangsmatrix für das aktuelle Gegnermodell:

$$a_t^i(z_t^i) = \begin{cases} \mathbf{P}[a_{1,t}^i | z_t^i] = 1 - m_{z_t^i}^{*i} \\ \mathbf{P}[a_{2,t}^i | z_t^i] = m_{z_t^i}^{*i} \end{cases} \tag{3.34}$$

Hervorzuheben ist, dass AgentM die geführten Gegnermodelle über den gesamten Spielverlauf hinweg aktualisiert, also auch wenn die Anwendung einer simulierten Antwort-Strategie aufgrund einer perfekten Kooperationsserie verzichtet wird.

3.3.2.4 Zusammenfassung

Die generalisiert unter Kapitel 3.2.1.2 (siehe S. 36) beschriebene Schrittfolge für modellbasierte lernende Agenten wurde als AgentM erfolgreich operationalisiert. Die Funktionsweise des Markovagenten lässt sich in zwei Phasen gliedern; (1) vor Durchführung von Spielen und (2) während der Durchführung von Spielen.

Vor einer Interaktion in einem Spiel kann zur Laufzeitoptimierung eine vollständige Datenbank aller Antwort-Strategien für mögliche Gegnermodelle auf Basis des erwarteten Nutzens erzeugt werden (siehe Kapitel 3.3.2.3, S. 62). Während einer Interaktion wird die Aktionswahl des Markovagenten rundenweise auf Basis kontinuierlich aktualisierter Gegnermodelle und der Strategiedatenbank in einem Dreitakt durchgeführt:

1. **Aktualisierung der Gegnermodelle**: Der Markovagent kann mittels der Lernfunktion L^i multiple Gegnermodelle parallel schätzen. Diese unterscheiden sich in Bezug auf die zugrunde gelegte Gedächtnistiefe $O \in \Omega$ (siehe Kapitel 3.3.1, S. 46). Alle Gegnermodelle werden kontinuierlich auf Basis der verfügbaren Zuginformationen aktualisiert (siehe Kapitel 3.3.2.1, S. 51).

2. **Auswahl eines Gegnermodells**: In jeder Runde wird das Gegnermodell anhand der Selektionsfunktion Q^i ausgewählt, welches in Hinblick auf das gewählte Anpassungsgütemaß das tatsächliche Gegnerverhalten am besten beschreibt (siehe Kapitel 3.3.2.2, S. 59).

3. **Auswahl einer Antwort**: Auf Basis des Gegnermodells wird schließlich mittels der Antwortfunktion B^i eine passende Markovantwort ermittelt. Die ausgewählte Antwort-Strategie und der aktuelle Markovzustand legen die Aktion des Markovagenten fest, wobei kooperative Aspekte gesondert berücksichtigt werden (siehe Kapitel 3.3.2.3, S. 62).

Tabelle 3.7: Exemplarischer Spielverlauf von AgentM01 mit $O = (0,1)$ im wiederholten Chicken Game unter Verwendung der empirischen Werte für den Prior $\hat{M}_0^j$ nach Tabelle 3.4 mit einer graduellen Aktualisierungsregel $\gamma_0 = 1$. Quelle: Eigene Darstellung.

t	$\hat{M}^j_{(0,1)}$	z^i	$M^{*i}_{(0,1)}$	a^i	a^j	r^i	r^j	$\bar{\rho}^i$	$\bar{\rho}^j$
1	$(0.20, 0.54)$	–	Always Cooperate	a_1^i	a_2^j	1	5	20.0	100.0
2	$(0.20, 0.54)$	a_2^j	$(0.0, 1.0)$	a_2^i	a_1^j	5	1	60.0	60.0
3	$(0.10, 0.54)$	a_1^j	$(0.0, 1.0)$	a_1^i	a_1^j	3	3	60.0	60.0
4	$(0.10, 0.27)$	a_1^j	$(1.0, 1.0)$	a_2^i	a_1^j	5	1	70.0	50.0
5	$(0.07, 0.27)$	a_1^j	$(1.0, 1.0)$	a_2^i	a_2^j	0	0	56.0	40.0
6	$(0.07, 0.51)$	a_2^j	$(0.0, 1.0)$	a_2^i	a_2^j	0	0	46.7	33.3
7	$(0.07, 0.64)$	a_2^j	$(0.0, 0.0)$	a_1^i	a_2^j	1	5	42.9	42.9
8	$(0.07, 0.71)$	a_2^j	$(0.0, 1.0)$	a_2^i	a_1^j	5	1	50.0	40.0
9	$(0.05, 0.71)$	a_1^j	$(0.0, 1.0)$	a_1^i	a_2^j	1	5	46.7	46.7
10	$(0.05, 0.76)$	a_2^j	$(0.0, 1.0)$	a_2^i	a_1^j	5	1	52.0	44.0

Tabelle 3.7 illustriert die Funktionsweise des Markovagenten. Beispielhaft wird hier eine tatsächliche Interaktion zwischen AgentM01 und einem Menschen im wiederholten Chicken Game nachgestellt. Da der Gegner sich in Runde 1 nicht auf das Kooperationsangebot einlässt, wechselt der Markovagent von einer Always Cooperate Strategie zur adaptiven Markovlogik. Folglich wählt AgentM01 auf Basis des Gegnermodells in Runde 2 bis 3 auf eine Tit-for-Tat Strategie $M^{*i}_{(0,1)} = (0,1)$ aus. Als der Gegner in Runde 3 trotz vormaligem Abweichen $a^i_{2,t-1}$ durch den Markovspieler eine Kooperation spielt, wechselt der Bot auf eine Always Defect Strategie $M^{*i}_{(0,1)} = (1,1)$. Grund ist, dass die Kooperation des Gegners trotz vorangegangener Ausbeutung nahelegt, dass sich dieser rechnerisch lohnenswert ausbeuten lässt. Zu diesem Schluss kommt der Markovspieler anhand des aktualisierten Gegnermodells $\hat{M}^j_{(0,1)}$. Zugrunde liegende Annahme dabei ist, dass der Schätzer des Gegnerverhaltens $(0.10, 0.27)$, beziehungsweise $(0.07, 0.27)$, das tatsächliche Gegnerverhalten vollständig abbildet. Diese Annahme stellt sich für den Gegner als unzutreffend heraus. Die Ausbeutungsversuche in Runde 4 und 5 scheitern, beide Spieler gehen mit einer Auszahlung von $r = 0$ leer aus. Das aktualisierte Gegnermodell $(0.07, 0.51)$ in Runde 6 vermutet daher einen deutlich vergeltungsbereiteren Gegner. Der Markovagent erwidert mit einer Tit-for-Tag Strategie. In Runde 7 berechnet der Markovagent auf Basis des neuen Gegnermodells, dass eine Always Cooperate Strategie payoffmaximierend ist. Als der Gegner daraufhin erneut ausbeutet, wechselt der Markovagent zurück zu einer Tit-for-Tat Strategie. Die beiden Spieler einigen sich dann auf eine alternierende Spielfolge, deren erwartete durchschnittliche Auszahlung äquivalent zur Kooperationslösung ist. Ergebnis in Runde 10 ist eine durchschnittliche normierte Auszahlung von 52.0 für AgentM01, während der

Gegenspieler einen Durchschnittswert von 44.0 ausweisen kann. Die Leistungsfähigkeit des in diesem Kapitel beschriebene Markovagenten soll nachfolgend anhand empirischer Erhebungen statistisch untersucht werden.

4 Methodik der experimentellen Untersuchung

Ein wesentlicher Unterschied zur Beurteilung der Leistungsfähigkeit von interaktiven Methoden ist, dass im Gegensatz zu deskriptiven Methoden wie die von Müller (2018) aufgrund der Pfadabhängigkeit des Spielverlaufs nicht ohne Weiteres auf historische Daten zurückgegriffen werden kann. Um die Leistungsfähigkeit von AgentM validieren zu können, wird daher eine eigene empirische Untersuchung angestrebt. Es handelt sich um ein Experiment im Rahmen einer ökonomischen Umgebung, die durch die Aspekte *Institution* und *Agenten* beschrieben werden kann, wobei erstere Kontext für die Interaktion der letzteren bildet (vgl. D. Friedman & Sunder, 1994, S. 12). Die charakterisierenden Eigenschaften dieser Umgebung werden anhand der folgenden Kapitel abgearbeitet:

1. **Experimentdesign:** Welche Gestaltungsentscheidungen hinsichtlich Spielauswahl, Spielerpaarung, Abbruchbedingung und Anreizsystem werden getroffen (siehe Kapitel 4.1)?
2. **Umsetzung der Experimente:** Wie werden die Experimente technologisch abgebildet und welche Aufbau- und Ablauforganisation wird zugrunde gelegt (siehe Kapitel 4.2)?
3. **Teilnehmerauswahl:** Wie werden Teilnehmer ausgewählt, beziehungsweise eingeladen und wie lassen sich diese beschreiben (siehe Kapitel 4.3)?
4. **Datenauswertung:** Welche übergeordnete Herangehensweise an die nachfolgende Datenauswertung wird angestrebt (siehe Kapitel 4.4)?
5. **Probelauf:** Wie kann ein robuster Experimenthergang sichergestellt werden (siehe Kapitel 4.5)?

4.1 Experimentdesign

Maßgeblich für den Erfolg empirischer Vorhaben ist ein sorgfältiges methodisches Design der Datenerhebung, welches anhand der folgenden Aspekte adressiert werden soll:

1. **Spielauswahl & Auszahlungsmatrix:** Im Rahmen welcher Spiele mit welchen Auszahlungsmodalitäten finden die Interaktionen statt (siehe Kapitel 4.1.1)?
2. **Spielerpaarung:** Welchen Überlegungen muss die Spielerpaarung gerecht werden und wie kann diese operationalisiert werden (siehe Kapitel 4.1.2)?

3. **Abbruchbedingung:** Nach welcher Logik soll die Spiellänge unter Berücksichtigung möglicher Implikationen für Probandenverhalten und Experimentablauf bestimmt werden (siehe Kapitel 4.1.3)?

4. **Anreizsystem:** Wie kann eine effektive und effiziente Gewährleistung adäquater Präferenzen auf Teilnehmerseite umgesetzt werden (siehe Kapitel 4.1.4)?

4.1.1 Spielauswahl und Gestaltung der Auszahlungsmatrix

Die Validierung des Markovagenten wird maßgeblich durch die Spezifitäten des gewählten Spiels determiniert, anhand derer sie stattfindet. Diese Arbeit beschränkt sich auf 2x2 Spiele, da sie die kondensierteste Form eines Interaktionsproblems mit zwei Akteuren darstellen. Die schlanke Struktur der Spiele lässt dabei jedoch, insbesondere im Kontext wiederholter Spiele mit unbekanntem Ende, keinen unmittelbaren Rückschluss auf Trivialität zu. Das Prisoner's Dilemma verdeutlicht diesen Sachverhalt. Im Stufenspiel wollen zwei rationale Spieler stets abweichen. Auch im wiederholten Spiel mit bekanntem Ende, kann jede Runde als eigenständiges Stufenspiel betrachtet werden. Findet jedoch eine wiederholte Interaktion mit unbekanntem Ende statt, greift die vorangegangene Logik nicht länger, sodass neben einer kontinuierlichen Abweichung auch Strategien wie Tit-for-Tat, Grim Trigger oder Pavlolv ein Nash-Gleichgewicht bilden können, in dem beide Spieler die gleiche Strategie spielen (vgl. Agrawal & Jaiswal, 2012, S. 1-2). Die Frage nach der *besten Strategie* für ein Spiel kann nicht länger eindeutig beantwortet werden, insbesondere, da die Leistungsfähigkeit einer Strategie von der Konstitution der Menge der anderen Strategien in der Gegnerpopulation abhängt (vgl. Axelrod, 1997). Ein derartiges Framing der Leistungsbeurteilung in Abhängigkeit von einer, sich möglicherweise verändernden, Gegnerpopulation wird im Kontext von MAL mit lernenden Akteuren als geeigneter Zugang zur Thematik bewertet (vgl. T. Sandholm, 2007; Shoham et al., 2007).

Die Menge der für die empirische Untersuchung zur Verfügung stehenden 2x2 Spiele ist groß, wobei sich die Spieltheorie bisweilen auf wenige, mit besonders interessanten Eigenschaften ausgestattete, Spieler konzentriert. Für die Validierung von AgentM sollen mehrere Spiele ausgewählt werden, die sich durch spezifische relevante Anreizstrukturen für die Spieler auszeichnen. Die für die Experimente getroffene Spielauswahl orientiert sich an der Systematik von Bruns (2015) und D. Robinson und Goforth (2006). In dieser werden analog zu Abbildung 4.1 sämtliche 144 ordinalen Anordnungen möglicher Auszahlungsrangkonstellationen von 2x2 Spiele anhand der daraus resultierenden Anreizstrukturen in die Familien *Win-Win*, *Cyclic*, *Inferior*, *Second Best*, *Unfair* und *Biased* klassifiziert.

	Nc	Ha	Pc	Co	As	Sh	Pd	Dl	Cm	Hr	Ba	Ch
	ChNc	ChHa	ChPc	ChCo	ChAs	ChSh	ChPd	ChDl	ChCm	ChHr	ChBa	**Ch**
Ch	2 3 3 4	2 2 3 4	2 1 3 4	2 1 3 4	2 2 3 4	2 3 3 4	2 4 3 3	2 4 3 2	2 4 3 1	2 4 3 1	2 4 3 2	2 4 3 3
	1 1 4 2	1 1 4 3	1 2 4 3	1 3 4 2	1 3 4 1	1 2 4 1	1 2 4 1	1 3 4 1	1 3 4 2	1 2 4 3	1 1 4 3	1 1 4 2
	BaNc	BaHa	BaPc	BaCo	BaAs	BaSh	BaPd	BaDl	BaCm	BaHr	**Ba**	BaCh
Ba	3 3 2 4	3 2 2 4	3 1 2 4	3 1 2 4	3 2 2 4	3 3 2 4	3 4 2 3	3 4 2 2	3 4 2 1	3 4 2 1	3 4 2 2	3 4 2 3
	1 1 4 2	1 1 4 3	1 2 4 3	1 3 4 2	1 3 4 1	1 2 4 1	1 2 4 1	1 3 4 1	1 3 4 2	1 2 4 3	1 1 4 3	1 1 4 2
	HrNc	HrHa	HrPc	HrCo	HrAs	HrSh	HrPd	HrDl	HrCm	**Hr**	HrBa	HrCh
Hr	3 3 1 4	3 2 1 4	3 1 1 4	3 1 1 4	3 2 1 4	3 3 1 4	3 4 1 3	3 4 1 2	3 4 1 1	3 4 1 1	3 4 1 2	3 4 1 3
	2 1 4 2	2 1 4 3	2 2 4 3	2 3 4 2	2 3 4 1	2 2 4 1	2 2 4 1	2 3 4 1	2 3 4 2	2 2 4 3	2 1 4 3	2 1 4 2
	CmNc	CmHa	CmPc	CmCo	CmAs	CmSh	CmPd	CmDl	**Cm**	CmHr	CmBa	CmCh
Cm	2 3 1 4	2 2 1 4	2 1 1 4	2 1 1 4	2 2 1 4	2 3 1 4	2 4 1 3	2 4 1 2	2 4 1 1	2 4 1 1	2 4 1 2	2 4 1 3
	3 1 4 2	3 1 4 3	3 2 4 3	3 3 4 2	3 3 4 1	3 2 4 1	3 2 4 1	3 3 4 1	3 3 4 2	3 2 4 3	3 1 4 3	3 1 4 2
	DlNc	DlHa	DlPc	DlCo	DlAs	DlSh	DlPd	**Dl**	DlCm	DlHr	DlBa	DlCh
Dl	1 3 2 4	1 2 2 4	1 1 2 4	1 1 2 4	1 2 2 4	1 3 2 4	1 4 2 3	1 4 2 2	1 4 2 1	1 4 2 1	1 4 2 2	1 4 2 3
	3 1 4 2	3 1 4 3	3 2 4 3	3 3 4 2	3 3 4 1	3 2 4 1	3 2 4 1	3 3 4 1	3 3 4 2	3 2 4 3	3 1 4 3	3 1 4 2
	PdNc	PdHa	PdPc	PdCo	PdAs	PdSh	**Pd**	PdDl	PdCm	PdHr	PdBa	PdCh
Pd	1 3 3 4	1 2 3 4	1 1 3 4	1 1 3 4	1 2 3 4	1 3 3 4	1 4 3 3	1 4 3 2	1 4 3 1	1 4 3 1	1 4 3 2	1 4 3 3
	2 1 4 2	2 1 4 3	2 2 4 3	2 3 4 2	2 3 4 1	2 2 4 1	2 2 4 1	2 3 4 1	2 3 4 2	2 2 4 3	2 1 4 3	2 1 4 2
	ShNc	ShHa	ShPc	ShCo	ShAs	**Sh**	ShPd	ShDl	ShCm	ShHr	ShBa	ShCh
Sh	1 3 4 4	1 2 4 4	1 1 4 4	1 1 4 4	1 2 4 4	1 3 4 4	1 4 4 3	1 4 4 2	1 4 4 1	1 4 4 1	1 4 4 2	1 4 4 3
	2 1 3 2	2 1 3 3	2 2 3 3	2 3 3 2	2 3 3 1	2 2 3 1	2 2 3 1	2 3 3 1	2 3 3 2	2 2 3 3	2 1 3 3	2 1 3 2
	AsNc	AsHa	AsPc	AsCo	**As**	AsSh	AsPd	AsDl	AsCm	AsHr	AsBa	AsCh
As	1 3 4 4	1 2 4 4	1 1 4 4	1 1 4 4	1 2 4 4	1 3 4 4	1 4 4 3	1 4 4 2	1 4 4 1	1 4 4 1	1 4 4 2	1 4 4 3
	3 1 2 2	3 1 2 3	3 2 2 3	3 3 2 2	3 3 2 1	3 2 2 1	3 2 2 1	3 3 2 1	3 3 2 2	3 2 2 3	3 1 2 3	3 1 2 2
	CoNc	CoHa	CoPc	**Co**	CoAs	CoSh	CoPd	CoDl	CoCm	CoHr	CoBa	CoCh
Co	2 3 4 4	2 2 4 4	2 1 4 4	2 1 4 4	2 2 4 4	2 3 4 4	2 4 4 3	2 4 4 2	2 4 4 1	2 4 4 1	2 4 4 2	2 4 4 3
	3 1 1 2	3 1 1 3	3 2 1 3	3 3 1 2	3 3 1 1	3 2 1 1	3 2 1 1	3 3 1 1	3 3 1 2	3 2 1 3	3 1 1 3	3 1 1 2
	PcNc	PcHa	**Pc**	PcCo	PcAs	PcSh	PcPd	PcDl	PcCm	PcHr	PcBa	PcCh
Pc	3 3 4 4	3 2 4 4	3 1 4 4	3 1 4 4	3 2 4 4	3 3 4 4	3 4 4 3	3 4 4 2	3 4 4 1	3 4 4 1	3 4 4 2	3 4 4 3
	2 1 1 2	2 1 1 3	2 2 1 3	2 3 1 2	2 3 1 1	2 2 1 1	2 2 1 1	2 3 1 1	2 3 1 2	2 2 1 3	2 1 1 3	2 1 1 2
	HaNc	**Ha**	HaPc	HaCo	HaAs	HaSh	HaPd	HaDl	HaCm	HaHr	HaBa	HaCh
Ha	3 3 4 4	3 2 4 4	3 1 4 4	3 1 4 4	3 2 4 4	3 3 4 4	3 4 4 3	3 4 4 2	3 4 4 1	3 4 4 1	3 4 4 2	3 4 4 3
	1 1 2 2	1 1 2 3	1 2 2 3	1 3 2 2	1 3 2 1	1 2 2 1	1 2 2 1	1 3 2 1	1 3 2 2	1 2 2 3	1 1 2 3	1 1 2 2
	Nc	NcHa	NcPc	NcCo	NcAs	NcSh	NcPd	NcDl	NcCm	NcHr	NcBa	NcCh
Nc	2 3 4 4	2 2 4 4	2 1 4 4	2 1 4 4	2 2 4 4	2 3 4 4	2 4 4 3	2 4 4 2	2 4 4 1	2 4 4 1	2 4 4 2	2 4 4 3
	1 1 3 2	1 1 3 3	1 2 3 3	1 3 3 2	1 3 3 1	1 2 3 1	1 2 3 1	1 3 3 1	1 3 3 2	1 2 3 3	1 1 3 3	1 1 3 2

Abbildung 4.1: Spieltypen gemäß des Periodensystem von 2x2 Spielen (Bruns, 2015; D. Robinson & Goforth, 2006) mit Rängen der Payoffs für die Spiele Chicken (Ch), Battle (Ba), Hero (Hr), Compromise (Cm), Deadlock (Dl), Prisoner's Dilemma (Pd), Stag Hunt (Sh), Assurance (Ar), Coordination (Co), Peace (Pc), Harmony (Ha), Concord/No Conflict (Nc); symmetrische Spiele liegen auf der Diagonalen und sind mit einem grünen Rahmen hervorgehoben. Quelle: Müller (2018, S. 46).

Die Auswahl der zu betrachtenden Spiele sollte analog zu Müller (2018) (1) symmetrisch sein, (2) interessante spieltheoretische Interaktionen durch strategische Heterogenität ermöglichen und (3) eine hinreichend große Datenlage erzeugen können:

1. **Symmetrie:** Die Verwendung ausschließlich symmetrischer Spiele ist für die wissenschaftliche Fragestellung der Arbeit unerlässlich. Ziel ist es, die Auszahlungsunterschiede zwischen Spielertypen statistisch zu analysieren. Symmetrische Spiele erlauben hier eine möglichst saubere Datenausgangslage, bei der die Position des Spielers in Bezug auf Spalte oder Zeile keinen Einfluss auf den erzielten Payoff hat (vgl. Willer & Walker, 2007). Darüber hinaus wurde die Effektivität der AgentM zugrundeliegenden deskriptiven Methodik von Müller (2018) bis dato lediglich für symmetrische Spiele gezeigt, weshalb eine Beschränkung auf derartige Spiele naheliegt. Zusätzlich vereinfachen symmetrische Spiele das Interaktionsverständnis für die Probanden, da ihr Gegenspieler identische Anreize besitzt. Infolgedessen würden nicht-symmetrische Spiele für die wissenschaftliche Fragestellung der Arbeit nicht zielführende Komplexität verursachen, dem kein unmittelbarer Mehrwert gegenübersteht.

2. **Strategische Heterogenität:** Die ausgewählten Spiele sollten im wiederholten Fall reichhaltige strategische Überlegungen erzeugen können. Dies ist insbesondere dann gegeben, wenn entweder *konfliktäre* oder *koordinative* Anreize vorliegen. Folglich sind Spiele der Kategorie Win-Win per Definition aufgrund ihrer trivialen Auszahlungsstruktur für den Untersuchungsgegenstand dieser Arbeit uninteressant, da für die Spieler kein Anreizkonflikt herrscht und beide Parteien stets gemeinsam den besten Zustand in einer trivialen Reinstrategie erreichen würden. Auch wird die Spielfamilie Second Best nicht weiter betrachtet. Neben beschränkten Ressourcen der Datenerhebung wird hier aufgrund einer dominanten Strategie für beide Spieler eine weniger reichhaltige Interaktionspalette erwartet.

3. **Datenlage:** Um eine hinreichend große Datenbasis zu generieren, wird auf die Untersuchung diverser Ausgestaltungen der gleichen Spielklasse verzichtet und nur eine spezifische Variante je ausgewählte Klasse betrachtet. Dadurch wird insbesondere sichergestellt, dass mit der Menge an Datenpunkten, die im Rahmen eines Experiments erzeugt werden können aussagekräftige quantitative Schlussfolgerungen möglich sind (vgl. Müller, 2018).

Unter Berücksichtigung der Kriterien werden lediglich Spiele auf der Diagonalen von Abbildung 4.1 unter Ausschluss der trivialeren Spielfamilien Win-Win und Second Best betrachtet. Aus der Inferior-Familie verbleibt lediglich das *Prisoner's Dilemma* auf der Diagonalen, während aus der Unfair-Familie lediglich das *Chicken Game* auf der Diagonalen verbleibt. Nach

Tabelle 4.1: Gleichgewichtseigenschaften im Stufenspiel der ausgewählten Spiele; mit Nash-Gleichgewicht des Stufenspiels (N), paretoeffizientem Zustand (grüne Farbe) und paretodominiertem Zustand (rote Farbe). Quelle: Eigene Darstellung mit Spielfamilien nach Bruns (2015) und D. Robinson und Goforth (2006).

(a) Prisoner's Dilemma aus der Inferior-Spielfamilie.

	a_1^j	a_2^j
a_1^i		
a_2^i		N

(b) Chicken Game aus der Unfair-Spielfamilie.

	a_1^j	a_2^j
a_1^i		N
a_2^i	N	

(c) Hero Game aus der Biased-Spielfamilie von .

	a_1^j	a_2^j
a_1^i		N
a_2^i	N	

der guten Abdeckung konfliktärer Spiele wird aus der Biased-Familie das *Hero Game* gewählt. Das Hero Game legt im Vergleich mit dem *Battle Game* als Alternative aus der Familie größeren Fokus auf konfliktäre Aspekte im Zusammenspiel mit koordinativen Aspekten.[41] Folglich ist eine erfolgreiche Koordination im Hero Game weniger leicht zu erreichen. Wird sie dennoch erreicht, spricht dies umso mehr für die Leistungsfähigkeit der Spieler (vgl. Bruns, 2015). Tabelle 4.1 illustriert diese Dynamiken anhand der Gleichgewichtseigenschaften der Spiele.

Tabelle 4.2: Auszahlungsmatrizen der ausgewählten Spiele. Quelle: Eigene Darstellung mit Spielfamilien nach Bruns (2015) und D. Robinson und Goforth (2006).

(a) Prisoner's Dilemma aus der Inferior-Spielfamilie.

	a_1^j	a_2^j
a_1^i	3/3	0/5
a_2^i	5/0	1/1

(b) Chicken Game aus der Unfair-Spielfamilie.

	a_1^j	a_2^j
a_1^i	3/3	1/5
a_2^i	5/1	0/0

(c) Hero Game aus der Biased-Spielfamilie.

	a_1^j	a_2^j
a_1^i	0/0	3/5
a_2^i	5/3	1/1

Die ausgewählten Spiele sollen nun hinsichtlich ihrer konkreten Auszahlungswerte definiert werden. Als Ausgangsbasis dient die, in Tabelle 4.2a gezeigte, klassische Definition des Prisoner's Dilemma gemäß der Strategieturniere von Axelrod (1984). Für die anderen Spiele wurden darauf aufbauend die identischen individuellen Auszahlungswerte bei unterschiedlicher Anordnung verwendet, um eine potentielle Inferenz durch mögliche Verhaltensimplikationen der Spieler über die Spielfamilien hinweg zu reduzieren. Weiterhin wurden ausschließlich positive Auszahlungswerte gewählt. Im Gegensatz zur klassischen Spieltheorie, legen empirische Erkenntnisse auf Basis der Prospect Theory von Kahneman und Tversky (1979) nahe, dass die Niveaus der Auszahlungswerte, insbesondere bei einem Vorzeichenwechsel (vgl. z.B. Cachon

[41] Auf den eigenen Höchstpayoff zu verzichten um sich mit dem anderen Spieler auf den zweithöchsten Payoff zu koordinieren birgt die Gefahr, dass beide Spieler leer ausgehen, sofern auch der Gegner zurückstecken möchte. Im *Battle Game* hingegen wird ein beidseitiger Verzichtsversuch weniger stark bestraft.

& Camerer, 1996; Feltovich et al., 2012), menschliches Verhalten beeinflussen. Tabelle 4.2 stellt die finalen Auszahlungsmatrizen der Experimente dar.

4.1.2 Spielerpaarung

Die Gestaltung der Paarungslogik der Spielerpopulation ist entscheidend für die Gewährleistung robuster Experimentdaten. Übergeordnetes Ziel ist die Minimierung und Kontrolle von Störfaktoren sowie Maximierung der Validität (vgl. D. Friedman & Sunder, 1994, S. 21). Folglich soll zunächst die Relevanz der Kohortenbildung in den Kontext des Forschungsziels eingebettet und anschließend hinsichtlich der konkreten Paarungslogik operationalisiert werden.

4.1.2.1 Spielerpaarung im Kontext des Forschungsgegenstandes

Das nachfolgende Kapitel beschäftigt sich mit der Gestaltung der Spielerpaarung in Hinblick auf die Anforderungen des Forschungsvorhabens. Den Probanden können künstliche und andere menschliche Spieler als Gegner zugewiesen werden. Unter *künstlichen Spielern* werden

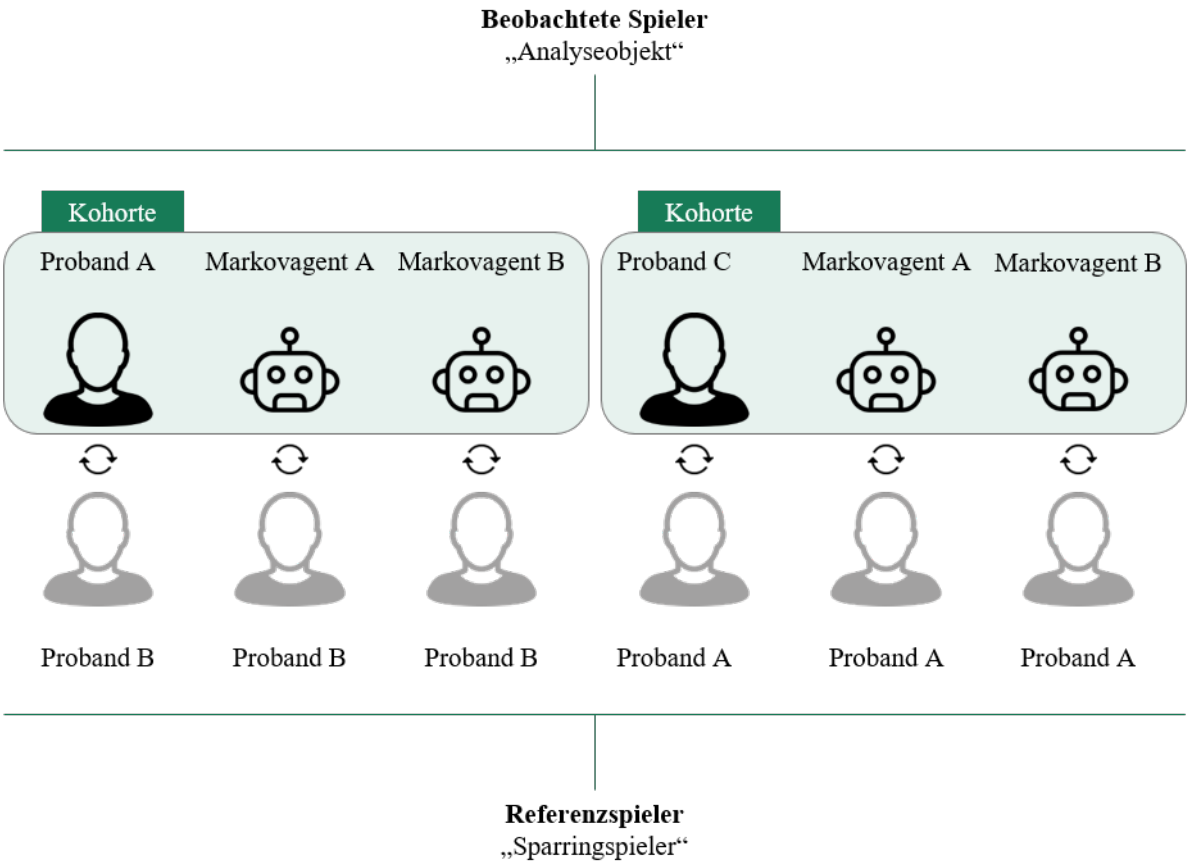


Abbildung 4.2: Exemplarische Darstellung von zwei Kohorten für die Paarung von drei menschlichen und zwei künstlichen Spielern. Jeder Proband spielt einmal gegen jeden Agenten und zweimal gegen einen nicht-identischen Menschen. Quelle: Eigene Darstellung.

sowohl MAL Algorithmen wie AgentM, als auch nicht-lernende algorithmische Strategien wie Tit-for-Tat zusammengefasst. Hierzu sei zunächst darauf hingewiesen, dass die Leistungsfähigkeit der Spieler nicht relativ zu der des unmittelbaren Gegenspielers betrachtet werden kann. Ein derartiger Vergleich wäre nicht konsistent mit der Zielfunktion der Spieler, die eigene Auszahlung *absolut*, also unabhängig von der Auszahlung des Gegenspielers, zu maximieren. Ein Vergleich der Auszahlung mit dem unmittelbaren Gegenspieler impliziert jedoch die Zielfunktion, die eigene Auszahlung *relativ* zu der des unmittelbaren Gegenspielers zu verbessern. Aus letzterem Fall würde neben der eigenen Payoffmaximierung ein unerwünschter expliziter Anreiz zur simultanen Payoffminimierung des Gegners resultieren.[42] Folglich können also nie die Auszahlungen zweier Spieler im Spiel gegeneinander verglichen werden. Gleichzeitig ist es erforderlich die Leistungen der Spielertypen relativ in Verbindung zu setzen, da sonst keine bewertende Aussage diesbezüglich möglich ist. Um also beispielsweise die Leistungsfähigkeit eines Tit-for-Tat-Spielers relativ zu menschlichen Spielern beurteilen zu können, muss die Leistung des Tit-for-Tat-Spielers mit der eines Menschen im jeweiligen Spiel gegen einen konstanten dritten Spieler bewertet werden.

Grundsätzlich besteht weiterhin die Wahl zwischen einem *within Subject* und einenem *between Subject* Design. Im within Subject Fall wird jeder Proband mehr als einer Spielsituation ausgesetzt, sodass eine Veränderung der beobachteten Variable mit der Veränderung der Situation in Verbindung gebracht werden kann, wobei die Situation im Rahmen dieser Arbeit durch den Gegenspieler charakterisiert wird. Im between Subject Design wird jeder Proband nur genau einer Spielsituation ausgesetzt und Effekte über unterschiedliche Spieler hinweg verglichen (vgl. Charness et al., 2012).

Im Kontext dieser Arbeit wird das within Subject Design vorgezogen. Es kontrolliert für die spezifischen Eigenschaften der Teilnehmer (vgl. D. Friedman & Sunder, 1994, S. 25) und reduziert die Menge der benötigten Teilnehmer für eine robuste Stichprobengröße. Das within Subject Design trägt dabei insbesondere der Abhängigkeit der realisierten Auszahlung von der Strategie des Gegners Rechnung. Spielt beispielsweise ein Tit-for-Tat-Spieler im wiederholten Prisoner's Dilemma gegen einen stets kooperierenden Spieler, ist dessen Auszahlungsergebnis klar unterschiedlich von dem eines Spiels gegen einen stets abweichenden Spieler. Um also die Leistungsfähigkeit des Tit-for-Tat-Spielers beurteilen zu können, muss dessen Auszahlung mit der Auszahlung anderer Spieler gegen den identischen Gegenspieler verglichen werden. So kann die Abhängigkeit der eigenen Auszahlung von den spezifischen Eigenschaften des Gegenspielers reduziert werden. Zusammenfassend wird die Leistungsfähigkeit der analysierten Spieler im relativen Vergleich untereinander aber im Spiel gegen den identischen Sparringspieler mit seinen spezifischen Eigenschaften und strategischen Überlegungen verglichen. Ziel

[42] Klar ist, dass die Payoffstrukturen der gewählten Spiele derart gestaltet sind, dass eine Maximierung der eigenen Payoffs einen negativen Effekt auf die Payoffs des Gegenspielers haben kann. Dieser Effekt besteht jedoch rein implizit und der eigenen Payoffmaximierung als Notwendigkeit untergeordnet.

ist es folglich, pro Proband im Sinne eines Sparringspielers eine Beobachtung bezüglich der gegnerischen realisierten Auszahlung je Spielertyp zu machen. Die realisierte Auszahlung der Sparringspieler wird folgerichtig nicht betrachtet.

Abbildung 4.2 stellt den Zusammenhang beispielhaft für ein Experiment mit menschlichen Spielern sowie einem Markovagent und einem Tit-for-Tat-Spieler dar. Die Rollen der Spielpartner gliedert sich in dem Beispiel wie folgt: Die Leistungsfähigkeit der verschiedenen Spielertypen wird anhand eines relativen Vergleichs innerhalb der eigenen Kohorte durchgeführt. Eine Kohorte wird dadurch charakterisiert, dass die Leistung der Spieler gegen den identischen menschlichen Sparringspieler beurteilt wird. Der Sparringspieler wird dabei als Teil der Experimentumgebung betrachtet, dessen Rolle es ist, die Varianz der Leistung der beobachteten Spieler möglichst auf den jeweiligen Spielertyp zu begrenzen, ohne dass große Inferenzen mit der gegnerischen Strategiewahl bestehen. Die Leistung des Sparringspielers wird nicht betrachtet. Jeder Proband tritt im abgebildeten Beispiel somit dreimal als Sparringspieler (zweimal gegen einen Markovagenten und einmal gegen einen Menschen) und einmal als Analyseobjekt (gegen einen anderen Menschen) auf.

4.1.2.2 Logistik der Spielerpaarung

Das folgende Kapitel befasst sich mit der operativen Abbildung der Spielerpaarungen im Experiment. Generell lassen sich die *Neupaarung während eines Spiels* und die *Neupaarung zwischen Spielen* unterscheiden (vgl. Müller, 2018). Ein Turniermodus, in dem jeder gegen jeden Spielt ist aufgrund der Kombinatorik der Paarungen logistisch nicht effizient.

Im laufenden Spiel ist eine Neupaarung nach jeder Runde eines wiederholten Spiels im Sinne einer Serie von oneshot Spielen möglich, alternativ können die Spielerpaare über alle Runden des wiederholten Spiels konstant gehalten werden (vgl. Müller, 2018). In einer vergleichenden Untersuchung stellen Andreoni und Croson (2008) fest, dass sich das beste Verfahren zur Betrachtung individuellen Spielerverhaltens nicht eindeutig bestimmen lässt. Hintergrund ist, dass erhöhte Kooperationsraten sowohl bei neu gepaarten Spielern (vgl. Andreoni, 1988; Palfrey & Prisbrey, 1996), als auch bei festen Paaren auftreten können (vgl. R. T. Croson, 1996; Keser & van Winden, 2000). Infolgedessen sollte sich die Spielerpaarung zwischen den Runden eines Spiels am Forschungsziel orientieren. Eine Neupaarung zwischen den Runden eines wiederholten Spiels ist im Rahmen dieser Arbeit nicht zielführend, da die Strategien für wiederholter Spiele der einzelnen Paarungen zentrales Betrachtungselement dieser Untersuchung sind (vgl. analog Müller, 2018). Folglich kommt kommt eine konstante Spielerpaarung über die Runden eines wiederholten Spiels zum Tragen.

Eine Neupaarung zwischen aufeinanderfolgenden Spielen wird nachfolgend anhand der Aspekte *Zulässigkeit der Gegner*, *Auswahl aus zulässigen Gegnern* und *Reihenfolge der ausgewählten Gegner* diskutiert und festgelegt. Zur Bestimmung der Zulässigkeit der Gegner muss zunächst

folgende Frage beantwortet werden: Welche Gegner kommen bei einer Neupaarung grundsätzlich in Frage? Hierzu lassen sich drei Ansätze in aufsteigender Robustheit unterscheiden (vgl. Dal Bo & Frechette, 2018):

- **Paarung mit Zurücklegen:** Nach jedem Spiel wird ein Gegner zufällig aus der Menge der anderen Probanden ausgewählt. Eine wiederholte Paarung mit einem bereits bekannten Gegner ist möglich.
- **Paarung ohne Zurücklegen:** Der ausgewählte neue Gegner muss ein *Fremder* sein und darf demnach noch nicht mit dem jeweils anderen Spieler auf Basis voriger Spiele bekannt sein. Wiederholte Paarungen sind somit ausgeschlossen.
- **Turnpike Protokoll:** Die Definition des Fremden wird um die Transitivitätseigenschaft erweitert, sodass Spieler A nicht mit Spieler B gepaart werden darf, sofern Spieler A in der Vergangenheit bereits mit Spieler C und Spieler C bereits mit Spieler B gepaart wurde (McKelvey & Palfrey, 1992).

Eine Paarung ohne Zurücklegen eignet sich als Kompromiss aus Reduktion von Spielverhalten auf Basis von Erfahrungswerten über die gegnerische Strategie aus vorigen Spielen und der Anzahl möglicher Paarungen (vgl. Müller, 2018). Insbesondere ist das Turnpike Protokoll ungeeignet, da menschliche Spieler nicht gegen andere Menschen spielen können, die gegen den gleichen künstlichen Gegner gepaart werden sollen. Infolgedessen würde nicht nur der Bedarf an Probanden steigen, sondern auch die Vergleichbarkeit der erhobenen Daten leiden, da die Leistung verschiedener Spielertypen nicht innerhalb des selben Menschen verglichen werden könnte. Eine Paarung ohne Zurücklegen ist aufgrund der vorhanden Probandenzahl zur Erzeugung ausreichender Datenmengen nicht erforderlich und in Bezug auf die Gewinnung aussagekräftiger Daten der Paarung ohne Zurücklegen aufgrund wiederholter Paarungen unterlegen.

Die Auswahl aus zulässigen Gegnern beantwortet die Kernfrage: Welche der insgesamt zulässigen Spieler werden als Gegner eines Probanden ausgewählt? Um unreflektiertes und unengagiertes Spielverhalten durch zu häufiges Spielen des gleichen Spiels zu vermeiden, beschränken sich die Experimente dieser Arbeit auf eine möglichst geringe Anzahl an Spielen je Gegner. Auf Basis dieser ressourceneffizienten Experimentplanung soll jeder Proband nur genau einmal mit jedem der $N_{künstlich}$ zu untersuchenden künstlichen Gegner gepaart werden. Darüber hinaus ist bei $N_{menschlich}$ Probanden die Paarung mit bis zu $N_{menschlich} - 1$ menschlichen Gegnern möglich. Unter Berücksichtigung der Ausführungen in Kapitel 4.1.2.1, erfolgt im Rahmen der Experimente dieser Arbeit eine Paarung eines jeden Probanden mit je 2 weiteren menschlichen Spielern (siehe insbesondere Abbildung 4.2). Jeder Mensch nimmt somit an $N_{künstlich} + 2$ Spielen teil.

Die Festlegung der Reihenfolge der ausgewählten Gegner beantwortet folgende Frage: In welcher Reihenfolge wird mit den ausgewählten Gegnern gespielt? Der Überlegung liegt zum einen eine Reduktion von *Lerneffekten* sowie *Reihenfolgeeffekten* zugrunde. Lerneffekte werden durch sich aufgrund wachsender Erfahrung systematisch über Zeit anpassendes Spielerverhalten charakterisiert. Derartige Lerneffekte können sich beispielsweise im Rahmen des wiederholten Prisoner's Dilemma in zunehmenden Kooperationsraten niederschlagen (vgl. Dal Bo & Frechette, 2018). Derartige Lerneffekte können sich sowohl über die Runden eines Spiels als auch über aufeinanderfolgende Instanzen des gleichen Spiels erstrecken. Weiterhin können Reihenfolgeeffekte im Sinne eines Einflusses des Verhaltens der vorigen Spielerpaarungen auftreten (vgl. Dal Bo & Frechette, 2018). Insbesondere beim gewählten within Subject Design ist den Gefahren von *Carry Over Effekten* und *Experimenter Demand Effekten* Rechnung zu tragen (vgl. Charness et al., 2012):

- **Carry Over Effekte:** Emotionen und Erfahrungen, die aus der Interaktion mit einem spezifischen Gegnerspieler resultieren, können sich auf darauffolgende Interaktionen auswirken. So kann sich beispielsweise eine Strategie gegen den ersten Gegner als vergleichsweise unfruchtbar entpuppen, wodurch die Wahrscheinlichkeit, dass der Spieler sie gegen den zweiten Gegner einsetzt, unabhängig von deren grundsätzlicher Sinnhaftigkeit, sinkt (vgl. Charness et al., 2012).
- **Experimenter Demand Effekte:** Probanden können bewusst oder unbewusst auf etwaige Muster in der Abfolge der Spielsituationen oder den von ihnen vermuteten Untersuchungsgegenstand des Experiments reagieren und ihr Verhalten dementsprechend anpassen (vgl. Charness et al., 2012).

Um die Auswirkungen dieser Effekte zu reduzieren, erfolgt die Reihenfolge der Paarungen randomisiert.[43] Weiterhin bleibt die Identität des Gegners geheim, sodass Störeffekte und Erfahrungen nicht zurechenbar sind und minimiert werden (vgl. Charness et al., 2012).

Zusammenfassend findet für diese Untersuchung nach jedem Spiel eine Neupaarung mit einem völlig fremdem Gegner statt, sodass jeder Proband gegen alle künstlichen Spieler und zwei weitere menschliche Gegner in zufälliger Reihenfolge spielt.

4.1.3 Abbruchbedingung

Die Abbruchbedingung spieltheoretischer Experimente kann Einfluss auf das Spielverhalten haben und ist daher wichtiges Gestaltungsmerkmal derartiger Untersuchungen (vgl. Dal Bo &

[43] Vereinzelt mussten die vorab zufällig festgelegten menschlichen Gegner beziehungsweise die Reihenfolge der Paarungen ad-hoc an den Experimentablauf angepasst werden. In diesen Fällen wurde das Paarungsschema aus technischen Gründen oder aufgrund des Nichterscheinens einzelner Probanden derart angepasst, dass dennoch erstens jeder Proband gegen alle künstlichen Gegner und zwei menschliche Gegner spielt und zweitens ein Paarungsmodus ohne Zurücklegen gewährleistet bleibt.

Frechette, 2018). Im Folgenden werden die drei Ansätze *bekanntes Ende*, *vollständig randomisiertes Ende* und *unbekannte feste Rundenzahl mit nachfolgend randomisiertem Ende* diskutiert. Letzterer Ansatz wird im Anschluss für die Verwendung in den Experimenten im ausgewählt:

- **Bekanntes Ende:** Aus theoretischer Sicht ist der Einfluss der Information über die Rundenzahl von wiederholten Spielen zentraler Bestandteil der Spieltheorie, wo Rückwärtsinduktion regelmäßig zur Bestimmung optimaler Aktionen in endlichen Extensivformspielen und sequentiellen Spielen herangezogen wird (Neumann & Morgenstern, 1944). Auch in der Praxis konnten derartige Endspieleffekte im Sinne von verändertem Spielverhalten bei bekanntem Ende in wiederholten Spielen von beispielsweise Normann und Wallace (2012) und Selten und Stoecker (1986) beobachtet werden.

- **Vollständig randomisiertes Ende:** Die zufällige Abbruchregel wurde erstmals von Roth und Murnighan (1978) und Murnighan und Roth (1983) vorgestellt. Ab der ersten Runde eines Spiels wird dieses mit einer spezifischen Wahrscheinlichkeit um die nächste Runde fortgeführt, wodurch sich eine zufällige Spiellänge ergibt. Die Verwendung einer rundenweisen Fortführungswahrscheinlichkeit induziert unter der Annahme von Risikoneutralität Präferenzen analog zu einem unendlich wiederholten Spiel mit einem Abzinsungsfaktor in Höhe der Fortführungswahrscheinlichkeit (vgl. Dal Bo & Frechette, 2018). Trotz der geläufigen Anwendung der zufälligen Abbruchregel (vgl. Aoyagi & Fréchette, 2009; Camera & Casari, 2009; Duffy & Ochs, 2009; Fudenberg et al., 2012) zur Modellierung unendlicher Spiele im Labor können folgende Kritikpunkte angeführt werden (vgl. Dal Bo & Frechette, 2018): Erstens könnten Probanden zu einer Fehleinschätzung über die Auswirkung der Abbruchwahrscheinlichkeit auf die Verteilung der Spiellänge kommen. Jedoch deuten die Ergebnisse von Dal Bó (2005) und Murnighan und Roth (1983) auf ein ausreichendes Verständnis der Probanden über diesen Zusammenhang hin. Zweitens könnten Probanden eine subjektive Erwartung über die Abbruchwahrscheinlichkeit hegen, welche von der implementierten Abbruchwahrscheinlichkeit abweicht. Der Einfluss subjektiver Erwartungen über die Abbruchwahrscheinlichkeit spielt im Rahmen dieser Untersuchung eine nachrangige Rolle, da keine direkten Zusammenhänge zwischen Abbruchwahrscheinlichkeit und Spielverhalten untersucht werden sollen (vgl. Dal Bo & Frechette, 2018). Drittens könnten Probanden von der Risikoneutralitätsannahme abweichen, wobei Untersuchungen von Sherstyuk et al. (2013) darauf hindeuten, dass eine derartige Abweichung keinen Einfluss auf das Spielverhalten hat.

- **Unbekannte feste Rundenzahl mit nachfolgend randomisiertem Ende:** Im Rahmen dieser von Sabater-Grande und Georgantzis (2002) vorgestellten Variante spielen die Probanden eine fixe Anzahl an Runden, auf welche nahtlos eine Fortführung auf Basis einer rundenweisen Fortführungswahrscheinlichkeit stattfindet. Es handelt sich demnach um

eine Kombination der ersten beiden Ansätze, die derart in Cabral et al. (2014) und Vespa (2011) zur Anwendung kommt, um unerwünscht kurze Spielverläufe im Falle einer vollständig randomisierten Abbruchregel zu vermeiden. Bei angenommener Risikoneutralität ist die Methode äquivalent zu einem unendlich wiederholten Spiel (vgl. Vespa, 2011).

Aufgrund der Implikationen auf das Spielverhaltens kommt ein Modus mit bekanntem Ende für diese Arbeit nicht weiter in Betracht (vgl. Agrawal & Jaiswal, 2012, S. 1-2). Für den zufälligen Abbruch stellen Fréchette und Yuksel (2017) Unterschiede in Spielverhalten abhängig vom konkret verwendeten Abbruchmodus fest. Normann und Wallace (2012) hingegen stellen keinen signifikanten Einfluss der Abbruchbedingung auf die durchschnittliche Kooperationsrate im Prisoner's Dilemma fest. Folglich muss sich der gewählte Abbruchmechanismus aufgrund nicht eindeutiger Erkenntnisse stets an dem jeweiligen Anwendungskontext orientieren (vgl. Dal Bo & Frechette, 2018). Klar ist, dass eine randomisierte Abbruchbedingung mit oder ohne eine vorangehende feste Rundenzahl in der Praxis nicht mit einem unendlich wiederholten Spiel gleichzusetzen ist. Gleichwohl kann und wird ein solcher Mechanismus in der wissenschaftlichen Praxis regelmäßig erfolgreich als eine hinreichende Näherung für das Konzept unendlich wiederholter Spiele verwendet (vgl. Dal Bo & Frechette, 2018). Kritisch anzumerken ist, dass die Verwendung einer Abbruchwahrscheinlichkeit zu stark unterschiedlichen Spiellängen führen mag, was erstens zu Heterogenität der gewonnen Paneldatenstruktur beitragen, zweitens im Extremfall den Experimenthergang behindern und drittens bei außergewöhnlich langen Spielverläufen das Engagement der Teilnehmer mindern kann. Viertens kann die Länge von vorangegangenen Spielen Erwartungen über die Länge zukünftiger Spiele bilden und so das Spielverhalten beeinflussen (Müller, 2018).

Den genannten Risiken wird im Rahmen dieser Arbeit durch die Verwendung (1) einer feste Rundenzahl T_{min} mit nachfolgend randomisiertem Ende (2) auf Basis eines im Vorhinein festgelegten gleichverteilten Rundenintervalls $[T_{min}, T_{max}]$ in Kombination mit (3) einer Analyse ausschließlich der Daten, welche im Rahmen des anfänglichen Blocks mit fester Rundenzahl erhoben wurde. Durch eine vorgelagerte feste Rundenzahl T_{min} werden kurze Spielverläufe mit geringem Erklärungsvermögen vermieden, sodass das Potential der Experimente besser genutzt werden kann (vgl. Cabral et al., 2014). Die Verwendung einer endlichen gleichverteilten Spanne

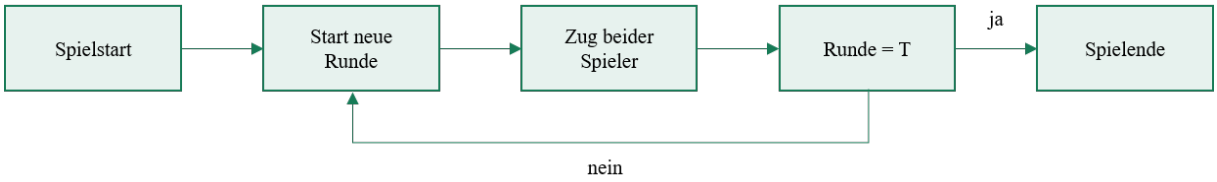


Abbildung 4.3: Ablauf eines Spiels mit einer zufälligen Rundenzahl T mit $T_{min} \leq T \leq T_{max}$. Quelle: Eigene Darstellung.

des zufälligen Teils der Spiellänge erlaubt für eine Kontrolle der maximalen Spieldauer T_{max}, sodass ein möglichst planbarer Experimentablauf möglich ist und die Motivation der Probanden nicht vereinzelt durch unverhältnismäßig Lange Spiele geschmälert wird (vgl. Davis & Holt, 1993). Durch die Beschränkung der Ergebnisanalyse auf den fixen Rundenteil wird darüber hinaus die Homogenität der Datenbasis gefördert und der Einfluss von möglichen Erwartungen über das Spielende aus vorigen Spielen geschmälert. Auf die Verwendung immer kürzer werdende Intervalle zur weiteren Reduktion einer derartigen Erwartungshaltung (Müller, 2018) wurde verzichtet, da folgerichtig auch darüber eine Erwartung gebildet werden kann. Zur Vermeidung von Erwartungshaltungen über die Spiellänge findet keine Kommunikation über die konkrete Abbruchbedingung und mögliche Spanne der Runden statt. Den Probanden ist lediglich bekannt, dass dem Spiel ein zufälliges Ende zugrunde liegt.

Folglich wird im Rahmen dieser Arbeit für die Anzahl Runden T eines Spiels in Anlehnung an Müller (2018) eine Kombination aus einem fixen Block von $T_{min} = 21$ Runden und zufällig festgelegten 0 bis 5 weiteren Runden verwendet (siehe Abbildung 4.3):

$$P(X = T) = \begin{cases} \frac{1}{6} & \forall T \in [21...26] \\ 0 & \text{sonst} \end{cases} \tag{4.1}$$

4.1.4 Anreizsystem

Eine geläufige Basis für das Anreizsystem ökonomischer Experimente ist die *Induced Preference Theory* von Smith (1976). Die Methode strebt an, die Nutzenfunktion der Probanden an die zugrunde gelegte Nutzenfunktion der Agenten des ökonomischen Modells anzugleichen. Die Vergütung sollte dabei die Eigenschaften *Dominanz*, *Salienz* und *Monotonie* aufweisen:

- **Salienz:** Die Vergütung des Probanden hängt von dessen Verhalten ab und dem Probanden ist der Kausalzusammenhang bewusst.
- **Dominanz:** Der aus der Vergütung resultierende sollte stärker sein als andere individuelle (de)motivierende Faktoren.
- **Monotonie:** Der Proband bevorzugt eine höhere stets gegenüber einer niedrigeren Vergütung.

In der Wirtschaftswissenschaft sind leistungsabhängige Vergütungsstrukturen geläufig, da diese durch publizierende Institutionen durchgesetzt werden. Im Fachbereich der Psychologie ist diese institutionelle Überwachung gewöhnlich nicht vorhanden, sodass die Methodenlandschaft hinsichtlich der Anreize weitaus heterogener ist (vgl. Hertwig & Ortmann, 2001). Eine Literaturanalyse von Hertwig und Ortmann (2001) kommt zu dem Schluss, dass eine leistungsabhängige Vergütung die Gründlichkeit steigert, mit der Probanden sich der Aufgabe widmen.

Gleichwohl weisen die Autoren darauf hin, dass erfolgsabhängige Vergütungsmodelle zwar eine Kontrollmöglichkeit für Experimente darstellen, die jedoch ungewollte Konsequenzen nach sich ziehen können. Beispielweise kann variable Vergütung dann schädlich sein, wenn sie die Teilnehmer sich ihrer selbst bewusst macht. Beispielweise beschreibt *Choking* eine hieraus resultierende schlechtere Leistung (vgl. Camerer & Hogarth, 1999). Auch Betsch und Haberstroh (2001) verweist auf ungewollte Konsequenzen leistungsabhängige Vergütung. Beispielweise steigt das Risiko einer verzerrten subjektiven Repräsentation der Experimentalaufgabe aufgrund des mit der variablen Vergütung zunehmenden Einflusses von Vorwissen, sodass dieses die zur Verfügung gestellten Informationen überschattet und maladaptive Verhaltensweisen induziert (vgl. Betsch, Haberstroh et al., 2001; Betsch, Plessner et al., 2001). Im Rahmen wiederholter Spiele wird hier die Gefahr des zunehmenden Einflusses von oneshot Strategien antizipiert, da diese häufiger Bestandteil wirtschaftswissenschaftlicher Grundausbildung ist. Ein derartiger Einfluss würde die Interaktionsqualität im wiederholten Setting der Experimente stark beeinträchtigen und die Relevanz der Ergebnisse gefährden.

Wird dennoch eine variable Vergütungsstruktur eingesetzt, kann nicht immer ein Verhaltenseffekt gezeigt werden. Beispielsweise finden einige Studien keinen Effekt der Vergütungsmodalitäten (vgl. z.B. Dawes, 1988; Hogarth et al., 1991; Stone & Ziebart, 1995). In einer Metaanalyse empirischer Forschung finden Jenkins et al. (1998), dass finanzielle Anreize zwar mit der Leistungsmenge (z.B. benötigte Zeit), aber nicht mit der Leistungsqualität (z.B. Genauigkeit) zusammenhängt. Der Literaturüberblick von Hertwig und Ortmann (2001) legt nahe, dass ein Hauptmehrwert variabler Vergütungsmodelle lediglich in der Reduktion der Variabilität der Ergebnisse liegt. In Märkten und Spielen stellen Camerer und Hogarth (1999) im Rahmen einer Literaturanalyse vornehmlich keinen Einfluss von finanziellen Anreizen auf das durchschnittliche Verhalten der Probanden fest. In den Fällen, in denen sich das Verhalten dennoch durch Anreize ändert, findet oft eine Verschiebung hin zu weniger risikofreudigem Verhalten statt. Häufig angeführter Grund ist, dass die Dominanzeigenschaft in Laboruntersuchungen nicht erfüllt ist, sodass das leistungsabhängige Vergütungsmodell für Individuen keine ausreichende Anreizstruktur bietet. Konkret wird die Sinnhaftigkeit variabler Vergütungsmodelle durch eine unzureichende Kaufkraft der Marginalvergütung bei Mehrleistung untergraben (vgl. Harrison & List, 2004; Jenkins et al., 1998). Eine Beispielrechnung macht den Sachverhalt deutlich. Unter Berücksichtigung der finanziellen Ressourcenausstattung dieses Forschungsvorhabens, der üblichen Vergütung von Probanden, der Anzahl der notwendigen Experimente, der benötigten Anzahl von Teilnehmern je Experiment sowie des Zeitaufwandes je Teilnehmer ist eine erwartete Vergütung von 10 bis 12 Euro angemessen. Jeder Proband nimmt an 5 über im Schnitt 23.5 Runden ablaufende Interaktionen teil, dass durchschnittlich 117.5 Interaktionsentscheidungen je Proband getroffen werden müssen. Bei einer Vergütung von 12 Euro für eine durchschnittli-

che Punktzahl von 2.5 über alle Spiele[44] ergibt sich somit je Runde eine Marginalvergütung von 0.04 Euro je erzieltem Punkt für die Teilnehmer. Es wird nicht erwartet, dass eine derartige Vergütung sich als dominant in dem Sinne erweist, dass die Probanden (1) allein von einer durch die monetäre Vergütung induzierten Profitmaximierung angetrieben werden oder sich (2) die Interaktionsqualität aufgrund einer höheren Aufwandsbereitschaft signifikant ändert. Vielmehr ist es nicht unwahrscheinlich, dass Probanden trotz erfolgsabhängiger Vergütung zu einem hohen Maße von Motiven jenseits der Profitmaximierung beeinflusst werden. Dazu zählen zum Beispiel der Wunsch, sich angemessen zu verhalten, den Erwartungen des Experimentleiters gerecht zu werden, schlau zu wirken, eine gute Person zu sein, ein Gewinner zu sein und viele mehr. Insbesondere spielen Reputationseffekte in Bezug auf das Selbstbild der Teilnehmer selbst dann eine Rolle, wenn sie nicht öffentlich sind (vgl. Bodner & Prelec, 2008).

Die Größe des variablen Anreizes ist aufgrund begrenzter Ressourcen seitens der Experimentatoren häufig unzureichend, um die hypothetische Interaktionssituation des Labors in eine Entscheidungssituation mit für den Teilnehmer signifikanten monetären Nutzendifferenzen zu transformieren (vgl. Read, 2005). Infolgedessen ist die Dominanzeigenschaft nur schwer zu erfüllen, da beispielsweise die Beobachtung durch den Experimentleiter oder Reputationseffekte einen nicht zu vernachlässigenden Einfluss auf das Probandenverhalten haben, insbesondere, wenn diese den Probanden nicht viel kosten (vgl Bohm, 2002). Speziell im Rahmen wiederholter Markt- und Spielsituationen wird der Anreizeffekt variabler Vergütung dadurch gemindert, dass eine Lösung des Problems für die Probanden zu schwer ist, um durch monetäre Vergütung herbeigeführt zu werden (vgl. Camerer & Hogarth, 1999). Vielmehr liegen die Schattenkosten einer variablen Vergütung in einer geringeren Forschungsineffizienz, sodass eine derartige Methodik die Machbarkeit von Forschungsvorhaben einschränkt. Unter Berücksichtigung der vorgenannten Faktoren folgt diese Arbeit insofern Read (2005), dass eine machbare Studie auf Basis hypothetischer Interaktionen, die auf die Fähigkeit der Probanden zurückgreift, sich wie in einer realen Situation zu verhalten, der Nichtdurchführung derselben vorzuziehen ist. Damit deckt sich die Vorgehensweise mit den Ausführungen von Bardsley et al. (2020), der vor dem Hintergrund potentiell nachteiliger Sekundäreffekte sowie einem möglichen Ausbleiben des gewünschten Anreizeffekts eine reflektierte Verwendung variabler Vergütung in ökonomischen Experimenten vorschlägt. Gleichwohl eine leistungsabhängige Vergütung die externe Validität eines Experiment vermutlich nicht reduziert, ist ihr Wertbeitrag zur externen Validität gegeben der Tatsache, dass sie nicht Haupttreiber alltäglicher ökonomischer Interaktion ist, voraussichtlich als gering einzuschätzen (vgl. Loewenstein, 1999, S. F31-F32). Variable Vergütung soll die kognitive Anstrengung und die Motivation der Teilnehmer erhöhen. Diese Effekte können auch

[44] Maximal können 5 Punkte und minimal 0 Punkte je Runde erzielt werden. Der Mittelwert von 2.5 wird für die Beispielrechnung als Referenzpunktzahl zugrunde gelegt.

ohne monetäre Anreize erzielt werden. Eine grundsätzliche Notwendigkeit variabler Vergütung in ökonomischen Experimenten ist nicht gegeben (vgl. Read, 2005).

Jenseits der bestehenden Forschungsliteratur ist weiterhin ganz maßgeblich der anekdotische Erfahrungswert von empirischen Untersuchungen mit Studentengruppen, der nahelegt, dass eine variable Vergütung zu einer kooperativeren Spielweise führen kann. Grund ist, dass sich die Studierenden als häufig solidarische soziale Gruppe gegenseitig nur ungern verschlechtern wollen und daher eher sozialen Konsens anstreben.

Auf Basis der vorangegangenen Ausführungen wurde im Rahmen der Experimente auf eine fixe Vergütung zurückgegriffen. Grund ist, dass im Rahmen der begrenzten finanziellen Ressourcen der Experimentdurchführung nicht erwartet werden konnte, selbst bei vollständig variabilisierter Vergütung eine ausreichende Grenzvergütung im Verhältnis zur erforderlichen Mehrleistung bereitstellen zu können. Infolgedessen ist die Erfüllung der Dominanzeigenschaft nicht zu erwarten. Zur Herstellung einer adäquaten Anreizkulisse für die Teilnehmer wurde besonderer Fokus auf nicht-monetäre Motivatoren gelegt. Beispielweise wurde mitgeteilt, dass es elementar für den Erfolg des Experiments und des Dissertationsvorhabens ist, dass die Probanden bestrebt sind, eine möglichst hohe Punktzahl zu erzielen. Weiterhin wurde hervorgehoben, dass den Mitteilnehmern aufgrund der fixen Vergütungsstruktur kein finanzieller Nachteil entsteht, wenn eine nichtkooperative nur einseitig einträgliche Strategie verfolgt wird. Eine Auswertung des Fragebogens (siehe Abbildung A.2) bestärkt die Überlegungen dieses Kapitels. Anhand von Tabelle 4.3 wird deutlich, dass selbst ohne variable Vergütung 69% der Probanden anga-

Tabelle 4.3: Abfrageergebnis zur Zielsetzung im Rahmen einer Selbsteinschätzung der Probanden in Relation zu Klopfer (2018, S. 122). Mehrfachnennungen möglich. Quelle: Eigene Darstellung.

Datenbasis	Anzahl	Antwort I	Antwort II	Antwort III	Antwort IV	Antwort V
Prestudy II	48	71%	35%	6%	2%	–
Experiment I	42	67%	26%	7%	–	5%
Experiment II	46	61%	46%	4%	2%	–
Experiment III	43	79%	21%	–	–	–
Gesamt	179	69%	32%	4%	1%	1%
Klopfer (2018)	93	59%	34%	–	5%	1%

Antwort I	Ich habe versucht, meine eigene Punktzahl zu maximieren.
Antwort II	Ich habe versucht, die Gesamtpunktzahl aller Spieler zu maximieren.
Antwort III	Ich habe versucht, die Punktzahl meiner Mitspieler zu minimieren.
Antwort IV	Ich habe ein anderes Ziel verfolgt.
Antwort V	Keine Angabe.

ben, die Maximierung der eigenen Payoffs verfolgt zu haben. 32% der Probanden gaben an, die Maximierung der Payoffs aller Spieler angestrebt zu haben. Die Ergebnisse sind insbesondere in Relation zur Auswertung identischer Fragen bei Klopfer (2018, S. 122) beachtlich. Klopfer (2018) verwendet für das Experimentdesign im Rahmen der Induced Preference Theory eine *vollständig* variable Vergütung. Dennoch gaben hier ähnlicherweise 59% der Probanden an, die eigenen Payoffs maximiert zu haben, während der Anteil der Probanden, die angeben die Gesamtpayoffs maximiert zu haben ebenfalls bei vergleichbaren 34% liegt.

4.2 Umsetzung der Experimente

Das nachfolgende Kapitel beschreibt die operative Implementierung des Experimentdesigns, welche sich aufgrund tiefgreifender Parallelen im Anforderungsprofil maßgeblich an Klopfer (2018) orientiert. Dabei werden folgende Aspekte aufgegriffen:

1. **Labortechnologie:** Wie werden die Experimente informationstechnisch abgebildet (siehe Kapitel 4.2.1)?
2. **Aufbau- und Ablauforganisation:** Welche organisatorische Aspekte müssen für einen sauberen Ablauf berücksichtigt werden (siehe Kapitel 4.2.2)?
3. **Terminübersicht:** Welche Termine wurden zur Bewertung prozessualer und funktionaler Aspekte angesetzt (siehe Kapitel 4.2.3)?

4.2.1 Labortechnologie

Die Datenerhebung erfolgt unter Verwendung der Experimentplattform *BMind*, welche im Rahmen einer Abschlussarbeit am Institut für Unternehmensführung am Karlsruher Institut für Technologie von Kabakcha (2017) entwickelt wurde. Serverseitig greift BMind auf das *Django REST Framework* in einer *Python* Programmierumgebung zurück. Klientenseitig interagieren menschliche Spieler mittels einer in *Java* programmierten Applikation über Android-basierte Tablets. Anfallende Daten werden in einer *PostgreSQL* Datenbank gespeichert (Kabakcha, 2017). Der Markovspieler *AgentM* ist ebenfalls in Python umgesetzt und wird auf einem zentralen Experimentrechner ausgeführt, von wo aus er mit BMind kommuniziert.

Abbildung 4.4 stellt die für die experimentelle Erhebung gewählte technologische Infrastruktur dar. BMind wurde originär lediglich für die Interaktion zwischen menschlichen Spielern (siehe Spiel A, Abbildung 4.4) konzipiert. Die Einbindung des in *Python* programmierten Markovagenten erfolgt daher über einen Klientenrechner. Auf dem Rechner wird, sobald sich ein menschlicher Spieler in ein dementsprechend gekennzeichnetes Spiel einwählt, automatisch eine separate Instanz des Markovagenten gestartet. Über den Klientenrechner fragt diese den

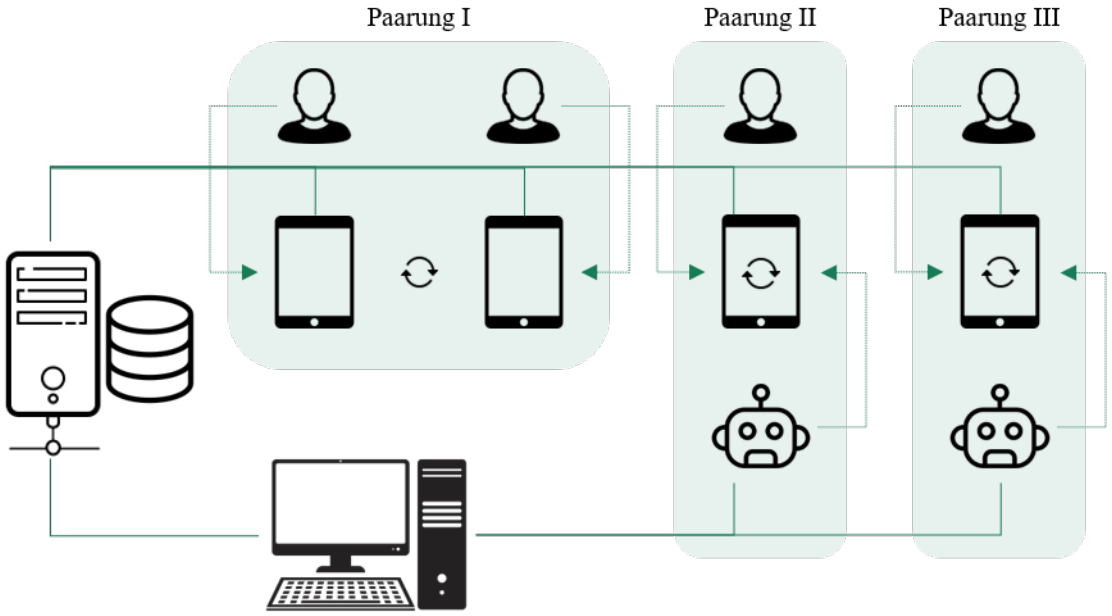


Abbildung 4.4: Schematische Darstellung der Spielerpaarung zwischen menschlichen Spielern (Paarung I) und Markovspielern und menschlichen Spielern (Paarung II und III). Quelle: Eigene Darstellung.

aktuellen Spielstand ab und gibt eigene Aktionen weiter (siehe Spiel B und C, Abbildung 4.4). Dem Markovagenten stehen hierbei die identischen Informationen wie dem Gegenspieler zur Verfügung.

Die nachfolgend beschriebene Methodik der Experimente wurde in der Experimentplattform BMind umgesetzt und lässt sich wie folgt charakterisieren (vgl. Klopfer, 2018):

- **Spiel, Spieler und Aktionsraum:** Je nach Experimentsitzung wurde die entsprechende Auszahlungsstruktur der zwei Spieler mit je zwei Aktionen in Form eines 2x2 Spiels hinterlegt (siehe Kapitel 4.1.1).
- **Spielerpaarung:** Spieler werden einander bei der Initialisierung eines Spiels anhand des festgelegten Paarungs-Mechanismus zugewiesen (siehe Kapitel 4.1.2).
- **Zugfolge:** Spieler ziehen simultan.
- **Auszahlungen:** Die Payoffs werden rundenweise kumuliert.
- **Spielende:** Abbruch erfolgt nach der vorab zufällig bestimmten Rundenzahl (siehe Kapitel 4.1.3).
- **Informationen:** Spieler sehen alle Spieler und deren Aktionen sowie die vollständige Auszahlungsmatrix. Es findet keine Anzeige der (kumulierten) vergangenen Auszahlungen oder der Rundenzahl statt (siehe Abbildung 4.5 und Tabelle 4.4).

Die Klienten-Applikation wurde in den Experimenten auf Amazon Fire Tablets ausgeführt. Initial zeigt der Startbildschirm den Probanden eine Auswahl der verfügbaren Interaktionen in Form von nummerierten Symbolen. Auf Anweisung des Experimentleiters wählen die Probanden das jeweilige Spiel aus. In Folge werden die Probanden vor dem Spielstart entsprechend dem in Kapitel 4.1.2 beschriebenen Verfahren mit einem Gegner gepaart und entweder der Rolle des Zeilen- oder Spaltenspielers zugewiesen.

Nach der Initialisierung des Spiels zeigen die Tablets den Spielern ihre eignen Handlungsalternativen und Payoffs in der oberen und die des Gegners in der unteren Bildschirmhälfte (siehe Abbildung 4.5). In jeder Zugfolge eines Spiels wählt der Spieler seine gewünschte Aktion durch Berühren des jeweiligen Kästchens. Nach Aktionswahl durch beide Spieler werden deren Handlungen aufgedeckt und die realisierten Auszahlungen durch einen orangefarbenen Hintergrund hervorgehoben. Dieser Zyklus wird bis zum Erreichen der zufälligen Rundenzahl wiederholt, nach welcher das Spiel durch die Meldung *"Game finished!"* beendet wird und der Spieler zum Startbildschirm zurückkehrt und bereit für das nächste Spiel ist.

4.2.2 Aufbau- und Ablauforganisation

Nachfolgend findet eine Darstellung der Ablauf- und Aufbauorganisation der Experimente statt. Ein robuster Hergang ist für die Gewinnung belastbarer empirischer Daten unabdingbar. Insbe-

Abbildung 4.5: Probandenseitige Spieloberfläche der Tablets im Verlauf einer Zugfolge im exemplarischen wiederholten Prisoner's Dilemma aus Sicht von Spieler A (mit ergänzten Kommentarfeldern in roter Farbe). Quelle: Eigene Darstellung.

Tabelle 4.4: Informationsausstattung der Experimentteilnehmer über die gesamte Experimentsitzung, die einzelnen Interaktionen bzw. wiederholten Spiele innerhalb der Sitzung und die einzelnen Runden innerhalb der Interaktionen. Quelle: Eigene Darstellung in Anlehnung an Chinczewski (2019).

Information	Verfügbarkeit	Quelle
Experimentsitzung		
- Experimentablauf	öffentlich	Instruktionen
- Spielregeln	öffentlich	Instruktionen
- Zielfunktion	öffentlich	Instruktionen
- Zielfunktion	öffentlich	Instruktionen
- Gesamtzahl der Interaktionen	öffentlich	Instruktionen
- Präsenz nichtmenschlicher Gegner	unbekannt	Instruktionen
Spezifische Interaktion		
- Realisierbare Aktionen	öffentlich	Spielmatrix, Instruktionen
- Realisierbare Auszahlungen	öffentlich	Spielmatrix, Instruktionen
- Verbleibende Anzahl von Interaktionen	implizit	Experimentverlauf
- Gesamte Rundenzahl	unbekannt	–
- Gegneridentität	unbekannt	–
Spezifische Runde		
- Aktionshistorie	implizit	Interaktionsverlauf
- (Kumulierte) Auszahlungshistorie	implizit	Interaktionsverlauf
- Gegnerstrategie	implizit	Interaktionsverlauf
- Verbleibende Rundenzahl	unbekannt	–

sondere wird hierbei dem Ziel Rechnung getragen, dass alle Spieler über dieselbe Informationsausstattung verfügen (vgl. D. Friedman & Sunder, 1994). Tabelle 4.4 bietet einen Überblick über die Informationsverfügbarkeit. Besonders hervorgehoben sei, dass die Probanden zu keinem Zeitpunkt Informationen darüber verfügen, dass nichtmenschliche Gegner Teil der Gegnermenge sind. Weiterhin ist nicht bekannt, gegen welchen spezifischen Gegner die einzelnen Spiele stattfinden.[45]

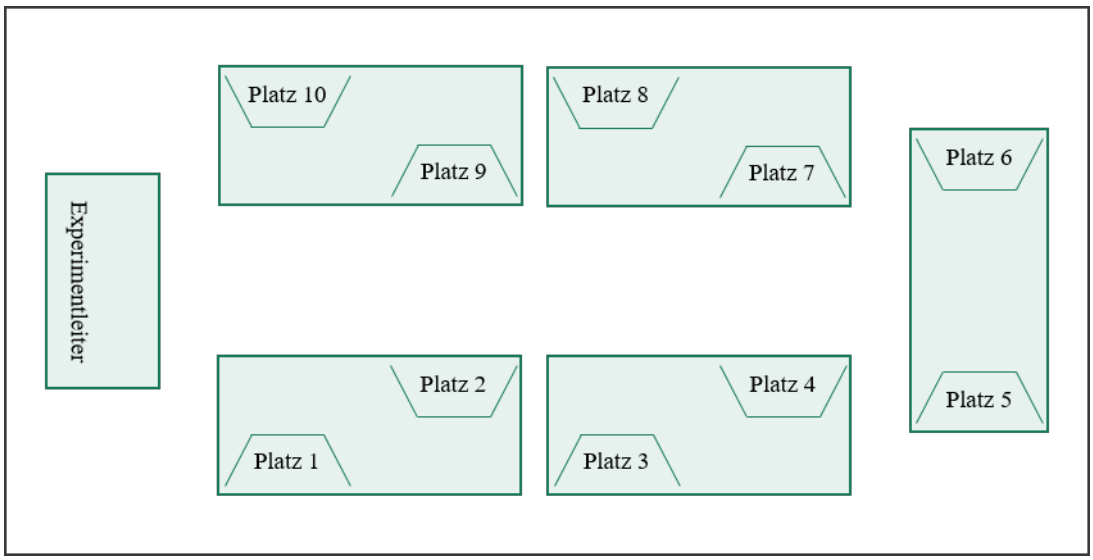


Abbildung 4.6: Räumlicher Aufbau der Laborumgebung mit Sichtschutz. Quelle: Eigene Darstellung.

4.2.2.1 Aufbauorganisation: Laborumgebung

Das Labor ist so aufgebaut, dass Teilnehmer möglichst wenig Informationen über andere Teilnehmer erlangen können (siehe Abbildung 4.6). Die Spieler werden mit großzügigem Abstand im Raum verteilt. Darüber hinaus ist an jedem Platz ein Sichtschutz angebracht, sodass keine Interaktion zwischen den Probanden durch nonverbale oder verbale Kommunikation stattfinden kann. Einblicke in das Spiel, den Spielverlauf, das (Aktions-)Verhalten oder die Gestik und Mimik der anderen Probanden werden so verhindert. Der Experimentleiter stellt darüber hinaus mit Nachdruck sicher, dass jegliche Kommunikationsversuche vor, während oder zwischen den Spieldurchläufen sofort unterbunden werden. Abbildung A.4 im Anhang dokumentiert die Aufbauorganisation ergänzend.

Gleichwohl wird durch das Verhindern von Kommunikation die Unkenntnis über Gegenspieler gewahrt, sodass sich keine Carry-Over-Effekte aus Erfahrungen mit identifizierbaren Gegnern ergeben können (vgl. Charness et al., 2012). Insbesondere wird so vermieden, dass Informationen einzelner Probanden über andere Teilnehmer, zum Beispiel in Form von Persönlichkeitsbildern, Einfluss auf die eigene Aktionswahl hat. Letzteres ist aufgrund der guten Vernetzung und Bekanntschaft von studentischen Probanden insbesondere hervorzuheben.

4.2.2.2 Ablauforganisation: Experimenthergang

Im Vorfeld einer jeden Experimentsitzung wurde den Teilnehmern zufällig ein Laborplatz (siehe Abbildung 4.6) zugewiesen. Darüber hinaus wurde für jeden Spieler für jedes der Spiele

[45] Es wurden zu keine falschen oder irreführenden Informationen übermittelt. In Tabelle 4.4 als *unbekannt* gekennzeichnete Informationen wurden lediglich ausgelassen. Die naheliegendste implizite Annahme bezüglich der Gegner für die Probanden ist, dass letztere gegen andere Probanden spielen.

eine zufällige Paarung mit einem Gegnerspieler entsprechend Kapitel 4.1.2 vorgenommen. Die Informationen wurden in einer zentralen Datei gespeichert, zu denen die Probanden zu keiner Zeit Zugang hatten.

Die Durchführung einer Experimentsitzung dauert maximal 75 Minuten und lässt sich anhand der folgenden Phasen mit ihrer zeitlichen Obergrenze charakterisieren (vgl. Klopfer, 2018):

1. **Platzzuweisung** (10 Minuten): Bei Ankunft wird den Probanden nach Vorlage eines Identitätsnachweises der festgelegte Laborplatz zugewiesen. Um eine möglichst homogene Informationsausstattung zu gewährleisten, sind den Teilnehmern zu diesem Zeitpunkt keine Informationen zum Experiment bekannt und jegliche Kommunikation wird durch den Experimentleiter unterbunden (vgl. D. Friedman & Sunder, 1994, S. 75).

2. **Instruktionen** (10 Minuten): Nach Platznahme durch alle Teilnehmer findet eine kombinierte Instruktion der Probanden in Schrift und dann in Wort statt (vgl. D. Friedman & Sunder, 1994, S.77). Zunächst werden dazu schriftliche Instruktionen ausgeteilt, für die eine stille Einlesezeit von fünf Minuten gewährt wird. Danach wird die Instruktionsunterlage durch den Experimentleiter verbal und visuell anhand der selben Instruktionsunterlage vorgenommen. Die Instruktionen sind in Abbildung A.1 im Anhang zu finden. Sie enthalten Informationen zu Verhaltensregeln und Kommunikation, Vergütung, Experimentablauf, Zielfunktion sowie dem technischen Ablauf eines Spiels in der Experimentplattform.

3. **Klärung von Fragen** (5 Minuten): Der Experimentleiter bittet die Teilnehmer offene Fragen zu stellen. Diese werden öffentlich beantwortet, um keinen privaten Informationsvorsprung zu erzeugen (vgl. D. Friedman & Sunder, 1994, S. 77). Fragen, die über die Bedienung der Experimentplattform sowie den Experimentablauf hinausgehen, werden nicht beantwortet. Insbesondere wurden keine Informationen zu dem Untersuchungsgegenstand des Experiments oder Verhaltensempfehlungen für die Spiele gegeben.

4. **Spieldurchführung** (50 Minuten): Es folgt die Durchführung der Spieldurchläufe. Zu Beginn jedes Spiels konfiguriert der Experimentleiter die Probanden-Tablets verdeckt entsprechend der nach Kapitel 4.1.2 festgelegten Paarung. Die Identität des Gegners ist den Teilnehmern nicht bekannt. Weiterhin wird zu Beginn eines jeden Spiels automatisch die in Kapitel 4.1.3 beschriebene, den Probanden ebenfalls unbekannte, Rundenzahl ermittelt und hinterlegt. Die Tablets werden ausgeteilt. Auf Hinweis des Experimentleiters wird das Spiel für die Probanden freigegeben. Die Teilnehmer spielen nun in simultaner Zugfolge bis die randomisierte Rundenzahl erreicht ist und die Teilnehmer über das Spielende informiert werden.

5. **Ausfüllen des Fragebogens** (5 Minuten): Nach Abschluss der Spieldurchläufe werden die Probanden aufgefordert einen Fragebogen zur Erhebung sozioökonomischer Informationen und Abfrage von Vorwissen und der während des Spielens angewandten Entscheidungslogik auszufüllen (siehe Abbildungen A.2 und A.3 im Anhang). Die (nicht) getätigten Angaben beeinflussen die Vergütung explizit nicht. Eine Incentivierung der Auswertung findet nicht statt, da das tatsächliche Spielverhalten im Vordergrund dieser Untersuchung steht und verfügbare Ressourcen dementsprechend auf das tatsächliche Spielen verwendet werden. Die Validität der gewonnen Daten ist folglich potentiell beeinträchtigt (vgl. R. Croson, 2005, S. 293). Jedoch hätte eine feste Vergütung für die Teilnahme am Fragebogen Falschangaben ebenso wenig verhindern können (vgl. Dingelstedt, 2015). Vorbehaltlich der genannten Limitationen kann der Datensatz dazu genutzt werden, Analysen über den konkreten Spielverlauf hinaus durchzuführen. So kann sichergestellt werden, dass die Probanden die Spielsituation tatsächlich verstanden haben.

6. **Auszahlung** (5 Minuten): Nachdem alle Teilnehmer den Fragebogen ausgefüllt haben, wird dieser einzeln am Platz des Experimentleiters abgegeben, wo die Probanden ihre Auszahlung (siehe Kapitel 4.1.4) empfangen und schließlich das Labor verlassen.

4.2.3 Terminübersicht

Aufgrund der kreativen Unsicherheit hinsichtlich der Leistungsfähigkeit des Markovagenten und den beschränkten Ressourcen der Datenerhebung wurde ein stufenweiser Zugang zur Validierung entsprechend Tabelle 4.5 gewählt. Die Termine gliedern sich nach Art wie folgt:

- **Ablauftest:** Unter Verzicht auf Laborbedingungen wurden diverse Ablauftests durchgeführt. Der Fokus lag hierbei auf der Stabilität der Labortechnologie, dem Verständnis

Tabelle 4.5: Übersicht der Erhebungstermine. Quelle: Eigene Darstellung.

	Termin	Zielsetzung	Agentenversion	Spiel (wiederholt)
Ablauftest I	29.06.2018	Prozessuale Validierung	AgentM-vα	Prisoner's Dilemma
Ablauftest II	05.12.2018	Prozessuale Validierung	AgentM-vα	Prisoner's Dilemma
Prestudy I	11.01.2019	Funktionaler Pretest	AgentM-vβ	Prisoner's Dilemma
	14.01.2019	Funktionaler Pretest	AgentM-vβ	Prisoner's Dilemma
Prestudy II	17.04.2019	Funktionaler Pretest	AgentM	Prisoner's Dilemma
Experiment I	16.07.2019	Funktionale Validierung	AgentM	Chicken Game
Experiment II	22.07.2019	Funktionale Validierung	AgentM	Hero Game
Experiment III	24.10.2019	Funktionale Validierung	AgentM	Prisoner's Dilemma

der Spielsituation durch die Teilnehmer und der Funktionsweise der Markovagenten. Die Stabilität der Labortechnologie wurde im Sinne der verwendeten Experimentplattform, des hinterlegten Paarungsmechanismus und des Spielabbruchs erprobt. Das Verständnis der Spielsituation durch die Teilnehmer wurde im Rahmen von Befragungen und Dialogrunden im Anschluss an die Spieldurchläufe bestätigt. Die korrekte Funktionsweise der Agenten wurde nachfolgend anhand von Analysen einzelner Spielverläufe sichergestellt. Die Durchführung fand als inhaltlicher Bestandteil von Seminarveranstaltungen des Lehrstuhls für Unternehmensführung statt, sodass keine explizite Vergütung stattfand.

- **Pre-Test:** Im Anschluss an die Validierung der Experimentgestaltung folgen Funktionstests, die eine erste Indikation über die grundsätzliche Leistungsfähigkeit der Bots unter Laborbedingungen geben sollen. Im Rahmen des ersten Funktionstests wurden darüber hinaus Verbesserungspotentiale in Bezug auf die Gestaltung des Agenten identifiziert. Nach Implementierung dieser Anpassungen wurde ein erneuter Funktionstest durchgeführt, um die antizipierte Leistungsverbesserung zu bestätigen. Kapitel B stellt diesen Prozess zusammenfassend dar.

- **Experiment:** Nach einer ersten funktionalen Untersuchung des Markovagenten im Rahmen von Pre-Tests wurden drei Experimente unter Laborbedingungen für die in Kapitel 4.1.1 ausgewählten Spiele durchgeführt.

4.3 Selektion und Deskription der Teilnehmer

Die Qualität der Teilnehmerbasis stellt das Fundament dar, welches das Sammeln belastbarer Daten ermöglicht (vgl. z.B. Greiner, 2015, S. 114-115). Gemäß dieser Maxime beschäftigt sich dieses Kapitel mit folgenden Schwerpunkten:

1. **Beschreibung der Teilnehmer:** Wie gestaltet sich der Auswahlprozess der Teilnehmer (siehe Kapitel 4.3.1)?

2. **Auswahl und Einladung der Teilnehmer:** Wie lassen sich die Experimentteilnehmer charakterisieren (siehe Kapitel 4.3.2)?

4.3.1 Auswahl und Einladung der Teilnehmer

Dieses Kapitel beschreibt den Auswahl- und Einladungsprozess der Probanden. Dabei findet eine Beschränkung auf die Termine ab Prestudy I statt. Die vorgelagerten Termine dienen wie zuvor beschrieben primär der Sicherung robuster Experimentalprozesse statt der Datenerhebung.

Tabelle 4.6: Zusammenfassung der Teilnehmerrekrutierung. Quelle: Eigene Darstellung.

		Anzahl der Probanden		
	Plattform	Eingeladen	Registriert	Erschienen
Prestudy I	ILIAS	254	71	63
Prestudy II	KD2Lab	1,113	50	48
Experiment I	KD2Lab	1,118	50	43
Experiment II	KD2Lab	1,150	50	45
Experiment III	KD2Lab	1,462	50	43
	∑	5.097	271	242

Für die erste experimentelle Validierung im Rahmen von Prestudy I wurde aufgrund des Prototypenstadiums des Markovagenten eine ressourcenschonende Teilnehmerquelle angestrebt. Dementsprechend fand die Rekrutierung der Probanden aus den Teilnehmern der Lehrveranstaltung *Organisationsmanagement* des Instituts für Unternehmensführung am KIT statt. Die Teilnahme am Experiment erfolgte im Rahmen des Lehrinhaltes der Veranstaltung, weshalb jeder Sitzung eine offene Diskussions- und Reflexionsrunde mit den Studierenden bezüglich der Implikationen auf Management- und Organisationsthemen folgte. Die Organisation der ersten Prestudy sowie die Einladung der Teilnehmer wurde über die Lernplattform-Software *ILIAS* durchgeführt, da alle Organisationsmanagement-Kursteilnehmer dort registriert sind. Es handelt sich um eine open-source Lernplattform, die KIT-weit zur Kursverwaltung eingesetzt wird.[46] Eine monetäre Vergütung war aufgrund der Einbettung im Lehrbetrieb nicht möglich. Stattdessen wurde den Teilnehmern eine festen Größe von Bonuspunkten gutgeschrieben.

Nach der ersten Validierung der Betaversion des Markovagenten in Prestudy I, wurde für die Organisation und Teilnehmerauswahl der nachfolgenden Experimente ein Wechsel zur Experimentalverwaltungs-Software des *KD2LAbs* vorgenommen. Es handelt sich bei dem KD2Lab um ein durch die DFG-gefördertes, institutsübergreifendes Experimentallabor des KITs, welches auf der Verwaltungs-Software *HROOT* von Bock et al. (2014) fußt.[47]

Über den Zeitraum der Experimente waren circa 3,500 Individuen im Teilnehmerpool des KD2Labs registriert. Diese setzen sich hauptsächlich, jedoch nicht ausschließlich aus Studierenden zusammen. An diese Grundgesamtheit wurde nach Anwendung der folgenden Kriterien eine elektronische Einladung versendet:

- **Sprache:** Es wurden nur Teilnehmer eingeladen, die laut eigener Angabe der Deutschen Sprache mächtig sind. Nur so kann sichergestellt werden, dass die Instruktionen durch

[46] Siehe https://www.ilias.de/.
[47] Siehe http://www.kd2lab.kit.edu/.

alle Probanden nachvollzogen werden und der Experimentablauf möglichst reibungsarm verläuft.

- **Allgemeine Erfahrung:** Es wurden nur Teilnehmer eingeladen, die bereits an mindestens einem Experiment des KD2Labs teilgenommen hatten. Hierdurch wird sichergestellt, dass Probanden sich bereits mit den grundsätzlichen Abläufen eines wirtschaftswissenschaftlichen Experiments vertraut machen konnten. Weiterhin findet insofern eine self-selection statt, als dass die Probanden nach dem vorangegangenen Experiment weiterhin für die Teilnahme an derartigen Veranstaltungen motiviert sind.

- **Zuverlässigkeit:** Es wurden nur Teilnehmer eingeladen, die bisher noch nie nach Registrierung für ein Experiment über das KD2Lab unentschuldigt an dem Termin gefehlt hatten. Durch das Herausfiltern augenscheinlich unzuverlässiger Probanden wird die Stabilität des Experimentablaufs weiter erhöht. Für Experiment III wurde die Grenze aufgrund initial geringer Rückmeldung auf einen erlaubten Fehltermin erhöht, um eine ausreichend große Zielgruppe ansprechen zu können.

- **Spezifische Erfahrung:** Es wurden nur solche Teilnehmer eingeladen, die noch nicht an einem der vorangegangenen Experimente dieser Arbeit des KD2Labs teilgenommen hatten. Teilnehmer von Prestudy II und Experiment I erhielten beispielsweise keine Einladung für Experimente II - III. Ziel ist es, etwaige Lerneffekte oder andere Beeinflussungsfaktoren aus vorigen Sitzungen auf das Spielverhalten möglichst auszuschließen.

Unter Anwendung der beschriebenen Vorgehensweise wurden gemäß Tabelle 4.6 für die fünf Experimente insgesamt 5,097 Einladungen ausgesprochen. Als Reaktion registrierten sich insgesamt 271 Teilnehmer, wovon 242 tatsächlich zu den Terminen erschienen. Davon entfallen 63 erschienene Probanden auf Prestudy I,[48] während zur Prestudy II 48 Teilnehmer erschienen. An den Experimenten I, II und III nahmen jeweils 43, 45 und 43 Probanden teil.

Ein nachteiliger Einfluss durch die Selektion primär studentischer Probanden wird als gering eingeschätzt. Forschung zum Verhaltensvergleich von studentischen und nicht-studentischen Experimentteilnehmern (vgl. z.B. Belot et al., 2015; Chan et al., 2011; Frechette, 2015) konnte keine systematische Differenz zwischen den beiden Teilnehmergruppen nachweisen (vgl. Davis & Holt, 1993; vgl. Holt, 1995, S. 6). Darüber hinaus ergeben sich aus dem Rückgriff auf studentische Probanden für universitäre Forschung organisatorische Vorteile. Beispielsweise sind Studierende aufgrund der organisationalen und räumlichen Nähe vergleichsweise leicht zu rekrutieren. Weiterhin ist es aufgrund der geringen Opportunitätskosten möglich, zahlreiche

[48] Aufgrund der Einbettung im Lehrbetrieb wurde die Experimentteilnahme jedem Kursteilnehmer zur Vermeidung von Diskriminierung offen gestellt. Hierdurch ergibt sich die ungleichmäßige Verteilung der registrierten Probanden in Tabelle 4.6.

Teilnehmer trotz eines beschränkten Budgets zu gewinnen. Vorteilhaft ist außerdem die Auffassungsgabe von Studierenden, die Voraussetzungen für die Umsetzung umfangreicher Experimente schafft (vgl. D. Friedman & Sunder, 1994, S. 39-40; vgl. Klopfer, 2018).

4.3.2 Beschreibung der Teilnehmer

Tabelle 4.7 gibt Aufschluss auf die Teilnehmerstruktur der Termine. Die ungleichmäßige Verteilung einzelner Parameter wie des Geschlechts der Prestudy-Teilnehmer ergibt sich primär aus der Beschaffenheit der Studierendenschaft des KIT, beziehungsweise der Teilnehmergesamtheit des KD2Labs (vgl. auch Klopfer, 2018). Infolgedessen fällt der Anteil der Abiturienten und Bacheloranden verhältnismäßig hoch aus. Auch der Anteil der Probanden mit wirtschafts- und ingenieurwissenschaftlichem Hintergrund fällt ausgesprochen hoch aus. Als direkte Konsequenz ist das Niveau der spieltheoretischen Vorkenntnisse mit nur 10% der Experimentteilnehmer ohne Vorwissen hoch. Klar ist weiterhin, dass die über die Experimentplattform KD2Lab rekrutierten Teilnehmer über eine gewisse Erfahrung mit spieltheoretischen Experimenten aus anderen Untersuchungen aufweisen.

Folgerichtig ist die Teilnehmerbasis der empirischen Untersuchungen dieser Arbeit nicht repräsentativ für die Gesamtbevölkerung, sodass die Ergebnisse der statistischen Auswertung lediglich als erster Anhaltspunkt zu verstehen sind. Um Aussagen jenseits der Spezifika der vorliegenden Teilnehmerstruktur generalisieren zu können, sind Folgeuntersuchungen mit einer allgemeingültigeren Teilnehmerpopulation erforderlich (vgl. auch Klopfer, 2018).

Tabelle 4.7: Teilnehmerstruktur auf Basis des demographischen Fragebogens. Quelle: Eigene Darstellung in Anlehnung an Klopfer (2018).

		Prestudy		Experiment			
Parameter	Wert	I	II	I	II	III	Ges.
Geschlecht	männlich	69%	69%	57%	46%	63%	55%
	weiblich	31%	31%	43%	54%	37%	45%
Bildungsgrad	Abitur	84%	56%	52%	61%	53%	56%
	Ausbildung	3%	0%	0%	0%	2%	1%
	Bachelor	11%	35%	43%	22%	40%	34%
	Master/Diplom	2%	4%	5%	17%	5%	8%
	Promotion	0%	0%	0%	0%	0%	0%
	Andere	0%	4%	0%	0%	0%	0%
Fachrichtung	Wirtschaftswiss.	82%	46%	56%	43%	42%	47%
	Ingenieurswiss.	15%	31%	22%	33%	35%	30%
	Naturwiss.	2%	10%	10%	13%	9%	11%
	Sozial- & Geisteswiss.	2%	6%	7%	9%	12%	9%
	Andere	0%	4%	5%	2%	2%	3%
	Keine Angabe	0%	2%	0%	0%	0%	0%
Einschätzung spieltheoretischer Vorkennnisse	Keine	3%	15%	14%	9%	7%	10%
	Gering	13%	31%	26%	22%	30%	26%
	Grundlegend	53%	25%	29%	54%	42%	42%
	Erweitert	24%	27%	21%	15%	19%	18%
	Sehr gut	6%	2%	5%	0%	2%	2%
	Keine Angabe	0%	0%	5%	0%	0%	2%
Anzahl bisheriger spieltheoretischer Experimente	Keine	53%	15%	20%	2%	30%	17%
	1	15%	13%	18%	7%	23%	16%
	2	10%	13%	28%	22%	12%	20%
	3	8%	15%	13%	11%	7%	10%
	$4-9$	14%	40%	20%	54%	23%	33%
	>10	0%	4%	3%	4%	5%	4%

4.4 Gestaltung der Datenauswertung

Die Beurteilung der Spieler-Performance erfolgt analog zur Zielfunktion der Payoffmaximierung anhand eines Vergleichs der realisierten Auszahlung aller Spielertypen gegen menschliche Sparringspieler. Die Verwendung des realisierten Auszahlungsdurchschnitts ist hierbei repräsentativ für die Nutzenfunktion des wiederholten Spiels (vgl. Shoham & Powers, 2014a, S. 4) Um die Interpretierbarkeit der beobachteten Auszahlungswerte zu vereinfachen und möglichst intuitiv zu gestalten, wird zunächst eine normierter Payoffkenngröße $\bar{\rho}^i$ je Spieler definiert. Es handelt sich dabei um den den durchschnittlichen $\bar{r}^i$ Payoff über T Runden, welcher auf Basis des höchstmöglichen Auszahlungswertes r^i_{max} und des geringstmöglichen Auszahlungswertes r^i_{min} der Spielmatrix normiert wird:

$$\begin{aligned} \bar{\rho}^i &= 100\frac{\bar{r}^i - r^i_{min}}{r^i_{max} - r^i_{min}} \\ \bar{r}^i &= \frac{\sum_{t=1}^{T} r^i_t}{T} \\ r^i_{max} &= max(R^i) \\ r^i_{min} &= min(R^i) \end{aligned} \tag{4.2}$$

Die Interpretation der Ergebnisse wird so von den spezifischen Auszahlungswerten der Spiele und der Spiellänge gelöst. Aufgrund von $\bar{\rho}^i \in [0,100]$ kann der normierte Payoff als Prozentwert der Spanne zwischen dem minimal und maximal möglichen Durchschnitts-Payoffs eines Spiels verstanden werden.

Es wird weiterhin ausschließlich die kleinste gemeinsame Rundenzahl für alle betrachteten Spielverläufe betrachtet. Hierdurch werden mögliche Varianzeffekte der Rundenzahl über verschiedene Spiele hinweg eliminiert. Gemäß der Abbruchbedingung in Kapitel 4.1.3 ist die kleinste gemeinsame Rundenzahl als 21 definiert, sodass bei der nachfolgenden Auswertung stets $T = T_{min} = 21$ gilt.

4.5 Probelauf zur Validierung der Laborbedingungen

Aufgrund des neu entwickelten und insbesondere neu in die technologische Infrastruktur der Experimentplattform BMind eingebundenen Software des Markovagenten (siehe Kapitel 4.2.1) wurde vorab ein technischer Probelauf durchgeführt, um einen robusten Experimentablauf zu gewährleisten (vgl. D. Friedman & Sunder, 1994, S. 74-75).

Ablauftest I und II fanden mit Teilnehmern des Moduls *Strategie und Management: Fortgeschrittene Themen* am Institut für Unternehmensführung des KITs statt (siehe Tabelle 4.5). Die Probanden besitzen einen wirtschaftswissenschaftlichen Studienhintergrund und befanden sich

im Bachelor- und Masterstudium. Um den Fokus auf prozessuale Fragen zu legen, wurde das wiederholte Prisoner's Dilemma aufgrund seiner Bekanntheit für den Test ausgewählt. Design und Ablauf des Tests orientierten sich an den Überlegungen in Kapiteln 4.1 und 4.2.2.2, finden jedoch nicht unter Laborbedingungen statt. Jeder Proband spielte in zufälliger Reihenfolge gegen den die Alphaversion des Markovagenten, gegen einen Tit-for-Tat-Spieler und gegen zwei andere Menschen. Auf monetäre Vergütung und Anreize wurde aufgrund rechtlicher Auflagen bezüglich des Lehrbetriebs verzichtet. Die Probeläufe bestätigen die Verständlichkeit der Instruktionen und der Klienten-App für menschliche Spieler. Weiterhin machte er Schwächen in der Anbindung des Markovagent-Klienten an den Server sowie eine technisch unzureichende Netzwerkstabilität deutlich. In Folge kam es zu teilweise vorzeitigen oder verzögerten Spielabbrüchen sowie fehlenden Spielaufzeichnungen. Die genannten Probleme konnten im Anschluss durch eine verbesserte Netzwerkinfrastruktur sowie eine robustere Einbindung des Markovagent-Klienten und eine sicherheitsbedingt redundante Spielaufzeichnung gelöst werden.

5 Empirische Validierung des Markovagenten

Das sich anschließende Kapitel befasst sich mit der statistischen Auswertung der empirischen Daten zur Vollversion des Markovagenten aus Prestudy II und Experimenten I bis III (siehe Tabelle 4.5). Die Analyseergebnisse zur Betaversion des Markovagenten aus Prestudy I finden sich im Anhang unter Kapitel C wieder. Im Zentrum der Betrachtung steht die Leistung der Markovagenten im Vergleich zu der von menschlichen Spielern. Die relative Betrachtung des Leistungsbegriffes ist ein zentrales Element der Analyse. Grund ist, dass es keinen objektiven Leistungsbenchmark für die Auszahlungsleistung eines einzelnen Spielers im wiederholten Spiel gibt.[49] Insbesondere hängt das überhaupt erzielbare Ergebnis maßgeblich von der Grundgesamtheit der Gegenspieler ab. Eine absolute Aussage hinsichtlich eines *hohen* oder *niedrigen* Payoffs ist nicht möglich. Die Bewertung kann stets nur relativ zu einer Menge an Vergleichsspielern getroffen werden (vgl. Axelrod, 1984). Im Rahmen dieser Arbeit wurden menschliche Spieler im Sinne eines natürlichen Urtypus eines lernenden Agenten als Bezugsrahmen ausgewählt.

Zusammenfassende Lesehilfe der empirischen Untersuchung

Die sich anschließende empirische Auswertung ist aus Gründen der methodischen Stringenz und Vollständigkeit umfassend gestaltet und folglich von wechselndem Erkenntnisgewinn geprägt. Kapitel 5.1 zeigt dementsprechend lediglich, dass es es keine initialen Kontraindikationen für die Verwendung des Markovagenten im antizpierten Anwendungskotenxt gibt. Erst Kapitel 5.2 widmet sich dem Kern der emprischen Untersuchung hinsichtlich eines umfassenden Leistungsbenchmarks. Innerhalb der Kapitel gibt die *deskriptive Datenauswertung* einen abstrahierenden Überblick über die aufgezeichneten Spielverläufe. Erste Leistungsvergleiche werden anhand von *Hypothesentests* generiert, wobei deren Ergebnisse aufgrund einer Nichtberücksichtigung von Lerneffekten begrenzt sind. Als gehaltvoller präsentieren sich die *Regressionsanalysen* auf Basis einer Panelstruktur. Diese zeigen eine über verschiedene Spiele robuste signifikante Überlegenheit des Markovagenten über menschliche Vergleichsspieler (siehe Kapitel 5.2.4).

[49] Die Auszahlung des Spielers mit der des Gegners zu vergleichen ist keine Option, da dies *nicht* der Zielfunktion entspricht, die eigene Auszahlung zu maximieren.

Die Ergebnisauswertung gliedert sich zunächst auf eine erste Validierung der Vollversion des Markovagenten in einer eingeschränkten Prestudy II.[50] Im Anschluss findet eine Betrachtung der weiterführenden Untersuchung des Markovagenten anhand der vollumfänglichen Experimente statt. Der Aufbau beider hier beschriebenen Unterkapitel gliedert sich anhand der folgenden Leitfragen:

- **Experimenthergang:** Welche organisatorischen Eigenschaften zeichnen den Erhebungstermin aus?
- **Deskriptive Datenauswertung:** Welche Rückschlüsse auf die Leistung der Spielertypen kann aufgrund statistischer Kenngrößen getroffen werden?
- **Hypothesentests:** Welche Rückschlüsse auf die Leistung der Spielertypen kann aufgrund von Hypothesentests getroffen werden?
- **Regressionsanalyse:** Welche Rückschlüsse auf die Leistung der Spielertypen kann aufgrund von Regressionsanalysen getroffen werden?

5.1 Erste Validierung des Markovagenten in Prestudy II zum wiederholten Prisoner's Dilemma

Im folgenden Kapitel werden die Ergebnisse der initialen Validierung des Markovagenten vorgestellt. Die Untersuchung findet im Rahmen einer Prestudy statt, die ressourcenschonend eine erste Indikation über die Leistung des Markovagenten geben soll, bevor sich ein vollumfängliche Untersuchung über mehrere Spiele anschließt. Ziel ist es dabei insbesondere, die Anpassungen, welche von der Betaversion hin zur Vollversion Markovagenten vorgenommen wurden direktional zu bekräftigen.[50] Eine Übersicht der verschiedenen Versionierungen und deren empirische Untersuchung findet sich in Tabelle 4.5.

Das Kapitel kennzeichnet sich durch zwei Hauptergebnisse. Erstens zeigt es, dass aufgrund starker Reihenfolgeeffekte über die Spieldurchläufe hinweg eine Regressionsanalyse, die für derartige Effekte kontrolliert, zielführend ist. Der Sachverhalt erschließt sich bereits erstmalig anhand der deskriptivem Ergebnisse in Tabelle 5.3. Hier ist die Leistung von AgentM1 in die erste und die zweite Interaktion innerhalb der Vergleichsgruppen aufgeschlüsselt. Die zweite Interaktion weist dabei eine um mehr als 10% höhere durchschnittliche Auszahlung bei gleichem Markovmodell aus. Zweitens zeigt das Kapitel anhand der Regressionsanalysen, dass die Markovagenten menschlichen Spielern im wiederholten Prisoner's Dilemma nicht unterlegen ist. Vielmehr zeichnet sich direktional eine Überlegenheit ab, die aufgrund des Rauschen und der

[50] Kapitel B stellt die Änderungen von der Betaversion hin zur Vollversion des Markovagenten zusammenfassend dar.

hier vergleichsweise kleinen Fallzahl noch nicht signifikant ist. Derartige Signifikanzen ergeben sich in der anschließenden umfassenderen Untersuchung in Kapitel 5.2.

5.1.1 Experimenthergang

Der Termin für *Prestudy II* zur initialen Validierung des Markovagenten im wiederholten Prisoner's Dilemma fand am 17. April 2019 am Institut für Unternehmensführung des Karlsruher Instituts für Technologie statt. Die 50 registrierten Teilnehmer wurden auf fünf Sitzungen von je 75 Minuten mit bis zu 10 Teilnehmern verteilt. Insgesamt erschienen 48 Teilnehmer zum Experiment. Abbildung 5.1 stellt den Sachverhalt dar.[51] Analog zu Kapitel 4.2.2 dienen Aufteilung und Abstände zwischen den Sitzungen dem Ziel, die Kommunikation zu minieren und den Versuchsablauf robuster gegen eventuelle technische oder prozessuale Komplikationen zu machen (Chinczewski, 2019). Wie auch in Prestudy I wurde das wiederholte Prisoner's Dilemma aus den unter Kapitel 4.1.1 priorisierten Spieltypen aufgrund seiner weitläufigen Erforschung und dem großen Literaturkorpus für Prestudy II ausgewählt.

Innerhalb der Sitzungen spielt jeder Teilnehmer gegen vier Gegenspieler unter Berücksichtigung der Paarungslogik aus Kapitel 4.1.2. Die sich ergebende Paarung wird in Tabelle 5.1 präsentiert. Die Probanden spielen in zufälliger Reihenfolge in folgendem Aufbau:

- Einmal in der Rolle des Sparringspielers gegen einen anderen Menschen.
- Zweimal in der Rolle des Sparringspielers gegen AgentMx1 in identischer Parametrisierung. Die Buchstaben *a* (AgentMx1a) und *b* (AgentMx1b) referenzieren lediglich die erste beziehungsweise zweite Begegnung des Gegenspielers mit AgentMx1.

[51] Eine ungerade Teilnehmerzahl innerhalb einer Sitzung ist unproblematisch, da die Probanden nicht nur untereinander, sondern auch gegen Markovagenten spielen.

Abbildung 5.1: Organisatorische Übersicht der Termine zu Prestudy II. Quelle: Eigene Darstellung.

Tabelle 5.1: Übersicht der Spielerpaarungen von Prestudy II gemäß Kapitel 4.1.2. Die tatsächliche Reihenfolge wurde randomisiert. Quelle: Eigene Darstellung.

Interaktionen Proband A	Beobachteter Spieler	Sparringspieler	Kohorte
1	Proband B	Proband A	Sparringspieler A
2	AgentMx1a	Proband A	
3	AgentMx1b	Proband A	
4	AgentM01	Proband A	
5	Proband A	Proband C	Sparringspieler C
	...	...	

- Einmal in der Rolle des Sparringspielers gegen AgentM01.
- In der Rolle als beobachteter Spieler gegen einen menschlichen Sparringspieler, welcher nicht dem vorigen menschlichen Partner entspricht.

Die Parametrisierung der Markovagenten erfolgt nach Kapitel 3. AgentMx1 wählt auf Basis der historischen Vorhersagequalität rundenweise zwischen einem Gegnermodell mit Gedächtnistiefe $O^i = (0,1)$ oder $O^i = (1,1)$ aus. AgentM01 hingegen verwendet stets ein Gegnermodell mit $O^i = (0,1)$. Ein Spiel gegen AgentM11 mit einem Gegnermodell von $O^i = (1,1)$ wurde in der initialen Untersuchung von Prestudy II nicht einbezogen. Grund ist, dass das Sammeln von Informationen zur Bestätigung der grundsätzlich korrekten Funktionsweise des Markovagenten höher priorisiert wurde. Dieses sekundäre Untersuchungsziel lässt anhand AgentMx1 aufgrund der Auswahl zwischen mehreren Gegnermodellen umfassender realisieren. Infolgedessen wurde eine umfangreichere Stichprobe einer breiteren Untersuchung vorgezogen.

Von den 48 Teilnehmern konnten die Daten der 9 Probanden aus Sitzung 2 nicht verwendet werden. Während der Sitzung kam es aufgrund von Netzwerkproblemen zu unwiderruflich kompromittierten Daten. So konnten für die insgesamt 39 verbleibenden Probanden 39 Kohorten-Datensätze erzeugt werden.

5.1.2 Deskriptive Datenauswertung

Erste Anhaltspunkte über das Spielverhalten der verschiedenen Spielertypen in Prestudy II soll eine deskriptive Analyse liefern. Dabei werden Aktionsverhalten der Spieler sowie die mit dem Gegner realisierten Zustände betrachtet, um Aufschluss über deren Sentiment zu geben. Im zweiten Schritt findet eine Untersuchung statistischer Momente der normierten durchschnittlichen Auszahlung statt.

5.1.2.1 Aktionsverhalten und realisierte Zustände

Die in Tabelle 5.2 veranschaulichte Analyse der von den verschiedenen Spielertypen erreichten Spielzustände deutet auf Unterschiede im jeweiligen Spielverhalten hin. Auffällig ist, dass menschliche Spieler weitaus häufiger Aktion a_2^i spielen als in Prestudy I (siehe Tabelle C.2). Dies deutet auf eine aggressivere Spielweise der menschlichen Spieler und Gegenspieler in der Grundgesamtheit von Prestudy II hin. Trotz der stärkeren Tendenz der Menschen zu Aktion a_2^i verzeichnen die beiden Markovagenten im Spiel mit den Menschen als Gegner eine höhere Kooperationsrate als die Betaversion des Markovagenten in Prestudy I. Infolgedessen realisiert AgentM01 trotz der aggressiveren Gegenspieler häufiger den Kooperationszustand $a_1^i a_1^j$ als in Prestudy I. AgentMx1 hingegen bietet die Kooperation etwas seltener an, wird jedoch von den menschlichen Gegnern in $a_1^i a_2^j$ häufiger ausgebeutet als AgentM01 oder die Betaversion des Markovagenten in Prestudy I. Trotz der tendenziell aggressiveren Gegenspieler und der höheren Kooperationsrate der Markovagenten schaffen es letztere nicht weniger häufig als in Prestudy I den Gegner in $a_2^i a_1^j$ auszubeuten. Insofern lässt sich eine performantere Interaktionslogik für die Vollversion von AgentM im Vergleich zur Betaversion in Prestudy I vermuten. Die veränderte Spielweise von AgentM wird dabei hauptsächlich auf die hinzugekommene Verwendung von empirischen Priors sowie deren graduelles Update zurückgeführt, da es sich hierbei um den womöglich stärksten Eingriff in die Gegnermodelle handelt. Die Auswirkungen auf den erzielten normierten Payoff werden im nächsten Kapitel betrachtet.

Tabelle 5.2: Verteilung der Aktionswahl und der erreichten Spielzustände der Spielertypen nach aufsteigend sortierter normierter Auszahlung in Prestudy II zum Prisoner's Dilemma; mit Nash-Gleichgewicht des Stufenspiels (N), paretoeffizientem Zustand (grüne Farbe) und paretodominiertem Zustand (rote Farbe); exemplarisch um Kooperation (C) und Abweichung (D) ergänzt. Quelle: Eigene Darstellung.

	Aktionswahl				Realisierter Spielzustand			
	Spieler		Gegner (Mensch)					
	a_1^i	a_2^i	a_1^j	a_2^j	$a_1^i a_2^j$	$a_2^i a_2^j$	$a_1^i a_1^j$	$a_2^i a_1^j$
	C	D	C	D	CD	DD	CC	DC
Spielertyp								
- AgentMx1	73%	27%	70%	30%	10%	20%	63%	7%
- AgentM01	75%	25%	73%	27%	6%	21%	68%	5%
- Mensch	59%	41%	60%	40%	7%	33%	53%	7%
Normierte Auszahlung								
- Spieler					0	20	60	100
- Gegner (Mensch)					100	20	60	0
Zustandseigenschaft						N		

5.1.2.2 Normierte Auszahlung

Die durchschnittlichen Auszahlungen der Spielertypen in Tabelle 5.3 zu Prestudy II grenzen sich deutlicher zwischen den Spielern ab als in Prestudy I (siehe Tabelle C.3). Die AgentM-Spieler weisen im Mittel einen höheren durchschnittlichen normierten Payoff aus als menschliche Spieler. Auffällig ist, dass das Auszahlungsniveau über alle Spielertypen geringer ausfällt als in Prestudy I. Dies wird wie im vorigen Kapitel geschildert auf das systematisch andere Spielverhalten der menschlichen Spieler und Gegenspieler zurückgeführt, denn Menschen in Prestudy II zeichnen sich durch ein aggressiveres Spielverhalten aus. Hierdurch wird das erzielen hoher Auszahlungen für alle Teilnehmer erschwert. Die deskriptiv bessere Leistung der Markovagenten im Vergleich mit menschlichen Spielern wird als Indikator für eine zielführende Anpassung der Betaversion von AgentM gewertet. Im Rahmen von Hypothesentests soll dies im nächsten Kapitel weiter betrachtet werden. Da in Prestudy II AgentMx1 in identischer Konfiguration zweimal gegen jeden Menschen spielt, wurde die erste (AgentMx1a) und zweite (AgentMx1b) Interaktion in der Tabelle ergänzend separat ausgewiesen. Dabei fällt auf, dass die durchschnittliche Auszahlung von AgentMx1b mit 6.01. Punkten Unterschied deutlich über der von AgentMx1a liegt. Der Medianwert der normierten Auszahlung von AgentMx1b von 60.00 entspricht dem Punktwert der beidseitigen Kooperationslösung und zeigt, dass die Spieler sich ceteris paribus in mindestens 50% der Runden kooperativ einigen konnten. Dies deutet bereits jetzt auf die Präsenz signifikanter Lerneffekte auf Seiten der menschlichen Spielerpopulation hin. Der Sachverhalt findet nach Abhandlung der Hypothesentests im Rahmen der Regressionsanalysen Beachtung.

5.1.3 Hypothesentests

Im Folgenden werden die Ergebnisse von Prestudy II zur initialen Validierung des Markovagenten im wiederholten Prisoner's Dilemma anhand von Hypothesentests präsentiert. Der zu zei-

Tabelle 5.3: Deskriptive Auswertung der durchschnittlichen normierten Auszahlungen $\bar{\rho}^i$ der beobachteten Spielertypen für Prestudy II. *AgentMx1a* entspricht der ersten und *AgentMx1b* der zweiten Interaktion des identisch konfigurierten AgentMx1. Quelle: Eigene Darstellung.

			Perzentile				
Spielertyp	n	$\bar{x}$	25%	$\tilde{x}$	75%	s	$\frac{s}{\bar{x}}$
AgentMx1	78	48.52	40.00	56.19	60.00	14.13	0.29
- AgentMx1a	39	45.52	36.19	46.67	60.00	14.45	0.32
- AgentMx1b	39	51.53	47.62	60.00	60.00	13.31	0.26
AgentM01	39	49.69	41.90	57.14	60.00	14.05	0.28
Mensch	39	45.79	27.62	50.48	60.00	16.01	0.35

gende Sachverhalt ist eine signifikante positive Abweichung der Leistung des Markovagenten im Vergleich der von menschlichen Spielern. Ausgangsbasis für sämtliche Tests ist die Datenlage aus Tabelle 5.3.

5.1.3.1 Zentrale Analysen

Zum Vergleich der Auszahlung der Spielertypen bieten sich parametrische und nicht-parametrische Verfahren an (vgl. Cleff, 2019, S. 145), wovon sich der zwei-Stichproben t-Test und der gepaarte t-Test als *parametrische Verfahren* im Folgenden als unpassend herausstellen. Die Verwendung eines zwei-Stichproben t-Tests wird aufgrund der unzureichenden Erfüllung grundlegender Annahmen ausgeschlossen. Insbesondere die *Unabhängigkeit der Beobachtungen* ist innerhalb der Spieler-Kohorten nicht gegeben (vgl. Toutenburg & Heumann, 2008, S. 142-145). Aufgrund der Strategie des Sparringspielers weist die realisierte Auszahlung der beobachteten Spieler eine Abhängigkeit innerhalb der Gruppe auf. Die parametrische Alternative eines gepaarten t-Tests basiert auf der Überlegung, die Abhängigkeit der Beobachtungen aufzulösen, indem die paarweise Differenz der Auszahlungen zwischen den Spielertypen je Sparringspielerkohorte betrachtet wird.[52] So existiert je Sparringspielerkohorte und zu testendes Paar von Spielertypen nur noch eine Beobachtung (Cleff, 2019, S. 157-160). Der gepaarte t-Test kann jedoch ebenfalls aufgrund einer fehlenden Annahmeerfüllung ausgeschlossen werden (vgl. Toutenburg & Heumann, 2008, S. 145-147). Insbesondere ist eine *Normalverteilung der Differenz* der Auszahlungen der Spielertypen innerhalb der Sparringspielerkohorten nicht gegeben, wie die Ergebnisse des Shapiro-Wilk-Tests (vgl. Royston, 1992) mit Anpassung durch Shapiro und Wilk (1965) in Tabelle 5.5 zeigt. Die Nullhypothese H_0, dass bei der Differenz der Auszahlungswerte eine Normalverteilung vorliegt konnte bei einem Signifikanzniveau von 5% für alle paarweisen Vergleiche bis auf jene mit AgentMx1a verworfen werden.[53]

[52] Es wird die Differenz zwischen der durchschnittlichen Auszahlungsleistung der *beobachteten Spieler* innerhalb einer Kohorte gebildet (siehe Tabelle 5.1)

[53] AgentMx1a und AgentMx1b beziehen sich wie in Tabelle 5.1 dargestellt auf den identischen AgentMx1. Es wird lediglich differenziert, ob es sich um die erste (Buchstabe *a*) oder zweite (Buchstabe *b*) Begegnung des Gegenspielers mit AgentMx1 handelt.

Tabelle 5.4: Übersicht von Mittelwerttestverfahren für den Vergleich von zwei Stichproben. Quelle: Eigene Darstellung in Anlehnung an Cleff (2019, S. 144-145).

	Testverfahren	
Stichproben	Parametrisch	Nicht-parametrisch
Unabhängig	Zwei-Stichproben t-Test	Mann-Whitney-U-Test
Abhängig	Gepaarter t-Test	Wilcoxon-Vorzeichen-Rang-Test

Tabelle 5.5: Ergebnisse des Shapiro-Wilk-Tests (Shapiro & Wilk, 1965) auf H_0, dass bei der Differenz der Auszahlungswerte $D = X_1 - X_2$ für Prestudy II eine Normalverteilung vorliegt (rote Kennzeichnung für verworfene Normalverteilungsannahme (NvA); indikativ: † $p < 10\%$; signifikant: * $p < 5\%$, ** $p < 1\%$, *** $p < 0.1\%$). Quelle: Eigene Darstellung.

Vergleich				
Spielertyp 1	Spielertyp 2	p		H_0
AgentMx1a	Mensch	17.8%		✓
AgentMx1b	Mensch	0.3%	***	×
AgentM01	Mensch	4.7%	*	×

Eine Auswertung anhand *nicht-parametrischer* Verfahren ist bis hierhin nicht ausgeschlossen. Grundsätzlich kommen der Mann-Whitney-U-Test und der Wilcoxon-Vorzeichen-Rang-Test in Betracht (vgl. Cleff, 2019, S. 145). Der Mann-Whitney-U-Test (Mann & Whitney, 1947) wird dabei identisch zum zwei-Stichproben t-Test aufgrund der fehlenden Unabhängigkeit der Beobachtungen ausgeschlossen (vgl. Cleff, 2019, S. 181-185). Ein Wilcoxon-Vorzeichen-Rang-Test untersucht analog zum gepaarten t-Test die Differenz zwischen den abhängigen Stichproben (Wilcoxon, 1945, 1947).[54] Im konkreten Anwendungsfall kommt der Wilcoxon-Vorzeichen-Rang-Test in Frage, da keine Verletzung der Symmetrie-Annahme (vgl. Toutenburg & Heumann, 2008, S. 182) festgestellt werden konnte. Die Nullhypothese H_0 bezüglich der Symmetrie der Differenz der paarweise nach Sparringspielern gruppierten Payoffs kann anhand des Tests von D'Agostino et al. (1990) mit Anpassung von Royston (1991) nicht überzeugend verworfen werden (siehe Tabelle 5.6).

Tabelle 5.7 zeigt die Ergebnisse des Wilcoxon-Vorzeichen-Rang-Tests mit Signifikanzniveau von 5%. Die Nullhypothese H_0, dass der Median der Auszahlungs-Differenzen im paarwei-

[54] Es wird die Differenz zwischen der durchschnittlichen Auszahlungsleistung der *beobachteten Spieler* innerhalb einer Kohorte gebildet (siehe Tabelle 5.1)

Tabelle 5.6: Ergebnisse des Tests auf Symmetrie der Differenzen (D'Agostino et al., 1990; Royston, 1991) mit H_0, dass Auszahlungsdifferenzen $D = X_1 - X_2$ für Prestudy II symmetrisch sind (rote Kennzeichnung für verworfene Symmetrieannahme (SA); indikativ: † $p < 10\%$; signifikant: * $p < 5\%$, ** $p < 1\%$, *** $p < 0.1\%$). Quelle: Eigene Darstellung.

Vergleich				
Spielertyp 1	Spielertyp 2	p		H_0
AgentMx1a	Mensch	24%		✓
AgentMx1b	Mensch	4%	*	×
AgentM01	Mensch	79%		✓

Tabelle 5.7: Ergebnisse des Wilcoxon-Vorzeichen-Rang-Tests (Wilcoxon, 1945) für gepaarte Daten aus Prestudy II auf H_0, dass zwei Stichproben aus der gleichen Verteilung gezogen wurden und demnach der Median der Auszahlungsdifferenzen $D = X_1 - X_2$ Null ist (grüne Kennzeichnung für positive Auszahlungsdifferenzen; indikativ: † $p < 10\%$; signifikant: * $p < 5\%$, ** $p < 1\%$, *** $p < 0.1\%$) unter Berücksichtigung der Erfüllung der zugrundeliegenden Symmetrieannahme (SA) aus Tabelle 5.6. Quelle: Eigene Darstellung.

Vergleich								
Spielertyp 1	Spielertyp 2	$\bar{D}$	$\tilde{D}$	z	p		H_0	SA
AgentMx1a	Mensch	-0.27	0.00	-0.53	59.5%		✓	✓
AgentMx1b	Mensch	5.74	0.95	2.05	4.1%	*	×	×
AgentM01	Mensch	3.91	0.00	1.09	27.7%		✓	✓

sen Vergleich Null ist, konnte für einen der paarweisen Vergleiche verworfen werden. Insofern konnte in Prestudy II eine signifikante Leistungsdifferenz lediglich zwischen dem überlegenen AgentMx1b und den menschlichen Spielern festgestellt werden. Dieses Ergebnis ist jedoch aufgrund der Variablenstruktur kritisch zu bewerten. Insbesondere das Fehlen einer signifikanten Differenz für AgentMx1a sowie die hochsignifikanten Lerneffekte aus Prestudy I (siehe Tabelle C.10) legen nahe, dass das Ergebnis durch hier nicht kontrollierte Lerneffekte verzerrt wurde. Der Sachverhalt wird in Kapitel 5.1.4 aufgearbeitet.

Zusammenfassend lässt sich festhalten, dass die im Durchschnitt dem Menschen überlegene Leistung des Markovagenten (siehe Tabelle 5.3) auf Basis von Hypothesentests in Prestudy II nicht signifikant abweicht. Die positive Tendenz der realisierten Markovagent-Payoffs im Vergleich zu Prestudy I sowie die nicht signifikante aber direktional vorhandene Überlegenheit der Markovagentenpayoffs im Vergleich zu den Menschen in Prestudy II wird als Indiz für zielführende Weiterentwicklung der Betaversion des Markovagenten gewertet. Im Anschluss an nachfolgende Robustheitsanalysen zu den zuvor ausgeschlossenen Hypothesentests findet eine Regressionsanalyse zu Prestudy II statt, welche Lerneffekte adressiert.

5.1.3.2 Zusätzliche Robustheitsanalysen

Die Ergebnisse der vorangegangenen Hypothesentests werden durch die ergänzenden Tests bestätigt (siehe Tabellen A.1 bis A.3 im Anhang). Tabelle A.1 stellt den zwei-Stichproben t-Test dar, während Tabelle A.2 den gepaarten t-Test präsentiert (Student, 1908). Tabelle A.3 hingegen zeigt die Ergebnisse des Mann-Whitney-U-Tests (Mann & Whitney, 1947). Die Ergebnisse der Robustheitsanalysen decken sich weitestgehend mit den Ergebnissen der Testhypothese und unterstreichen somit vorbehaltlich der beschriebenen Einschränkungen deren Validität.

5.1.4 Regressionsanalyse

Im Folgenden werden die Ergebnisse von Prestudy II zur initialen Validierung des Markovagenten im wiederholten Prisoner's Dilemma anhand von Regressionsanalysen präsentiert. Der zu zeigende Sachverhalt ist eine signifikante positive Abweichung der Leistung des Markovagenten im Vergleich der von menschlichen Spielern unter Berücksichtigung von Einflüssen wie Lerneffekten auf Seiten der menschlichen Spieler sowie der Abhängigkeit der Leistung vom Spielverhalten des Gegenspielers.

5.1.4.1 Zentrale Analysen

Im Folgenden sollen die Daten der Prestudy II im Rahmen einer Regression analysiert werden. Als *abhängige Variable* wird der normierte Payoff (siehe Kapitel 4.4) der betrachteten Spieler als Maß für dessen Fähigkeit in den entsprechenden Spielen zielführend zu agieren verwendet.

Aufgrund der Interdependenz zwischen dem erzielbaren Payoffs des beobachteten Spielers und der Strategie des Gegenspielers (siehe Kapitel 4.1.2.1) ist eine Panelregression sachdienlich. Dementsprechend wird der menschliche Gegenspieler in Einklang mit dem Experimenthergang in Tabelle 5.1 (analog zu Abbildung 4.2) als *Panelvariable* definiert. Die Abhängigkeit wird im Spiel gegen zwei verschiedene exemplarische Gegner im wiederholten Prisoner's Dilemma deutlich. Der erste Gegner bedient sich einer *Always Defect* Strategie, sodass der maximal erzielbare Durchschnitts-Payoff gegen diesen Spieler bei 1 liegt. Der zweite Gegner bedient sich einer *Always Cooperate* Strategie, sodass der maximal erzielbare Durchschnitts-Payoff gegen diesen Spieler bei 5 liegt. Soll nun eine beliebige Menge an Teststrategien untereinander im Spiel gegen die Beispielstrategien verglichen werden, so wird schnell klar, dass ein Leistungsvergleich zwischen zwei Testspielern bei möglichst konstanter Gegnerstrategie zielführender ist. Hintergrund ist, dass der realisierte Spielverlauf nicht nur von eigenem Verhalten, sondern insbesondere auch vom Aktionsverhalten des Gegners abhängt. Folgerichtig werden Spielverläufe zwischen *verschiedenen* Spielertypen gegen den *gleichen* menschlichen Gegner verglichen.

Aufgrund dieser Abhängigkeit innerhalb der Kohorten eignet sich aus qualitativer Sicht als *Modelltyp* vornehmlich ein within Fixed Effects Panelmodell. Dieses maximiert die Varianzaufklärung innerhalb der Kohorte (siehe Kapitel 4.1.2.1), welche gegen den gleichen Gegenspieler gemessen wird (vgl. Das, 2019, S. 474-475). Der potentiellen Präsenz von Heteroskedastizität wird durch die Berücksichtigung clusterrobuster Standardfehler Rechnung getragen (vgl. Das, 2019, S. 481). Die Verwendung eines Fixe Effects Modells wird durch die vorangegangenen qualitativen Überlegungen unterstützt.[55]

[55] Die Ergebnisse quantitativer Spezifikationstests werden in Tabelle 5.9 dargestellt.

Die zu schätzenden Koeffizienten des Regressionsmodells differenzieren im konkreten Anwendungsfall zwischen folgenden Aspekten als erklärende *Kontrollvariablen*:

- **Wer spielt gegen den menschlichen Sparring-Spieler?** Durch die Verwendung von Dummyvariablen wird berücksichtigt, ob es sich bei dem beobachteten Spieler um einen Spieler vom Typ *Mensch*, *AgentMx1* oder *AgentM01* handelt. So können Rückschlüsse auf den Einfluss des Spielertyps auf dessen normierten Durchschnitts-Payoff getroffen werden. Als *Referenzspieler* wird dabei stets der menschliche beobachtete Spieler verwendet, sodass die Regressionskoeffizienten der anderen Spielertypen als relative Leistungsdifferenz zu diesen zu verstehen ist.[56]
- **Wie viel Erfahrung im Experiment haben menschliche Gegenspieler?** Um temporale Lerneffekte der Gegenspieler zu berücksichtigen, findet darüber hinaus die Zählvariable *Anzahl Spiele (Gegner)* Einfluss in das Modell. Sie gibt an, auf das wievielte Spiel für nicht beobachteten menschlichen Sparringspieler sich die Interaktion bezieht. Die Lerneffekte werden an der Erfahrung des Sparringspielers gemessen, da bei einer Berücksichtigung der Lerneffekte anhand des beobachteten Spielers eine ungleiche Varianzzurechnung zwischen den Spielertypen vorläge. Grund ist, dass AgentM niemals Lerneffekte aufweist, da er vor jeder neuen Interaktion vollständig zurückgesetzt wird. Eine derartige asymmetrische Kontrollvariable würde die Vergleichbarkeit innerhalb der Spielertypen des Modells verzerren.

Das insgesamt signifikante Modell in Tabelle 5.8 zeigt, dass die Leistung beider Markovagenten für das wiederholte Prisoner's Dilemma in Prestudy II nicht signifikant von der Leistung der menschlichen Referenzspieler abweicht. Insbesondere die signifikant überlegene Leistung des AgentMx1b, also der zweiten Begegnung des Gegenspielers mit AgentMx1 zeigt sich in der Regression für die Gesamtheit der Spiele von AgentMx1 nicht. Der vormalige Leistungsunterschied wird wie bereits in Kapitel 5.1.3.1 diskutiert auf die hier kontrollierten signifikanten Lerneffekte zurückgeführt. Die Robustheit der Regressionsergebnisse wird anhand alternativer Modelle im folgenden Kapitel 5.1.4.2 sichergestellt.

Zwischenfazit

Die Markovagenten können in Prestudy II zum wiederholten Prisoner's Dilemma direktional bessere aber nicht signifikant bessere Ergebnisse als menschliche Referenzspieler erzielen.

[56] Die menschlichen Referenzspieler entsprechen den Probanden in Spalte *Beobachtete Spieler* in Tabelle 5.1. Die Leistung der Sparringspieler wird zu keiner Zeit analysiert.

Tabelle 5.8: Ergebnisse des Fixed Effects Panelregressionsmodells mit clusterrobusten Standardfehlern (vgl. Das, 2019, S. 482) zu Prestudy II im wiederholten Prisoner's Dilemma (indikativ: † $p < 10\%$; signifikant: * $p < 5\%$, ** $p < 1\%$, *** $p < 0.1\%$). Quelle: Eigene Darstellung.

	Koeffizient	Standardfehler	t	p	
Spielertyp					
- Mensch (Referenz)	–				
- AgentMx1	2.57	2.12	1.21	23.3%	
- AgentM01	3.82	2.28	1.68	10.2%	
Lerneffekte: Anzahl Spiele (Gegner)	3.21	0.67	4.80	0.0%	***
Konstante	36.24	2.61	13.86	0.0%	***
Beobachtungen	156				
Gruppen	39				
R^2_{within}	0.25				
$R^2_{between}$	0.00				
$R^2_{overall}$	0.09				
$F(3,38)$	8.26				
Signifikanz ($\mathbf{P} > F$)	0.0%				***

Gleichwohl konnte die die Vollversion des Markovagenten in Prestudy II im Gegensatz zu der Betaversion des Markovagenten in Prestudy I (siehe Tabelle C.10) relativ zu den menschlichen Spielern eine positive Effektstärke bei aussagekräftigeren p-Werten erzielen. Auf Basis dieser Indikation folgt nach der Präsentation der Robustheitsanalysen zu Prestudy II die weiterführende Validierung des Markovagenten in Experiment I bis III.

Tabelle 5.9: Ergebnisse von Tests zur Modellspezifikation der Panelregression zu Prestudy II (* $p < 0.1$, ** $p < 5\%$, *** $p < 1\%$). Quelle: Eigene Darstellung.

Test	Statistik	p	Empfohlenes Modell
Hausman Test	$\chi^2(3) = 0.86$	83.6%	Random Effects
Robuster Hausman Test	$F(1,38) = 0.73$	39.9%	Random Effects
Sargan-Hansen Test	$\chi^2(1) = 0.90$	34.2%	Random Effects
Robuster Sargan-Hansen Test	$\chi^2(1) = 0.98$	32.3%	Random Effects

5.1.4.2 Zusätzliche Robustheitsanalysen

Quantitative Tests für Empfehlungen zur Gestaltung des Modells legen ein Random Effects Modell nahe (siehe Tabelle 5.9). Es findet daher eine Random Effects Panelregression zur Berücksichtigung zeitinvarianter unbeobachteter Heterogenität Eingang (vgl. Das, 2019, S. 487). Außerdem wird eine Ordinary Least Squares (OLS) Regression durchgeführt, welche den Einfluss des Gegenspielers als Panelvariable vernachlässigt. So wird den anderslautenden Empfehlungen quantitativer Spezifikationstests wie dem gewöhnlichen Hausman Test (Hausman, 1978), dem robusten Hausman Test (Greene, 2008; Hoechle, 2007; Wooldridge, 2009) und dem (robusten) Sargan-Hansen Test (Arellano, 1993; Wooldridge, 2002) Rechnung getragen.

Die in Kapitel 5.1.4.1 festgestellte, aus qualitativer Sicht als besser bewertete Eignung eines Fixed Effects Modells ist dies jedoch nicht abträglich. Die in Tabelle A.4 (siehe Anhang) dargestellten Ergebnisse einer Random Effects Regression bestärken, beziehungsweise übertreffen die Ergebnisse des Fixed Effects Modells. Tabelle A.5 (siehe Anhang) zeigt die Ergebnisse einer OLS Regression, welche keine Paneleffekte berücksichtigt.

5.2 Weiterführende Validierung des Markovagenten Experimente zu ausgewählten Spielen

Im folgenden Kapitel werden die Ergebnisse der weiterführenden Validierung des Markovagenten in den wiederholten Spielen *Prisoner's Dilemma*, *Chicken Game* und *Hero Game* (zur Spielauswahl siehe Kapitel 4.1.1) vorgestellt. Die Kapitelstruktur gestaltet sich analog zur Übersicht in Kapitel 5. Wie in Tabelle 4.5 dargestellt, wird die identische Version und Gestaltung des Markovagenten *AgentM* wie in Prestudy II eingesetzt. Einziger Unterschied ist der Transfer auf das Chicken Game und das Hero Game über das Prisoner's Dilemma hinaus. Das Prisoner's Dilemma wird dennoch erneut untersucht, da die Datenbasis aus Prestudy II aufgrund des Testcharakters der Erhebung vergleichsweise klein ausfällt. Wesentliches Ergebnis ist die signifikant höhere Leistung des Markovagenten gegenüber den menschlichen Vergleichsspielern, welche über verschiedene Spiele hinweg robust ist (siehe Kapitel 5.2.4).

5.2.1 Experimenthergang

Die Experimente I, II und III zur weiterführenden Validierung des Markovagenten fanden am 16. Juli 2019 (Chicken Game), 22. Juli 2019 (Hero Game) und 24. Oktober 2019 (Prisoner's Dilemma) am Institut für Unternehmensführung des Karlsruher Instituts für Technologie statt. Die je Experiment 50 registrierten Teilnehmer wurden auf fünf Sitzungen von je 75 Minuten mit bis zu zehn Teilnehmern verteilt. Insgesamt erschienen 131 Teilnehmer zu den Experimenten. Davon verteilen sich 43 Probanden auf das Chicken Game, 45 Probanden auf das Hero Game und 43 auf das Prisoner's Dilemma. Abbildung 5.2 stellt den Sachverhalt dar.[51] Wie auch in

16. Juli 2019 – Experiment I: Chicken Game

08:30 – 09:45	10:00 – 11:15	11:30 – 12:45	13:00 – 14:15	14:30 – 15:45
Sitzung 1	Sitzung 2	Sitzung 3	Sitzung 4	Sitzung 5
Registriert: 10	Registriert: 10	Registriert: 10	Registriert: 10	Registriert: 10
Teilgenommen: 7	Teilgenommen: 10	Teilgenommen: 8	Teilgenommen: 9	Teilgenommen: 9

22. Juli 2019 – Experiment II: Hero Game

08:30 – 09:45	10:00 – 11:15	11:30 – 12:45	13:00 – 14:15	14:30 – 15:45
Sitzung 1	Sitzung 2	Sitzung 3	Sitzung 4	Sitzung 5
Registriert: 10	Registriert: 9	Registriert: 10	Registriert: 8	Registriert: 8
Teilgenommen: 10	Teilgenommen: 9	Teilgenommen: 10	Teilgenommen: 8	Teilgenommen: 8

24. Oktober 2019 – Experiment III: Prisoner's Dilemma

08:30 – 09:45	10:00 – 11:15	11:30 – 12:45	13:00 – 14:15	14:30 – 15:45
Sitzung 1	Sitzung 2	Sitzung 3	Sitzung 4	Sitzung 5
Registriert: 10	Registriert: 10	Registriert: 10	Registriert: 10	Registriert: 10
Teilgenommen: 10	Teilgenommen: 10	Teilgenommen: 9	Teilgenommen: 5	Teilgenommen: 9

Abbildung 5.2: Organisatorische Übersicht der Experimente I bis III. Quelle: Eigene Dastellung.

Prestudy I und II dient die Aufteilung der Sitzungen dem Ziel, die Kommunikation zu minieren (Chinczewski, 2019).

Innerhalb jeder Sitzungen spielen die Teilnehmer unter Berücksichtigung der Paarungslogik aus Kapitel 4.1.2 gegen mehrere menschliche Gegenspieler. Die sich ergebenden Paarungen werden für das Chicken Game und den Hero Game in Tabelle 5.10 und für das Prisoner's Dilemma in Tabelle 5.11 präsentiert. Im Prisoner's Dilemma wird neben den menschlichen Spielern und den Markovagenten außerdem ein Tit-for-Tat-Spieler verwendet. Dieser soll im späteren Verlauf der Arbeit als geläufiger Benchmark für die Leistung des Markov-Bots dienen. Die Probanden spielen in zufälliger Reihenfolge in folgendem Aufbau:

- Einmal in der Rolle des Sparringspielers gegen einen Menschen.
- Einmal in der Rolle des Sparringspielers gegen den AgentMx1.
- Einmal in der Rolle des Sparringspielers gegen den AgentM01.
- Einmal in der Rolle des Sparringspielers gegen den AgentM11.
- Einmal In der Rolle als beobachteter Spieler gegen einen menschlichen Sparringspieler, welcher nicht dem vorigen menschlichen Partner entspricht.

- In Experiment III außerdem einmal in der Rolle des Sparringspielers gegen Tit-for-Tat.

Die Parametrisierung der Markovagenten erfolgt nach Kapitel 3. AgentMx1 wählt auf Basis der Anpassungsgüte der beiden Gegnermodelle rundenweise aus einem $O^i = (0,1)$ und einem $O^i = (1,1)$ Gegnermodell aus. AgentM01 hingegen verwendet stets ein $O^i = (0,1)$ Gegnermodell. AgentM11 wiederum verwendet stets ein $O^i = (1,1)$ Gegnermodell. Die Eröffnung im Chicken Game und Prisoner's Dilemma erfolgt gemäß der Markov for Tat Logik: Spiele so lange kooperativ, bis der Gegner das erste Mal abweicht und agiere danach auf Basis der Markovlogik. Im Hero Game ist kein eindeutig kooperativer Zustand vorhanden, sodass der AgentM in Runde 1 mit a_1^i eröffnet, um Koordinationsbereitschaft zu zeigen. Ab Runde 2 greift er auf die Markovlogik zurück (siehe Kapitel 3.3.2.3).

In den Experimenten zum Hero Game und dem Prisoner's Dilemma konnten die Daten von allen 45, beziehungsweise 43 Teilnehmern erfasst werden. Im Chicken Game konnten die Spieldaten eines Probanden in Sitzung 3 aufgrund einer technischen Fehlfunktion eines Tablets nicht korrekt erfasst, sodass nur 42 Datensätze von den insgesamt 43 Probanden verwendet werden können. Weiterhin wurde eines der Spiele gegen AgentMx1 nicht aufgezeichnet, sodass für diesen Spielertyp nur 41 Datensätze vorliegen. Auch ein Spielverlauf von AgentMx1 konnte im Chicken Game nicht aufgezeichnet werden.

5.2.2 Deskriptive Datenauswertung

Erste Anhaltspunkte über das Spielverhalten der verschiedenen Spielertypen in den Experimenten I bis III soll eine deskriptive Analyse liefern. Dabei werden Aktionsverhalten der Spieler sowie die mit dem Gegner realisierten Zustände betrachtet, um Aufschluss über deren Sentiment zu geben. Im zweiten Schritt findet eine Untersuchung statistischer Momente der normierten durchschnittlichen Auszahlung statt.

Tabelle 5.10: Spielerpaarungen der Experimente I und II zum Chicken Game und Hero Game gemäß Kapitel 4.1.2. Die tatsächliche Reihenfolge wurde randomisiert. Quelle: Eigene Darstellung.

Spiele für Proband A	Beobachteter Spieler	Sparringspieler	Kohorte
1	Proband B	Proband A	Sparringspieler A
2	AgentMx1	Proband A	
3	AgentMx1	Proband A	
4	AgentM11	Proband A	
5	Proband A	Proband C	Sparringspieler C
	...	...	

Tabelle 5.11: Spielerpaarungen des Experiments III zum Prisoner's Dilemma gemäß Kapitel 4.1.2. Die tatsächliche Reihenfolge wurde randomisiert. Quelle: Eigene Darstellung.

Spiele für Proband A	Beobachteter Spieler	Sparringspieler	Kohorte
1	Proband B	Proband A	Sparringspieler A
2	AgentMx1	Proband A	
3	AgentMx1	Proband A	
4	AgentM11	Proband A	
5	Tit-for-Tat	Proband A	
6	Proband A	Proband C	Sparringspieler C
	...	...	

5.2.2.1 Aktionsverhalten und realisierte Zustände

Im Folgenden findet eine differenzierte Analyse der in den drei Experimenten durch die verschiedenen Spielertypen jeweils erreichten Spielzustände statt. In chronologischer Reihenfolge werden so die Ergebnisse zu den wiederholten Spielen Chicken Game, Hero Game und Prisoner's Dilemma abgearbeitet.

Tabelle 5.12: Verteilung der Aktionswahl und der erreichten Spielzustände der Spielertypen nach aufsteigend sortierter normierter Auszahlung in Experiment I zum Chicken Game; mit Nash-Gleichgewicht des Stufenspiels (N), paretoeffizientem Zustand (grüne Farbe) und paretodominiertem Zustand (rote Farbe). Quelle: Eigene Darstellung.

	Aktionswahl				Realisierter Spielzustand			
	Spieler		Gegner (Mensch)					
	a_1^i	a_2^i	a_1^j	a_2^j	$a_2^i a_2^j$	$a_1^i a_2^j$	$a_1^i a_1^j$	$a_2^i a_1^j$
Spielertyp								
- AgentMx1	77%	23%	42%	58%	13%	45%	32%	10%
- AgentM01	83%	17%	51%	49%	10%	39%	44%	7%
- AgentM11	69%	31%	48%	52%	16%	37%	33%	15%
- Mensch	36%	64%	38%	62%	45%	17%	20%	19%
Normierte Auszahlung								
- Spieler					0	20	60	100
- Gegner (Mensch)					0	100	60	20
Eigenschaft						N		N

Die in Tabelle 5.12 veranschaulichte Analyse der von den verschiedenen Spielertypen erreichten Spielzustände in Experiment I zum *wiederholten Chicken Game*. Beobachtete *menschliche Spieler* grenzten sich mit 64% ihrer Züge in a_2^i mit deutlichem Abstand zu den Markovagenten als am konfrontationsfreudigsten ab. Hierdurch mussten sie in fast der Hälfte aller Runden Zustand $a_2^i a_2^j$ ohne Auszahlung hinnehmen, konnten aber in 19% der Fälle den Maximalpayoff realisieren. Mit ähnlicher Häufigkeit wurde eine Kooperationslösung in $a_1^i a_1^j$ oder alternativ der Zustand $a_1^i a_2^j$ erreicht. Als starker Kontrast dazu stellt sich *AgentM01* dar, welcher nur in 17% aller Runden a_2^i spielte. Dadurch ging er nur in 10% aller Runden leer aus und konnte im Vergleich mit Menschen mit 44% mehr als doppelt so häufig die Kooperationslösung realisieren. Dennoch wurde er auch mehr als doppelt so häufig in $a_1^i a_2^j$ ausgenutzt und konnte nur in 7% der Runden die höchste Auszahlung erwirtschaften. *AgentMx1* zeigte sich mit 23% der Züge in a_2^i ähnlich zurückhaltend, wenn auch etwas fordernder. Folglich wurde der Bestzustand mit 10% häufiger erzielt als bei AgentM01. Dennoch musste der Spielertyp Verluste durch das häufigere Erreichen der unerwünschten Zustände $a_2^i a_2^j$ und $a_1^i a_2^j$ zu Lasten des Kooperationszustandes $a_1^i a_1^j$ hinnehmen. *AgentM11* zuletzt gab sich mit 31% der Züge in a_2^i als der konfrontativste Markovagent, auch wenn er nur weniger als halb so oft a_2^i spielte als Menschen. Der Bestzustand $a_2^i a_1^j$ wurde in 15% der Runden erreicht, also mit einer Differenz von nur 4% zu den Menschen. Trotz des vergleichsweise häufigen Ausbeuten des Gegners im vorgenannten Zustand wurde der schlechteste Zustand $a_2^i a_2^j$ in nur 16% der Runden erreicht. Im Vergleich dazu mussten Menschen dies in 45% der Runden hinnehmen. Außerdem wurde trotz offensiverer Spielweise vergleichbar öfter die Kooperationslösung als bei AgentMx1 erzielt und gleichzeitig der Anteil an erlittenen Ausbeutungen in $a_1^i a_2^j$ deutlich reduziert.

Tabelle 5.13 veranschaulicht die Analyse der von den verschiedenen Spielertypen erreichten Spielzustände in Experiment II zum *wiederholten Hero Game*. Beobachtete *menschliche Spieler* bestanden mit 69% ihrer gespielten Züge unter allen Spielertypen am häufigsten auf die Durchsetzung ihrer Bestlösung, welche sie mit 26% auch am häufigsten erreichen konnte. Gleichwohl wurde so der unerwünschte Zustand $a_1^i a_1^j$ mit nur 2% am besten vermieden. Insgesamt konnten so in 55% aller Züge die attraktivsten beiden Zustände $a_1^i a_2^j$ und $a_2^i a_1^j$ realisiert werden. *AgentM01* verfolgte mit 50% seiner Züge in a_2^i die durchsetzungswilligste Herangehensweise unter den Markovagenten. Trotz der im Vergleich zu den Menschen entgegenkommenderen Spielweise realisierte er den Bestzustand nur 2% seltener als erstere. Darüber hinaus wurde der zweitbeste Zustand $a_1^i a_2^j$ mit 45% wesentlich öfter erreicht, wohingegen der unattraktive Konfliktzustand $a_2^i a_2^j$ wesentlich seltener in Kauf genommen werden musste. *AgentM11* zeigte sich mit 42% a_2^i wiederum entgegenkommender als AgentMx1. Infolgedessen erreichte er den Zustand $a_2^i a_1^j$ insgesamt 7% seltener als der Mensch, konnte so jedoch den zweitbesten Zustand $a_1^i a_2^j$ mit 54% auch häufiger realisieren. *AgentMx1* war AgentM11 im durchschnittlichen Zugverhalten mit 43% der Züge in a_2^i ähnlich. Trotz der zurückhaltenden Spielweise wurde der

Tabelle 5.13: Verteilung der Aktionswahl und der erreichten Spielzustände der Spielertypen nach aufsteigend sortierter normierter Auszahlung in Experiment II zum Hero Game; mit Nash-Gleichgewicht des Stufenspiels (N), paretoeffizientem Zustand (grüne Farbe) und paretodominiertem Zustand (rote Farbe). Quelle: Eigene Darstellung.

	Aktionswahl				Realisierter Spielzustand			
	Spieler		Gegner (Mensch)					
	a_1^i	a_2^i	a_1^j	a_2^j	$a_1^i a_1^j$	$a_2^i a_2^j$	$a_1^i a_2^j$	$a_2^i a_1^j$
Spielertyp								
- AgentMx1	57%	43%	25%	75%	4%	21%	54%	22%
- AgentM01	50%	50%	28%	72%	5%	27%	45%	24%
- AgentM11	58%	42%	24%	76%	4%	22%	54%	19%
- Mensch	31%	69%	28%	72%	2%	43%	29%	26%
Normierte Auszahlung								
- Spieler					0	20	60	100
- Gegner (Mensch)					0	20	100	60
Eigenschaft							N	N

beste Zustand $a_2^i a_1^j$ insgesamt bei nur 4% seltener als beim Mensch erreicht. Dennoch konnte der zweitbeste Zustand ebenfalls mir 54% realisiert werden. In Summe erreichte AgentMx1 die besten beiden Zustände in 76% der Runden, wohingegen Menschen dies nur in 55% der Runden erwirtschaften konnten. Interessant ist, dass das Verhalten der *menschlichen Gegenspieler* im Vergleich zu den Ergebnissen zu Experiment I (Tabelle 5.12) und Experiment III (Tabelle 5.14) eine gewisse Beständigkeit über die Spielertypen hinweg aufweist. Dies deutet darauf hin, dass die Angst vor Bestrafung im koordinativen Hero Game geringer sein könnte, als in den konfliktären Spielen Prisoner's Dilemma und Chicken Game.

Die in Tabelle 5.14 veranschaulichte Analyse der von den verschiedenen Spielertypen erreichten Spielzustände in Experiment III zum *wiederholten Prisoner's Dilemma*. Beobachtete *menschliche Spieler* zeigen sich im Vergleich mit Prestudy II (siehe Tabelle 5.2) ähnlich kooperativ, wenn auch Aktion a_1^i im Schnitt 4% häufiger gespielt wurde. Auffällig ist, dass die Zustände $a_1^i a_1^j$ und $a_2^i a_1^j$ für menschliche Spieler in Experiment III mit identischer Häufigkeit wie in Prestudy II erreicht wurden. Auch die Zustände $a_1^i a_2^j$ und $a_2^i a_2^j$ wurden vergleichbar of erreicht, wobei die Summe der Häufigkeit der beiden Zustände identisch mit Prestudy II verbleibt. Im Vergleich zwischen den Markovagenten fällt auf, dass *AgentM01* mit 61% am seltensten kooperiert, sodass er unter den nicht-menschlichen Spielern den geringsten Anteil der Kooperationslösung $a_1^i a_1^j$ realisieren konnte. Als Kehrseite dieses Verhaltens konnte AgentM01 jedoch am häufigsten den Maximalpayoff in $a_2^i a_1^j$ realisieren und wurde am seltensten vom Gegenspieler

Tabelle 5.14: Verteilung der Aktionswahl und der erreichten Spielzustände der Spielertypen nach aufsteigend sortierter normierter Auszahlung in Experiment III zum Prisoner's Dilemma; mit Nash-Gleichgewicht des Stufenspiels (N), paretoeffizientem Zustand (grüne Farbe) und paretodominiertem Zustand (rote Farbe); exemplarisch um Kooperation (C) und Abweichung (D) ergänzt. Quelle: Eigene Darstellung.

	Aktionswahl				Realisierter Spielzustand			
	Spieler		Gegner (Mensch)					
	a_1^i	a_2^i	a_1^j	a_2^j	$a_1^i a_2^j$	$a_2^i a_2^j$	$a_1^i a_1^j$	$a_2^i a_1^j$
	C	D	C	D	CD	DD	CC	DC
Spielertyp								
- AgentMx1	78%	22%	71%	29%	12%	17%	66%	5%
- AgentM01	61%	39%	64%	36%	5%	31%	56%	8%
- AgentM11	73%	27%	65%	35%	15%	20%	58%	7%
- Mensch	63%	37%	60%	40%	10%	30%	53%	7%
- Tit-for-Tat	79%	21%	77%	23%	6%	17%	73%	5%
Normierte Auszahlung								
- Spieler					0	20	60	100
- Gegner (Mensch)					100	20	60	0
Eigenschaft						N		

in $a_1^i a_2^j$ ausgebeutet. Vergleichbar dazu weist auch *AgentM11*mit 58% nur ein leicht höheren Anteil der Kooperationslösung $a_1^i a_1^j$ auf und auch der beste Zustand $a_2^i a_1^j$ wurde nur 1% seltener realisiert. Jedoch konnte sich AgentM11 im Gegensatz zu AgentM01 weitaus weniger effektiv gegen eine Ausbeutung im Zustand $a_1^i a_2^j$ mit einer Auszahlung von 0 schützen. Im Schnitt wurde AgentM11 mit 15% dreimal so oft ausgenutzt wie AgentM01. Zwar weist *AgentMx1* mit 12% einen nur leicht besseren Schutz gegen Ausbeutung vor, kann aber in Relation zu den anderen beiden Markovagenten mit 66% unter diesen mit Abstand am häufigsten die Kooperationslösung $a_1^i a_1^j$ verhandeln und konnte ein häufiges beidseitiges Abweichen $a_2^i a_2^j$ so vermeiden. Diese Leistung wird nur vom *Tit-for-Tat-Spieler* übertroffen, der zwar gleich oft in $a_2^i a_2^j$ verweilte, aber dennoch nur halb so oft ausgebeutet wurde und dennoch noch häufiger die Kooperationslösung erzielte. Die Auswirkungen auf den erzielten normierten Payoff werden im nächsten Kapitel betrachtet.

5.2.2.2 Normierte Auszahlung

Tabelle 5.15 zeigt statistische Kennzahlen zu den durchschnittlichen normierten Auszahlungen der Spielertypen in den Experimenten I bis III. Im wiederholten *Chicken Game* (Experiment

Tabelle 5.15: Deskriptive Auswertung der durchschnittlichen normierten Auszahlungen $\bar{\rho}^i$ der beobachteten Spielertypen in Experimenten I bis III zum Chicken Game (CG), Hero Game (HG) und Prisoner's Dilemma (PD). Quelle: Eigene Darstellung.

				Perzentile				
Spiel	Spielertyp	n	$\bar{x}$	25%	$\tilde{x}$	75%	s	$\frac{s}{\bar{x}}$
CG	AgentMx1	41	37.98	20.00	39.05	54.29	16.56	0.44
	AgentM01	42	41.22	25.71	43.81	54.29	16.81	0.41
	AgentM11	42	42.11	29.52	45.71	58.10	15.77	0.37
	Mensch	42	33.99	15.24	38.10	48.57	19.63	0.58
HG	AgentMx1	45	58.05	54.29	56.19	64.76	11.83	0.20
	AgentM01	45	56.02	47.62	56.19	63.81	11.05	0.20
	AgentM11	45	56.19	52.38	56.19	60.00	10.41	0.19
	Mensch	45	51.81	31.43	54.29	72.38	21.44	0.41
PD	AgentMx1	43	47.73	37.14	60.00	60.00	15.86	0.33
	AgentM01	43	47.49	34.29	53.33	60.00	15.38	0.32
	AgentM11	43	45.98	28.57	57.14	60.00	17.19	0.37
	Mensch	43	44.94	27.62	52.38	60.00	16.09	0.36
	Tit-for-Tat	43	51.65	41.90	60.00	60.00	13.52	0.26

I) weisen die Markovspieler im Mittel einen höheren durchschnittlichen normierten Payoff aus als menschliche Spieler. Insbesondere wurde die Leistung der Markovagenten im Vergleich mit menschlichen Spielern bei einer deutlich geringeren Standardabweichung konsistenter erzielt. Von den nicht-menschlichen Spielern weist AgentMx1 ist aufgrund der wie im vorigen Kapitel beschriebenen häufigen Realisierung der geringstauszahlenden Zustände den geringsten Payoff aus (siehe Tabelle 5.13). Im wiederholten *Hero Game* (Experiment II) weisen die Markovagenten im Mittel einen höheren durchschnittlichen normierten Payoff aus als menschliche Spieler. Die von den Markovagenten erreichten Zustände (siehe Tabelle 5.13) deuten darauf hin, dass sich diese besser mit dem Gegenspieler koordinieren konnte. Auffällig ist außerdem die im Vergleich zu menschlichen Spieler nur circa halb so große Standardabweichung der Payoffs. Im wiederholten *Prisoner's Dilemma* (Experiment III) wurden mit Prestudy II (siehe Tabelle 5.3) konsistente Ergebnisse beobachtet. Die Markovagenten schnitten im Mittel besser ab als die menschlichen Spieler. Wie auch in Prestudy I (siehe Tabelle C.3) konnte Tit-for-Tat im Mittel die besten Ergebnisse erzielen. Die Relation der Medianwerte zwischen den Spielertypen gestaltet sich analog.

Zusammenfassend konnten in Abhängigkeit des Spiels deskriptiv homogene Ergebnisse bezüglich der relativen Leistungsfähigkeit der Markovagenten beobachtet werden, welche im Mittel in allen Spielen besser abschnitten als Menschen der Referenzgruppe. Über die Varianten der Markovagenten hinweg gibt es Unterschiede zwischen den Spielen. Im Chicken Game stellte sich AgentM11 als am performantesten heraus, wohingegen AgentMx1 im Hero Game und knapp im Prisoner's Dilemma besser abschließen konnte. Nachfolgend sollen diese Unterschiede anhand von Hypothesentests untersucht werden.

5.2.3 Hypothesentests

Im Folgenden werden die Ergebnisse von Experiment I bis III zur weiterführenden Validierung der Markovagenten anhand von Hypothesentests präsentiert. Der zu zeigende Sachverhalt ist eine signifikante positive Abweichung der Leistung des Markovagenten im Vergleich der von menschlichen Spielern. Ausgangsbasis für sämtliche Tests ist die Datenlage aus Tabelle 5.15.

5.2.3.1 Zentrale Analysen

Die zugrundeliegenden Überlegungen zu den Hypothesentests gestalten sich analog zu Prestudy II (siehe Kapitel 5.1.3.1). Daher wird der zwei-Stichproben t-Tests aufgrund der Abhängigkeit der Beobachtungen innerhalb der Spielerkohorten als unpassend eingestuft (vgl. Toutenburg & Heumann, 2008, S. 142-145). Ein gepaarter t-Test als parametrische Alternative zur Betrachtung der Payoffdifferenz zwischen Spielertypen ist nur eingeschränkt möglich, da die Annahme einer Normalverteilung der Differenzen teilweise verletzt wird (vgl. Toutenburg & Heumann, 2008, S. 145-147). Tabelle A.6 (siehe Anhang) stellt diesen Sachverhalt auf Basis des Shapiro-Wilk-Tests (vgl. Royston, 1992) mit Anpassung durch Shapiro und Wilk (1965) dar. Die Nullhypothese, dass bei der Differenz der Auszahlungswerte eine Normalverteilung vorliegt konnte bei einem Signifikanzniveau von 5% für eine Vielzahl paarweiser Vergleiche insbesondere im Prisoner's Dilemma verworfen werden.

Als nicht-parametrische Alternative wird der Mann-Whitney-U-Test (Mann & Whitney, 1947) gleich dem zwei-Stichproben t-Test aufgrund der fehlenden Unabhängigkeit der Beobachtungen ausgeschlossen (vgl. Cleff, 2019, S. 181-185). Der Wilcoxon-Vorzeichen-Rang-Test (Wilcoxon, 1945, 1947) kommt bei Erfüllung der Symmetrieannahme für die Auszahlungsdifferenzen in Frage (vgl. Toutenburg & Heumann, 2008, S. 182). Im Symmetrietest von D'Agostino et al. (1990) mit Anpassung von Royston (1991) legt die Nullhypothese eine Symmetrie der Differenzen der paarweise nach Sparringspielern gruppierten Payoffs zugrunde. Die Ergebnisse in Tabelle A.7 (siehe Anhang) konnten Verletzungen der Annahme aufzeigen, jedoch beschränken diese sich auf den nachrangigen Vergleich zwischen nichtmenschlichen Spielern. Die vordergründige Frage, wie nichtmenschliche im Vergleich zu menschlichen Spielern abschneiden ver-

bleibt im Einklang mit der Annahme, sodass der Wilcoxon-Vorzeichen-Rang-Test angewendet werden kann.

Tabelle 5.16: Ergebnisse des Wilcoxon-Vorzeichen-Rang-Tests (Wilcoxon, 1945) für gepaarte Daten der Experimente I bis III zum Chicken Game (CG), Hero Game (HG) und Prisoner's Dilemma (PD) auf H_0, dass zwei Stichproben aus der gleichen Verteilung gezogen wurden und demnach der Median der Auszahlungsdifferenzen $D = X_1 - X_2$ Null ist (grüne Kennzeichnung für positive, rote für negative Auszahlungsdifferenzen; indikativ: † $p < 10\%$; signifikant: * $p < 5\%$, ** $p < 1\%$, *** $p < 0.1\%$) unter Berücksichtigung der Erfüllung der zugrundeliegenden Symmetrieannahme (SA) aus Tabelle A.7. Quelle: Eigene Darstellung.

	Vergleich								
Spiel	Spielertyp 1	Spielertyp 2	$\bar{D}$	$\tilde{D}$	z	p		H_0	SA
CG	AgentMx1	Mensch	3.28	1.90	1.02	30.9%		✓	✓
	AgentM01	Mensch	7.23	4.76	2.00	4.6%	*	×	✓
	AgentM11	Mensch	8.12	6.67	2.56	1.1%	*	×	✓
HG	AgentMx1	Mensch	6.24	1.90	1.55	12.1%		✓	✓
	AgentM01	Mensch	4.21	4.76	1.16	24.7%		✓	✓
	AgentM11	Mensch	4.38	4.76	1.17	24.3%		✓	✓
PD	AgentMx1	Mensch	2.79	0.00	1.80	7.2%	†	✓	✓
	AgentM01	Mensch	2.55	0.00	0.83	40.6%		✓	✓
	AgentM11	Mensch	1.04	0.00	0.50	61.8%		✓	✓
	Tit-for-Tat	Mensch	6.71	3.81	3.01	0.3%	**	×	✓

Tabelle 5.16 zeigt die Ergebnisse des Wilcoxon-Vorzeichen-Rang-Tests zur Nullhypothese, dass der Median der Auszahlungsdifferenzen im paarweisen Vergleich Null ist. Für das *Chicken Game* konnte eine signifikant überlegene Leistung von AgentMx1 und AgentM01 über Menschen nachgewiesen werden. Die Leistungsdifferenzen im *Hero Game* fallen wiederum nicht-signifikant aus. Im *Prisoner's Dilemma* zuletzt, wurde eine signifikant überlegene Leistung des Tit-for-Tat-Spielers über Menschen gezeigt. Weiterhin gibt der Wilcoxon-Rang-Vorzeichen-Test an, dass der Tit-for-Tat-Spieler dem AgentM11 signifikant überlegen war. Jedoch ist für diesen Vergleich die notwendige Symmetrieannahme nicht erfüllt. Die Ergebnisse des alternativen Vorzeichentests (Arbuthnott, 1710; Snedecor & Cochran, 1991, vgl.), der keine Symmetrie in der Payoffdifferenz voraussetzt, bestätigt die Ergebnisse (siehe Tabelle A.11 im Anhang).

Zusammenfassend lässt sich festhalten, dass die Leistung der Markovagenten in jedem der Experimente direktional den menschlichen Spielern überlegen ist. Eine signifikante Abweichung konnte jedoch im Rahmen der Hypothesentests nur für das Chicken Game gezeigt werden. Im Anschluss an unterstützende Robustheitsanalysen zu den Hypothesentests schließt sich ei-

ne ergänzende Regressionsanalyse an, welche Interdependenzen zwischen dem Verhalten des Gegenspielers und der gemessenen Leistung sowie Lerneffekte berücksichtigt.

5.2.3.2 Zusätzliche Robustheitsanalysen

Die Ergebnisse ergänzender Untersuchungen (siehe Tabellen A.6 bis A.11 im Anhang) untermauern die Argumentation bezüglich der Annahmenerfüllung für verwendete Hypothesentests in Kapitel 5.2.3.1, beziehungsweise ergänzen deren Aussagen anhand ausgeschlossener Verfahren.

Tabelle A.6 untersucht zunächst das Vorliegen einer Normalverteilung für Differenz der gepaarten Auszahlungswerte anhand des Shapiro-Wilk-Tests vgl., Royston (1992), Shapiro und Wilk (1965). Diese konnte vor allem im Vergleich zwischen verschiedenen Markovagenten, aber auch generell häufig im Prisoner's Dilemma verworfen werden, sodass ein gepaarter t-Test für die betreffenden Vergleiche nur eingeschränkt aussagekräftig ist. Tabelle A.7 hingegen untersucht die Differenzen auf Symmetrie (D'Agostino et al., 1990; Royston, 1991). Die Symmetrieeigenschaft konnte weitestgehend nicht verworfen werden; lediglich im Vergleich zwischen Markovagenten wurde eine Verletzung gezeigt.

Analog zu Kapitel 5.1.3.2 über Prestudy II werden die Ergebnisse der Testhypothesen zu Experiment I bis III durch Anwendung der in Kapitel 5.2.3.1 verworfenen Tests bestätigt. Tabelle A.8 stellt den zwei-Stichproben t-Test dar, während Tabelle A.9 den gepaarten t-Test präsentiert (Student, 1908). Tabelle A.10 hingegen zeigt die Ergebnisse des Mann-Whitney-U-Tests (Mann & Whitney, 1947), während Tabelle A.11 die Ergebnisse des Vorzeichentests präsentiert (Arbuthnott, 1710; Snedecor & Cochran, 1991, vgl.). Die Ergebnisse der Robustheitsanalysen unterstreichen die Validität der zentralen Testhypothesen.

5.2.4 Regressionsanalyse

Im Folgenden werden die Ergebnisse von Experiment I bis III zur weiterführenden Validierung des Markovagenten anhand von Regressionsanalysen präsentiert. Der zu zeigende Sachverhalt ist eine signifikante positive Abweichung der Leistung des Markovagenten im Vergleich der von menschlichen Spielern unter Berücksichtigung von Einflüssen wie Lerneffekten auf Seiten der menschlichen Spieler sowie die Abhängigkeit der Leistung vom Spielverhalten des Gegenspielers.

5.2.4.1 Zentrale Analysen

Die Daten von Experiment I bis III zu den wiederholten Spielen *Chicken Game*, *Hero Game* und *Prisoner's Dilemma* werden im Rahmen einer gemeinsamen Panelregression analysiert, welche

sich in Teilen an der Modellierungslogik von Prestudy II orientiert (siehe Kapitel 5.1.4). Weiterhin werden ebenfalls die Daten aus Prestudy II zum Prisoner's Dilemma mit in den Datensatz einbezogen. Dies bietet sich aufgrund der konsistenten Laborbedingungen und konsistenten Konfiguration der Markovagenten an, um so die Datengrundlage zu verbessern. Infolgedessen wird der durchschnittliche normierte Payoff als *abhängige Variable* (siehe Kapitel 4.4) und der jeweilige menschliche Gegenspieler als *Panelvariable* (siehe Kapitel 4.1.2.1 sowie Tabelle 5.10 und 5.11) herangezogen. Dennoch sind aufgrund der integrierten Betrachtung der Spiele Anpassungen erforderlich. Neben den bisherigen erklärenden *Kontrollvariablen* des betrachteten Spielertyps und der Erfahrung des Gegenspielers (siehe Kapitel 5.1.4.1), soll außerdem der Spieltyp berücksichtigt werden. Infolgedessen wird in Summe für folgende Effekte kontrolliert:

1. Wie beeinflusst der *Spieltyp* die Leistung des beobachteten Spielers?
2. Wie beeinflusst der *Spielertyp* des beobachteten Spielers dessen Leistung?
3. Wie beeinflussen temporale *Lerneffekte* im Sinne der Erfahrung des menschlichen Gegenspielers die Leistung des beobachteten Spielers?

Wie auch in Prestudy II werden Spieltyp und Spielertyp über Dummyvariablen abgebildet, während es sich bei der Anzahl der Spiele des menschlichen Gegenspielers um eine kontinuierliche Zählvariable handelt.

Weiterhin wird für das beschriebene Vorgehen als *Modelltyp* ein Random Effects Modell bevorzugt. Qualitative Grundlage der Gestaltungsentscheidung ist, dass ein Fixed Effects Modell nicht in der Lage ist, den Spieltyp als Kontrollvariable zu berücksichtigen (siehe Tabelle A.12), da dieser innerhalb der Panelvariable invariant ist (vgl. Cameron & Trivedi, 2010; Kohler & Kreuter, 2017). Das Vorgehen wird durch die verwendeten quantitativen Spezifikationstests in Tabelle 5.20 nicht eindeutig widerlegt.

Das insgesamt signifikante Modell in Tabelle 5.17 zeigt in Bezug auf den *Spielertyp*, dass die Leistung der Markovagenten in Bezug auf durchschnittlichen normierten Payoff der menschlichen Leistung für die integrierte Panelregression über Prestudy II sowie Experiment I bis III signifikant überlegen ist. Konkret schneidet AgentMx1 um 4.03, AgentM01 um 4.62 und AgentM01 um 4.18 besser als die Menschen ab. Der Tit-for-Tat-Spieler, welcher lediglich in Experiment III eingesetzt wurde, übertrifft die menschliche Leistung sogar um 7.82 Punkte. Weiterhin zeigt die Regression, dass signifikante Unterschiede zwischen den *Spieltypen* vorliegen. Relativ zum Chicken Game erzielen die Spieler im Hero Game demnach im Schnitt 16.58 Punkte und im Prisoner's Dilemma 7.88 Punkte mehr. Auch die Präsenz von *Lerneffekten* über die Spiele konnte signifikant gezeigt werden. So erzielen die Spieler mit jedem Spiel, welches der Gegenspieler bereits absolvieren konnte 2.01 Punkte mehr.[57]

[57] Die Lerneffekte werden anhand des Gegenspielers bemessen, da dieser stets ein Mensch ist. Würden die Lerneffekte auf den beobachteten Menschen bezogen, käme es zu einer Verzerrung der Regression zu Gunsten

Tabelle 5.17: Integrierte Ergebnisse des Random Effects Panelregressionsmodells zu Prestudy II und Experiment I bis III im wiederholten Chicken Game, Hero Game und Prisoner's Dilemma (indikativ: † $p < 10\%$; signifikant: * $p < 5\%$, ** $p < 1\%$, *** $p < 0.1\%$). Quelle: Eigene Darstellung.

	Koeffizient	Standardfehler	z	p	
Spielertyp					
- Mensch (Referenz)	–				
- AgentMx1	4.03	1.18	3.41	0.1%	***
- AgentM01	4.62	1.23	3.76	0.0%	***
- AgentM11	4.18	1.34	3.13	0.2%	**
- Tit-for-Tat (nur PD)	7.82	2.05	3.82	0.0%	***
Spieltyp					
- Chicken Game (Referenz)	–				
- Hero Game	16.58	2.51	6.62	0.0%	***
- Prisoner's Dilemma	7.88	2.22	3.54	0.0%	***
Lerneffekte: Anzahl Spiele (Gegner)	2.01	0.29	7.04	0.0%	***
Konstante	29.67	2.13	13.92	0.0%	***
Beobachtungen	718				
Gruppen	169				
R^2_{within}	0.12				
$R^2_{between}$	0.21				
$R^2_{overall}$	0.17				
Wald $\chi^2(7)$	118.26				
Signifikanz ($\mathbf{P} > \chi^2$)	0.0%				***

Eine differenzierte Aussage über die Leistung der Spielertypen in Bezug auf die gespielten Spieltypen lässt sich anhand des Modells in Tabelle 5.17 nicht treffen. Unter Einbezug von *Interaktionseffekten* kann dieser Aspekt adressiert werden (vgl. Aiken & West, 1991). Dementsprechend findet eine Anpassung der Kontrollvariablen zur Berechnung eines Panelmodells mit Interaktion zur Beantwortung folgender Fragestellungen statt:

1. Wie beeinflusst der *Spieltyp* die Leistung des beobachteten Spielers?

der Markovagenten. Hintergrund ist, dass ein derartiger Lerneffekte zwar Varianz in der Leistung der beobachteten Menschen erklärt, dies jedoch nicht auf die Markovspieler übertragen werden kann, da letztere zu Beginn jedes Spiels zurückgesetzt werden. Infolgedessen würden die Koeffizienten nichtmenschlicher Spieler steigen. Aufgrund der Interdependenz des Spielergebnisses vom Gegenspieler wird die Betrachtung der Lerneffekte des Gegners als hinreichende Näherung herangezogen.

2. Wie beeinflusst der *Spielertyp* des beobachteten Spielers dessen Leistung *je Spieltyp*?

3. Wie beeinflussen temporale *Lerneffekte* im Sinne der Erfahrung des menschlichen Gegenspielers die Leistung des beobachteten Spielers *je Spieltyp*?

Gleichbleibend zur vorangegangen Regression wird ein Random Effects Modell zur Berücksichtigung invarianter Effekte des Spieltyps mit einzubeziehen (Fixed Effects Modell siehe Tabelle A.13). Die Verwendung eines Random Effects Modells wird durch diverse Spezifikationstests in Tabelle 5.21 bestätigt.

Tabelle 5.18 und 5.19 zeigen die Ergebnisse des signifikanten Regressionsmodells mit Interaktionseffekten zwischen dem Spieltyp und dem Spielertyp sowie dem Spieltyp und dem Lerneffekt zur Erklärung des erzielten normierte Payoffs. Die isolierten Effekte bezüglich des *Spieltyps* unterscheiden sich von den Ergebnissen des Modells ohne Interaktion (siehe Tabelle 5.17). Der weiterhin signifikante Unterschied zwischen dem Hero Game und dem Chicken Game fällt mit 13.23 geringer aus. Dahingegen weist der Unterschied zwischen dem Prisoner's Dilemma und dem Chicken Game bei vergleichbar geringer Effektstärke von 0.73 keine Signifikanz mehr auf. Bei nach Spieltyp gruppierten *Lerneffekten* zeigt sich ein differenziertes Bild. Im Chicken Game konnten keine signifikanten Effekte gezeigt werden. Im Hero Game nimmt die erzielte Auszahlung jedoch mit der Erfahrung des Gegners signifikant um 1.33 Punkte je Spiel zu. Im Prisoner's Dilemma wird der Zuwachs mit 3.06 als hochsignifikant geschätzt.[57] In Bezug auf *Spielertypen* kondensiert sich mit durchgehend positiver Effektstärke eine direktionale Überlegenheit der Markovagenten in allen Spieltypen, jedoch mit heterogenen Signifikanzniveaus. Im Chicken Game sind AgentM01 mit 7.18 Punkten und AgentM11 mit 8.10 Punkten den Menschen klar signifikant überlegen. AgentMx1 konnte trotz einem Leistungsvorsprung von 3.90 keine Signifikanz erzielen. Im Hero Game hingegen schneidet AgentMx1 mit 6.21 Punkten signifikanten Vorsprungs gegenüber den Menschen ab. AgentM01 und AgentM11 verpassen mit 4.12 und 4.35 Punkten Differenz zu menschlichen Spielern nur knapp das Erreichen eines Signifikanzniveaus. Im Prisoner's Dilemma wiederum leistete AgentMx1 mit 3.42 Mehrpunkten signifikant besser als die Menschen. AgentMx1 ist mit 2.83 Punkten Vorsprung zu menschlichen Spielern knapp nicht signifikant überlegen, wohingegen AgentM11 mit nur 0.82 Punkten als klar nicht signifikant eingestuft wird. Der als Benchmark geprüfte Tit-for-Tat-Spieler leistet eine den Menschen um 6.28 Punkten signifikant überlegene Leistung.

Zusammenfassend zeigen die integrierten Regressionsmodelle zum wiederholten Chicken Game, Hero Game und Prisoner's Dilemma eine tendenziell überlegene Leistung der Markovagenten relativ zu menschlichen Vergleichsspielern. Die Regression ohne Interaktionseffekte in Tabelle 5.17 weist dieses Ergebnis deutlicher aus, als jene mit Interaktionseffekten in Tabelle 5.18. Hierbei sei jedoch darauf verwiesen, dass auch bei Berücksichtigung der Interaktionseffekte stets eine positive Effektstärke zugunsten der Markovagenten vorliegt. Zumeist sind diese Ef-

Tabelle 5.18: Integrierte Ergebnisse des Random Effects Panelregressionsmodells mit Interaktionseffekten zu Prestudy II und Experiment I bis III (indikativ: † $p < 10\%$; signifikant: * $p < 5\%$, ** $p < 1\%$, *** $p < 0.1\%$). Quelle: Eigene Darstellung.

	Koeffizient	Standardfehler	z	p	
Spielertyp im Chicken Game					
- Mensch (Referenz)	–				
- AgentMx1	3.90	2.44	1.60	11.0%	
- AgentM01	7.18	2.42	2.97	0.3%	**
- AgentM11	8.10	2.42	3.35	0.1%	***
Spielertyp im Hero Game					
- Mensch (Referenz)	–				
- AgentMx1	6.21	2.33	2.66	0.8%	**
- AgentM01	4.12	2.33	1.77	7.7%	†
- AgentM11	4.35	2.33	1.86	6.2%	†
Spielertyp im Prisoner's Dilemma					
- Mensch (Referenz)	–				
- AgentMx1	2.83	1.60	1.78	7.6%	†
- AgentM01	3.42	1.73	1.98	4.8%	*
- AgentM11	0.82	2.16	0.38	70.3%	
- Tit-for-Tat	6.28	2.16	2.91	0.4%	**
Spieltyp					
- Chicken Game (Referenz)	–				
- Hero Game	13.23	4.18	3.17	0.2%	**
- Prisoner's Dilemma	0.73	3.67	0.20	84.1%	
Lerneffekte: Anzahl Spiele (Gegner)					
- Chicken Game	−0.21	0.63	−0.33	74.5%	
- Hero Game	1.33	0.60	2.22	2.6%	*
- Prisoner's Dilemma	3.06	0.37	8.29	0.0%	***
Konstante	34.62	3.03	11.42	0.0%	***

Tabelle 5.19: Ergänzende Informationen zum integrierten Random Effects Panelregressionsmodell mit Interaktionseffekten zu Prestudy II und Experiment I bis III in Tabelle 5.18 (indikativ: † $p < 10\%$; signifikant: * $p < 5\%$, ** $p < 1\%$, *** $p < 0.1\%$). Quelle: Eigene Darstellung.

Beobachtungen	718	
Gruppen	169	
R^2_{within}	0.17	
$R^2_{between}$	0.22	
$R^2_{overall}$	0.19	
Wald $\chi^2(15)$	151.15	
Signifikanz ($\mathbf{P} > \chi^2$)	0.0%	***

fekte weiterhin nur knapp unter dem Signifikanzniveau von 5%. Ein Grund hierfür könnte die reduzierte Effizienz des Modells mit Interaktionseffekten sein, da mit identischer Stichprobe statt 7 Koeffizienten 15 Koeffizienten geschätzt werden müssen.

Fazit

Die Markovagenten können in weiterführenden Experimenten zum wiederholten Chicken Game, Hero Game und Prisoner's Dilemma konsistent direktional bessere und häufig signifikant bessere Ergebnisse als menschliche Referenzspieler erzielen.

5.2.4.2 Zusätzliche Robustheitsanalysen

Im Folgenden werden Spezifikationstests zum Regressionsmodell *ohne* Interaktion (siehe Tabelle 5.17) und *mit* Interaktion (siehe Tabelle 5.18) präsentiert. Bezüglich des Modells ohne Interaktion geben die Spezifikationstests in Tabelle 5.20 keine eindeutige Empfehlung zur Verwendung eines Random oder Fixed Effects Modells. Für das Modell mit Interaktion bestätigen Spezifikationstests in Tabelle 5.20 die Verwendung eines Random Effects einstimmig.

Tabelle 5.20: Integrierte Ergebnisse von Tests zur Modellspezifikation der Panelregression ohne Interaktionseffekte zu Prestudy II und Experiment I bis III (* $p < 0.1$, ** $p < 5\%$, *** $p < 1\%$). Quelle: Eigene Darstellung.

Test	Statistik	p		Empfohlenes Modell
Hausman Test	$\chi^2(5) = 1.06$	95.8%		Random Effects
Robuster Hausman Test	$F(3,168) = 0.39$	75.7%		Random Effects
Sargan-Hansen Test	$\chi^2(4) = 10.89$	2.8%	**	Fixed Effects
Robuster Sargan-Hansen Test	$\chi^2(4) = 9.19$	5.7%	*	Fixed Effects

Tabelle 5.21: Integrierte Ergebnisse von Tests zur Modellspezifikation der Panelregression mit Interaktionseffekten zu Prestudy II und Experiment I bis III (* $p < 0.1$, ** $p < 5\%$, *** $p < 1\%$). Quelle: Eigene Darstellung.

Test	Statistik	p	Empfohlenes Modell
Hausman Test	$\chi^2(13) = 1.24$	100.0%	Random Effects
Robuster Hausman Test	$F(5,168) = 0.71$	61.9%	Random Effects
Sargan-Hansen Test	$\chi^2(5) = 1.59$	90.3%	Random Effects
Robuster Sargan-Hansen Test	$\chi^2(5) = 1.68$	89.1%	Random Effects

Die Ergebnisse der ergänzenden Regressionsanalysen finden sich im Anhang (sieh Tabellen A.12 bis A.17). Die Modellspezifikation ist der Validität der Regression ohne Interaktion (siehe Tabelle 5.17) nicht abträglich, wie die in Tabelle A.12 dargestellten Ergebnisse einer Fixed Effects Regression zeigen. Auch die Robustheit der Ergebnisse mit Interaktion (siehe Tabelle 5.18) wird anhand alternativer Modelle in Tabellen A.13 und A.14 sichergestellt. Ordinary Least Squares (OLS) Regressionen, welche den Einfluss des Gegenspielers als Panelvariable vernachlässigen runden das Bild mit und ohne Interaktion ab (siehe Tabellen A.15 bis A.17).

6 Diskussion

Das Kapitel setzt sich zum Ziel, die gewonnenen Erkenntnisse kritisch mit Bezug auf die in Kapitel 2.4 festgelegten Forschungsfragen zu diskutieren. Dabei findet zunächst eine Bewertung der formalen Validität nach Powers und Shoham (2005a) des entwickelten Markovagenten statt. Die Bewertung nach Formalkriterien stellt sicher, dass der etwickelte Markovagent konzeptionellen Gestaltungsanforderungen des Multi Agent Learning Kontextes genügt. Da eine derartige theoretische Untersuchung gegebenenfalls ein nur unvollständiges Leistungsbild geben kann, wird diese durch eine empirische Untersuchung angereichert. Daher wird im Anschluss die experimentelle Validierung thematisiert. Diese setzt sich aus einer ergänzenden Untersuchung zu einem Turnier mit algorithmischen Strategien nach Axelrod (1980) und einer Reflexion über die Ergebnisse der Untersuchungen mit Menschen (siehe Kapitel 5) zusammen. Die Ergebnisse der Turniere gegen algorithmische Gegner sind in Tabellen 6.3 sowie A.20 und A.21 zu finden. Es sei hervorgehoben, dass das Spiel gegen algorithmische und menschliche Gegner grundlegend verschiedene ist. Beispielsweise kann der zu entwickelnde Agent im Spiel gegen Algorithmen mit Kenntnis der Aktionslogik der zu schlagenden Algorithmen Informationsvorteile ausnutzen. Weiterhin wird so geährleistet, dass die Interaktionslogik des zu entwickelnden Agenten eine gewisse Relevanz aufweißt. Ein KI-Agent, der die komplexe und undurchsichtige Spielweise von Menschen schlagen kann, vermag darauf aufbauende Erkenntnisse über Mensch-Maschine-Interaktion zu fördern. Zu guter Letzt sollen die Ergebnisse zu einer Gesamtbetrachtung synthetisiert werden.

6.1 Diskussion der formalen Validität: Sind die Formalkriterien an MAL Agenten für AgentM erfüllt?

Nachfolgend wird die Erfüllung der Formalkriterien für nichtkooperativ präskriptive MAL Agenten von Powers und Shoham (2005a) geprüft, welche (1) das Erzielen einer ε-besten Antwort gegen eine Zielgegnerklasse und gleichzeitig (2) die Garantie einer Mindestauszahlung gegen jeden Gegner sowie (3) eine angemessene Leistung im Selbstspiel erfordern (siehe Kapitel 2.2.3). Zusammenfassend ergibt die Untersuchung, dass der Markvoagent Bedingung (1) erfüllt. Bedingung (2) wird aufgrund einer Designentscheidung nicht erfüllt, kann aber durch Verwenden eines *Sicherheitsmoduls* problemlos erreicht werden. Bezüglich Bedingung (3) kön-

nen spezifische Parametrisierungen des Markovagenten existieren, die bei bestimmten Spielen die definierte Leistungsgrenze im Selbstspiel nicht erreichen.

6.1.1 Gezielte Optimalität

Das Finden einer besten Antwort ohne die Einschränkung der Menge möglicher Gegner ist kaum möglich. Infolgedessen muss sich die Angemessenheit einer MAL Lösung stets nach dem angedachten Anwendungskontext richten (vgl. Powers & Shoham, 2005b). Dementsprechend besteht die Zielklasse des AgentM aus Markovspielern mit beschränktem endlichen Gedächtnis im Sinne von Gegnern, deren gemischte oder reine Strategien auf einer Partition der kürzlichen Spielhistorie bedingen. Es sei angemerkt, dass diese Menge insbesondere alle *deterministischen* finiten Automaten mit Aktionsfolgen als Zustände analog zur Zielgegnerklasse von Carmel und Markovitch (1998) enthält, aber darüber hinaus auch derartige *stochastischen* finiten Automaten beinhaltet.

Auf Basis dieser Spielerklasse soll die Erfüllung der *gezielten Optimalität* im Sinne des Findens einer Antwort-Strategie, welche langfristig maximal ε von der durchschnittlichen Auszahlung der besten Antwort gegen die tatsächliche Gegnerstrategie abweicht. Während es einfach ist, die beste Antwort für eine bekannte bedingte Gegnerstrategie zu finden, ist es umso schwieriger diese für eine konkrete aber unbekannte bedingte Gegnerstrategie oder gar eine Menge möglicher bedingter Gegnerstrategien zu finden. Powers und Shoham (2005a) gestatten bei der Bewertung der gezielten Optimalität daher insbesondere folgende zwei Schlupflöcher:

1. Die Verwendung einer beliebig langen und nicht gewerteten *Explorationsphase*.
2. Mögliche Negativeffekte, welche aus *Pfadabhängigkeiten* der Explorationsphase entstehen können, sind dem Begriff der gezielten Optimalität nicht abträglich.

Nachfolgend soll die Notwendigkeit dieser beiden Ausnahmen für die gezielte Optimalität diskutiert werden. Die grundlegendste Erklärung ist, dass eine ε-beste Antwort auf die *tatsächliche* Gegnerstrategie gefunden werden soll, während andere Kriterien lediglich die beste Antwort auf die empirische Verteilung des Gegnerverhaltens erfordern (vgl. Fudenberg & Levine, 1995, 1998). Insofern ist die Bildung eines hinreichend guten Gegnermodells Voraussetzung eines MAL, der das Kriterium der gezielten Optimalität erfüllt.

Die Verwendung einer beliebig langen und nicht gewerteten *Explorationsphase* soll dem untersuchten MAL Algorithmus ermöglichen, die tatsächliche Gegnerstrategie dementsprechend zu modellieren. Die erforderliche Länge der Explorationsphase ist, wenn AgentM eine vollständig gemischte Explorationsstrategie spielt, durch eine untere Schranke von $|Z^j|\tau^j + o^j_{max}$ begrenzt. Um für eine Markovstrategie mit Gedächtnistiefe O^j und Aktionsspeicher τ^j die Parameter

der Markovübergangsmatrix M^j korrekt zu schätzen, muss nach einer Initialisierungsphase von o^j_{max} Runden jeder der $|Z^j|$ Zustände τ^j mal besucht werden. Die tatsächliche Anzahl der Runden, welche dafür notwendig sind hängt von der stochastischen Aktionskette der beiden Spieler ab.

Nicht jede Strategie lässt sich durch Exploration vollständig parametrisieren. Daher werden mögliche Negativeffekte, welche aus *Pfadabhängigkeiten* der Explorationsphase entstehen können, nicht in die Optimalitätsbewertung eingeschlossen (vgl. Powers & Shoham, 2005a, S. 818-819). Der Sachverhalt wird anhand bedingter Gegenspieler im wiederholten Prisoner's Dilemma deutlich (siehe Tabelle 2.2). Sofern beide Spieler bis zur aktuellen Runde nicht abgewichen sind, kommen unter anderem sowohl *Grim Trigger*,[58] als auch *Always Cooperate* als Gegnerstrategie in Frage. Die beste Antwort gegen einen Grim Trigger Spieler ist ständiges Kooperieren mit einem Durchschnittspayoff von 3, während di e beste Antwort gegen einen Always Cooperate Spieler ständiges Abweichen mit einem Durchschnittspayoff von 5 ist. Es gibt jedoch keine mögliche Lernstrategie, welche die Leistung einer besten Antwort gegen beide Gegner erzielen kann. Grund hierfür ist, dass der Lernalgorithmus mindestens einmal Abweichung spielen müsste, um die beiden Strategien voneinander differenzieren zu können. Jedoch wäre danach die Option der vorteilhaften Kooperation mit dem Grim Trigger Spieler nicht mehr existent (vgl. Powers & Shoham, 2005a). Es handelt sich in Analogie zu Welle-Teilchen Dualismus der Physik um einen Gegner, der vor der Messung eine Superposition aus mehreren plausiblen Strategien darstellt. Durch den Akt der experimentelle Messung selbst kollabiert jedoch der Möglichkeitsraum und die Gelegenheit einer Fortgeführten Kooperation gegen bestimmte Gegner wird unmöglich, auch wenn der Test ebenso gut einen Gegner hätte offenbaren können, gegen den mit Abweichung bessere Ergebnisse erzielt worden wären.

Derartige Pfadabhängigkeiten können, wie im vorangegangenen Beispiel illustriert, nicht nur die Auszahlungsleistung eines Spielverlaufs beeinflussen, sondern auch verhindern, dass die tatsächliche gegnerische Markovstrategie überhaupt vom MAL-Spieler gelernt werden kann. Eine Markov-Gegnerstrategie s^j, deren bedingte Logik wie zum Beispiel im Falle von Grim Trigger nicht vollständig, das heißt in jedem Zustand, vom Zugverhalten der Strategie s^i des explorierenden Spielers kausal abhängt,[59] kann sich aufgrund ihrer Markovübergangsmatrix vor der vollständigen Parametrisierung durch Exploration in einer Teilmenge ihres Markovzustandsraumes festsetzen, aus welcher sie der explorierende Spieler wie im Falle von Grim

[58] Der Grim Trigger Spieler kooperiert stets, wechselt aber nach der ersten Abweichung des Agenten selbst zu permanenter Abweichung.

[59] Nach dem erstmaligen Abweichen durch den Spieler, bedingt Grim Trigger nicht länger auf dem Verhalten des Spielers. Der Sachverhalt gilt neben derartig partiell-unbedingter Zuglogiken auch für Markovstrategien, die partiell nur auf dem Zugverhalten des Gegenspielers bedingen. Dies ist immer dann gegeben, wenn der MAL-Agent nicht zumindest mit einer geringen Wahrscheinlichkeit dazu in der Lage ist, den Gegner durch sein Zugverhalten in einen bestimmten Markovzustand zu bringen.,

Trigger nicht mehr herauszulocken vermag. In Folge bleibt die tatsächliche Gegnerstrategie verdeckt.

Powers und Shoham (2005a, S. 818-819) schlagen zwei Alternativen vor, mit derartigen Irreversibilitäten umzugehen:

- **Ausschluss (partiell) unbedingter und nur partiell auf dem MAL-Agenten bedingender Gegner:** Nicht vollständig auf den explorierenden MAL-Agenten bedingende Markovstrategien werden aus der Zielgegnermenge ausgenommen.
- **Anpassung der ε-besten Auszahlung:** Die Bewertungsanforderung des Kriteriums der gezielten Optimalität ist nicht länger das Erzielen einer allgemein bestmögliche Auszahlungsleistung gegen den Gegner, die durch irgendeine MAL Lösung von Spielbeginn an möglich ist. Stattdessen soll der höchste durchschnittliche Payoff, der nach einer beliebig langen Aktionsfolge im Rahmen einer Explorationsphase zur Ermittlung der Gegnerstrategie noch möglich ist, erreicht werden. Im vorangegangenen Beispiel wäre demnach gegen einen Grim Trigger Spieler eine durchschnittliche Auszahlung von 1 trotz der existierenden Möglichkeit einer durchschnittlichen Auszahlung von 3 als gezielt optimal bewertet worden.

Die erste Alternative wurde im Zuge der Arbeit als nicht zielführend eingestuft. Hintergrund ist, dass vor allem aufgrund der empirischen Prävalenz der Grim Trigger Strategie bei menschlichen Spielern (Dal Bo & Frechette, 2018, S. 83) derartige Markovstrategien von besonderem Interesse für die Relevanz der Ergebnisse ist. Dementsprechend wird für die konzeptionelle die Herabsetzung der ε-besten Auszahlung auf das nach einer Explorationsphase noch mögliche Niveau herangezogen.

Für einen vollständig auf den Markovagenten bedingenden Markovgegner konvergiert AgentM bei einer hinreichend langen Explorationsphase gegen die korrekte Parametrisierung des Gegnermodells. Auf Basis des erzeugten Gegnermodells kann dann garantiert mit dem durchschnittlichen Erwartungsnutzen als Auswahlkriterium die exakt beste Markovantwort-Strategie gefunden werden, sofern $m_z^j \in \mathbb{Q} \forall z \in Z^j$. Wenn $\exists m^j \in \mathbb{R} \backslash \mathbb{Q}$ gilt, kann eine beste Antwort nicht notwendigerweise gefunden werden, da die exakte Parametrisierung der gegnerischen Übergangswahrscheinlichkeiten nicht länger möglich ist. Dennoch kann eine ε-beste Antwort gefunden werden. Für einen nicht vollständig auf den MAL-Agenten bedingenden Gegner wird eine im Sinne der Pfadabhängigkeit angepasste ε-beste Antwort gefunden. Die tatsächliche Gedächtnistiefe des Gegners ist dafür nicht relevant (vgl. Press & Dyson, 2012).

Zusammenfassend erfüllt AgentM anhand der bisherigen Untersuchungen bei (1) der Gewährung einer nicht gewerteten Explorationsphase und (2) der Nichtbewertung negativer Pfadabhängigkeiten gemäß der Vorgaben von Powers und Shoham (2005a) die Eigenschaft der gezielten Optimalität. Es sei hervorgehoben, dass im Rahmen der *empirischen* Untersuchung keine

Explorationsphase genutzt wurde. Diese scheint jenseits der *konzeptionellen* Bewertung nicht zielführend. Erstens könnten menschliche Gegenspieler von einem anfänglichen erratischen Explorationsverhalten irritiert werden, sodass die Qualität der nachfolgenden Interaktion leidet. Zweitens scheint die Nichtbewertung der Explorationsphase im Anwendungskontext ein asymmetrischer Vorteil des Agenten, der die eigentlichen Kosten der Explorationsphase vernachlässigt.

6.1.2 Sicherheit

Weiterhin soll ein AgentM *Sicherheit* bieten, sodass er auch von Gegnern außerhalb der Zielklasse nicht ausgenutzt werden kann, indem er gegen jene zumindest einen durchschnittlichen Payoff von maximal ε unter dem Sicherheitswert (Maximin-Payoff) garantiert (vgl. Powers & Shoham, 2005b).

Der Sachverhalt soll anhand einer Modifikation des Matching Pennies Beispiels erarbeitet werden, welches Fudenberg und Levine (1998) nutzen, um zu zeigen, dass Fictitious Play nicht sicher ist. Bei dem wiederholten Matching Pennies Spiel handelt es sich um ein Nullsummenspiel, in dem der Agent einen Payoff von 1 erzielt, wenn beide Spieler die gleiche Aktion wählen; andernfalls entspricht die Auszahlung -1 (siehe Tabelle 6.1). Der Sicherheitswert des Matching Pennies Spiels entspricht bei einer vollständigen Randomisierung durch den Spieler einer durchschnittlichen Auszahlung von 0. Wie Tabelle 6.2 exemplarisch zeigt, existieren Spielverläufe, für welche der Sicherheitswert im Durchschnitt durch AgentM nicht erreicht wird. Im dargestellten Beispiel interagiert ein Fictitious Play Spieler, welcher durch einen Markovagenten ohne Gedächtnistiefe $O = (0,0)$ werden kann mit einem deterministisch alternierenden Spieler. Im konkreten Beispiel verwendet AgentM00 einen Prior von $\hat{M}_0 = (\frac{\sqrt{2}}{1+\sqrt{2}})$ mit $\gamma_0 = 1$ als Gewicht für dessen Aktualisierung. Das Gegnermodell von AgentM00 ist dabei immer der Gestalt, dass zyklisch stets die falsche Aktion des Gegners als wahrscheinlicher antizipiert wird. Das Verletzen des Sicherheitskriteriums ist im Anhang (siehe Tabelle A.19 und A.20) anhand weiterer exemplarischer Spielverläufe, auch durch einen AgentM01 mit $O = (0,1)$ dokumentiert.

Tabelle 6.1: Normalform des Matching Pennies Spiels mit den Auszahlungen des Stufenspiels für den Zeilenspieler gefolgt von denen des Spaltenspielers. Quelle: Eigene Darstellung.

	a_1^2	a_2^2
a_1^1	$1/-1$	$-1/1$
a_2^1	$-1/1$	$1/-1$

Tabelle 6.2: Interaktionsverlauf zwischen einem exemplarischen AgentM00 mit $O = (0,0)$, Prior $\hat{M}_0 = (\frac{\sqrt{2}}{1+\sqrt{2}})$, Priorgewicht $\gamma_0 = 1$ und einem deterministisch alternierenden Gegner im wiederholten Matching Pennies Spiel.

	Aktionen		AgentM00 $i=1$			Auszahlung	
t	$\mathbf{P}[a_2^i]$	a^j	$\hat{M}^j_{(0,0)}$	z	$m^*(z_t)$	$\bar{\mathbf{E}}[r^i]$	$\bar{\mathbf{E}}[r^j]$
0			$(\frac{\sqrt{2}}{1+\sqrt{2}})$	–	1.00		
1	100%	a_1^j	(0.29)	–	0.00	-1.00	-1.00
2	0%	a_2^j	(0.53)	–	1.00	-1.00	-1.00
3	100%	a_1^j	(0.40)	–	0.00	-1.00	-1.00
4	0%	a_2^j	(0.52)	–	1.00	-1.00	-1.00
5	100%	a_1^j	(0.43)	–	0.00	-1.00	-1.00
...	...	...	...	...	...	...	...
$T-1$	0%	a_2^j	$(0.50+\varepsilon_t)$	–	1.00	-1.00	-1.00
T	100%	a_1^j	$(0.50-\varepsilon_t)$	–	0.00	-1.00	-1.00

Anhand des Beispiels konnte gezeigt werden, dass Konfigurationen des AgentM existieren, bei welche für bestimmte gegnerische Strategien keine Sicherheit gewährleisten. Gleichwohl kann die Sicherheitseigenschaft bei Bedarf ohne Weiteres ergänzt werden. Ein Lösungsansatz dafür besteht zum Beispiel darin, auf die Maximin-Strategie des Agenten zu wechseln, sobald dessen durchschnittlicher Payoff für eine bestimmte Anzahl an Runden unter den Sicherheitswert fällt (vgl. Powers & Shoham, 2005a). Steigt der durchschnittliche Payoff wieder über den Sicherheitswert, kann erneut auf die adaptive Verhaltensweise des Markov-Moduls AgentM zurückgegriffen werden.

Im Rahmen der empirischen Untersuchungen dieser Arbeit wurde auf ein derartiges *Sicherheitsmodul* verzichtet, da die Leistungsfähigkeit eines rein adaptiven Markovagenten untersucht werden soll. Ein Wechsel des Strategietypus in einzelnen Spielverläufen würde die Ergebnisdaten diesbezüglich verzerren. Gleichwohl kann eine derartige Funktionalität zur Wahrung von Formalkriterien für MAL Agenten ohne Weiteres integriert werden. Viel mehr noch, würde eine Absicherung der Payoffs nach unten hin die Leistung von AgentM erwartungsgemäß relativ zu den hier erzielten empirischen Ergebnissen tendenziell verbessern.

6.1.3 Kompatibilität

Zuletzt soll ein AgentM *Kompatibilität* mit sich selbst aufweisen. Ein Spieler ist autokompatibel, wenn der durchschnittliche Payoff im Selbstspiel höchstens ε unter der minimalen Aus-

zahlung der Nash-Gleichgewichten liegt, die nicht von anderen Nash-Gleichgewichten paretodominiert werden (vgl. Powers & Shoham, 2005b).

Für Gegner, die eine stationäre Strategie oder eine stationäre bedingte (Markov-)Strategie spielen, findet AgentM nach hinreichend langer Exploration stets die beste Antwort. Da im Selbstspiel beide Spieler entsprechend dieser Beschreibung agieren befindet sich das Spiel im Nash-Gleichgewicht. Da AgentM bei Existenz mehrere indifferente Alternativen für seine beste Antwort die erwartete Auszahlung des anderen Spielers maximiert (siehe Kapitel 3.3.2.3), handelt es sich dabei um ein paretoeffizientes Nash-Gleichgewicht.

Da jedoch im Selbstspiel von AgentM als Lernalgorithmus nicht in jedem Fall von Stationarität ausgegangen werden kann, ist eine genauere Betrachtung der Initialisierungs- und Markovlogik erforderlich. Die Autokompatibilitätseigenschaft im Selbstspiel hängt aufgrund der Eröffnungslogik von AgentM (siehe Kapitel 3.3.2.3) vom betrachteten Spiel ab. Wenn genau ein Zustand des Stufenspiels kooperativ ist, spielt AgentM diesen so lange, bis der Gegner das erste mal von der Kooperationslösung abweicht. Da im Selbstspiel beide Akteure ein AgentM sind, wird keiner initial abweichen, sodass die Kompatibilitätseigenschaft stets gegeben ist.

In allen anderen Fällen ist Kompatibilität immer dann gewährleistet, wenn die beiden AgentM Spieler gegen ein stationäres Markov-Gegnermodell konvergieren. Für dieses finden beide jeweils stets die beste Antwort, sodass sich ein Nash-Gleichgewicht einstellt. Da AgentM großzügig ist, also bei Existenz gleichwertiger bester Antworten stets die erwartete Auszahlung des Gegners maximiert, wird sichergestellt, dass das Gleichgewicht nicht paretodominiert wird und die Kompatibilitätseigenschaft ist gegeben. Da die Gegnermodelle jeweils vom Verhalten des anderen Spielers abhängen, welches jedoch wiederum vom Gegnermodell abhängt, lässt sich nicht ausschließen, dass es Parameterkombinationen des Markovagenten gibt, die für ein gegebenes Spiel zu nicht konvergierendem Selbstspiel führen. Es ist also möglich, dass es Situationen gibt, in denen sich die beiden Spieler in einem nichtkonvergenten Zyklus aus Gegnermodellen und Aktionsmustern befinden. Eine Auszahlung, die höchstens ε unter der minimalen Auszahlung der Nash-Gleichgewichte liegt, die nicht von anderen Nash-Gleichgewichten paretodominiert werden, kann dabei nicht garantiert werden. Die Kompatibilitätseigenschaft ist in diesem Spezialfall nicht gegeben.[60] Da der Markovagent primär zum Spiel gegen Menschen entwickelt wurde, handelt es sich hierbei um das relativ unwichtigste Kriterium, dessen stellenweise Nichterfüllung keine signifikante Tragweite zugeschrieben wird.

[60] In Nullsummenspielen resultiert das Sicherheitsmodul in einer Erfüllung der Kompatibilitätseigenschaft, da die Minimax-Strategie stets in einem Nash-Gleichgewicht resultiert.

6.2 Diskussion der experimentellen Validität: Wie gut schneidet AgentM in empirischen Untersuchungen ab?

Die Kriterien der konzeptionellen Validität gehen bisher von einer bekannten Zielgegnermenge aus. In der Realität ist jedoch bisweilen nicht nur der einzelne Gegner, sondern auch die Grundgesamtheit der potentiellen Gegner unbekannt. Darin besteht gewissermaßen die zentrale Herausforderung. Daher soll die konzeptionelle Diskussion um eine experimentelle Überprüfung der Leistungsfähigkeit ergänzt werden. Ein erfolgreicher präskriptiver MAL Algorithmus für nichtkooperative Anwendungen erzielt in einer gegebenen Landschaft eine hohe Auszahlung, wobei eine Haupteigenschaft der Landschaft die Klasse möglicher Gegner ist (vgl Shoham & Powers, 2014a). Um dieser Maxime entsprechend ein umfassendes Bild zu liefern, wird eine Zweischrittige experimentelle Betrachtung vorgenommen:

- **Algorithmische Gegner:** Im ersten Schritt sollen hierfür Gegner mit einer unbekannten aber explizit vordefinierten Strategie, also reguläre unbekannte algorithmische Gegner diskutiert werden. Sie erlauben eine Aussage darüber wie gut AgentM im Vergleich mit Gegnern abschneidet, welche nicht notwendigerweise in seiner Zielgegnerklasse sind.
- **Menschliche Gegner:** Im zweiten Schritt sollen Gegner mit einer unbekannten und lernenden Strategie betrachtet werden. Hierfür werden menschliche Gegner gegenüber MAL-Algorithmen aus zwei Gründen bevorzugt. Erstens stellen sie als *natürlicher* MAL-Agent einen intuitiven Benchmark dar. Zweitens verhindert die implizite und individuelle Natur der menschlichen Strategien, beziehungsweise Lernprozesse, dass im Sinne des Reverse Engineerings ein maßgeschneiderter MAL-Agent entworfen werden kann, was die Validität der Ergebnisse weiter steigert.

6.2.1 Experimente gegen algorithmische Spieler

Als naheliegende Testumgebung für experimentelle Untersuchungen von Strategien bietet sich das wegweisende Turnier von Axelrod (1980) an. Dabei spielte eine Population von 15 eingesandten algorithmischen Strategien in einem Round Robin Turnier jeweils 200 Runden des wiederholten Prisoner's Dilemmas. Dieses wird hier nachgestellt, wobei jede Strategiepaarung jeweils 10 Mal miteinander über die 200 Runden spielt, um stochastische Effekte zu reduzieren. Für die Algorithmen ist die Rundenzahl unbekannt. Die ursprüngliche Strategiepopulation [61] (siehe Tabelle A.22) wurde um den Markovagenten ergänzt. In Summe spielt jede Strategie somit 160 Spiele. Die Parametrisierung des Markovagenten erfolgt analog zu den vorangegan-

[61] Die Axelrodstrategien wurde anhand von Vince Knight et al. (2020) umgesetzt.

Tabelle 6.3: Deskriptive Auswertung der durchschnittlichen normierten Auszahlungen $\bar{\rho}^i$ von AgentM01 im nachgestellten Round Robin Turnier von Axelrod (1980) zum wiederholten Prisoner's Dilemma. Quelle: Eigene Darstellung.

			Perzentile				
Algorithmus	n	$\bar{x}$	25%	$\tilde{x}$	75%	s	$\frac{s}{\bar{x}}$
AgentM01	160	53.93	54.53	60.00	60.00	12.42	0.23
Stein & Rapoport	160	52.66	50.50	60.00	60.00	12.88	0.24
Grofman	160	52.09	45.08	60.00	60.00	13.52	0.26
Tit-for-Tat	160	51.49	45.10	60.00	60.00	13.36	0.26
Shubik	160	51.35	44.63	60.00	60.00	15.03	0.29
Tideman & Chieruzzi	160	50.80	47.60	60.00	60.00	15.52	0.31
Nydegger	160	50.61	51.00	60.00	60.00	16.60	0.33
Davis	160	49.70	38.80	60.00	60.00	16.01	0.32
Grim Trigger	160	49.64	38.35	60.00	60.00	16.36	0.33
Graaskamp	160	46.60	30.00	52.50	56.85	14.13	0.30
Downing	160	41.67	21.58	28.00	58.68	22.75	0.55
Feld	160	39.96	26.60	29.55	55.30	17.21	0.43
Joss	160	36.78	23.30	27.15	53.35	16.61	0.45
Tullock	160	35.63	25.08	29.70	47.30	13.84	0.39
Random	160	32.48	13.98	35.05	44.43	19.70	0.61
Anonymous	160	31.89	13.30	34.20	44.63	19.59	0.61

genen Experimenten mit Menschen. Einziger Unterschied ist die Verwendung eines Priors von $\hat{m}_0 = 0.5$ entsprechend des Indifferenzprinzips.[62]

Tabelle 6.3 zeigt die Ergebnisse für AgentM01. AgentM01 zeichnet sich mit 53.93 durch den höchsten Mittelwert der durchschnittlichen normierten Auszahlung über alle Spiele aus. Auch die ausgewiesenen Quartile zeichnen sich durch eine dominante Leistung aus. Die im Mittel überlegene Auszahlungsleistung von AgentM kann dieser mit einer geringen Varianz erzielen, sodass er mit 0.23 den geringsten Variationskoeffizienten aufweist. Die Ergebnisse zu AgentM11 und AgentMx1 (siehe Tabellen A.20 und A.21 im Anhang) bestätigen dies durch ihre Konsistenz; auch sie belegen jeweils den ersten Platz des Axelrod-Turniers.[63] Der Tit-for-

[62] Eine Verwendung der empirischen Priors in Tabelle 3.4 aus menschlichen Spielverläufen bietet sich hier per Definition nicht an.

[63] AgentM01 wird hier im Spiel gegen Algorithmen prominent behandelt, da dieser im Rahmen der Regressionsanalysen zu den Experimenten im Spiel gegen Menschen am besten abschneidet (siehe Tabellen 5.17 und 5.18).

Tat-Spieler, welcher im Experiment mit Menschen besonders gut abschneidet (siehe Kapitel 5.2.4) konnte in dieser Konstellation lediglich den vierten Platz belegen, wobei die Abstände im oberen Feld relativ gering auszufallen scheinen. Auffällig ist, dass alle in der Population performanten Strategien sich durch einen Median von 60.00 auszeichnen, was der normierten Auszahlung der Kooperationslösung entspricht und die Notwendigkeit von Kooperationsfähigkeit für erfolgreiche Strategien unterstreicht (vgl. Axelrod, 1984).

6.2.2 Experimente gegen menschliche Spieler

Als nächst schwieriger Benchmark bieten sich menschliche Gegner an. Aufgrund vielschichtiger und teilweise impliziter Entscheidungsprozesse gestaltet sich menschliches Handeln komplex. Insbesondere sind menschliche Strategien durch ein gewisses Maß an Rauschen charakterisiert (vgl. Müller, 2018). Infolgedessen gestaltet sich die effektive und effiziente Gegnermodellierung herausfordernd. Ein erfolgreicher MAL Algorithmus muss in der Lage sein im Kontext einer derartigen Dynamik agieren zu können. Die Aussagekraft der Ergebnisse wird durch die schwammige Struktur menschlichen Aktionsverhaltens jedoch bestärkt. Besonders hervorzuheben ist dabei, dass durch die implizite Natur der menschlichen Strategien die Entwicklung einer durch Reverse Engineering herbeigeführten Speziallösung kaum möglich ist.

Die Effizienz des Lernprozesses von AgentM wird sichergestellt, indem die Interaktion auf wiederholte Spiele mit maximal 26 Runden beschränkt ist und auf eine Explorationsphase verzichtet wird. Im Gegensatz dazu verwenden Carmel und Markovitch (1996) aber auch Powers und Shoham (2005a, 2005b) nicht gewertete Explorationsphasen von mehreren hundert Runden. Gründe hierfür ist neben der Effizienz der Lernmethode die Tatsache, dass nicht nur die Parametrisierung, sondern auch die Struktur (vgl. Carmel & Markovitch, 1996) oder der Typus des Gegners (vgl. Powers & Shoham, 2005a, 2005b) bestimmt werden muss. Dieser erhöhte Informationsbedarf macht die Gesamtlösung vergleichsweise schwerfällig. Trotz der vergleichsweise kurzen (vgl. z.B. Axelrod, 1980; Carmel & Markovitch, 1996, 1998; Powers & Shoham, 2005a, 2005b) Interaktion konnte AgentM entsprechend der Ergebnisse in Kapitel 5.2 durchweg den menschlichen Referenzspielern signifikant überlegene Auszahlungswerte erzielen (siehe Tabelle 5.17).

Viel mehr noch scheint die Verwendung einer randomisierenden Explorationsphase im Spiel gegen Menschen kontraproduktiv. Ein solcher Spieler kann schnell als beliebig eingestuft werden, sodass Menschen zum Beispiel im Prisoner's Dilemma schnell mit *Abweichung* reagieren können. Von einem derartigen Zustand wieder in eine Kooperationslösung zu koordinieren oder den Gegner gar ausbeuten zu können gestaltet sich voraussichtlich als nichttrivial.

6.3 Integrierte Diskussion der Gesamtergebnisse

Die bisherigen Ausführungen zu konzeptioneller und experimenteller Validierung zusammenführend, beschäftigt sich dieses Kapitel mit der Ableitung übergreifender Einblicke. Im Zuge dessen wird erstens die Relevanz von Lernkosten für MAL Algorithmen herausgearbeitet, zweitens die Kontextabhängigkeit des Leistungsbegriffes unterstrichen, drittens AgentM zu anderen MAL Lösungen abgegrenzt und viertens mit gestalterischen Abwägungen für MAL Algorithmen geschlossen.

6.3.1 Lernkosten und Exploration

Die Verwendung einer *Explorationsphase* im Sinne einer gegenüber der Maximierung der eigenen Auszahlung primär auf das Lernen des gegnerischen Verhaltens wurde bereits partiell aufgegriffen. Insgesamt lässt sich der Sachverhalt als eine mehrstufige Abwägung analog zu Abbildung 6.1 gliedern, deren Vor- und Nachteile kurz aufgeführt werden. Der Verzicht auf eine Explorationsphase ermöglicht, bereits ab der ersten Spielrunde der Zielfunktion der Auszahlungsmaximierung zu folgen. Anders gesagt, benötigt ein MAL Agent, der ohne eine Explorationsphase auskommt weniger Informationen, um performant zu agieren. Dennoch können nicht-explorative Agenten in einem lokalen Minimum gefangen sein (vgl. Carmel & Markovitch, 1998). Bei Verwendung einer Explorationsphase wird das Risiko, ein suboptimales, aber mit dem Verhalten des Gegners konsistentes Gegnermodell zu verwenden, reduziert. Jedoch

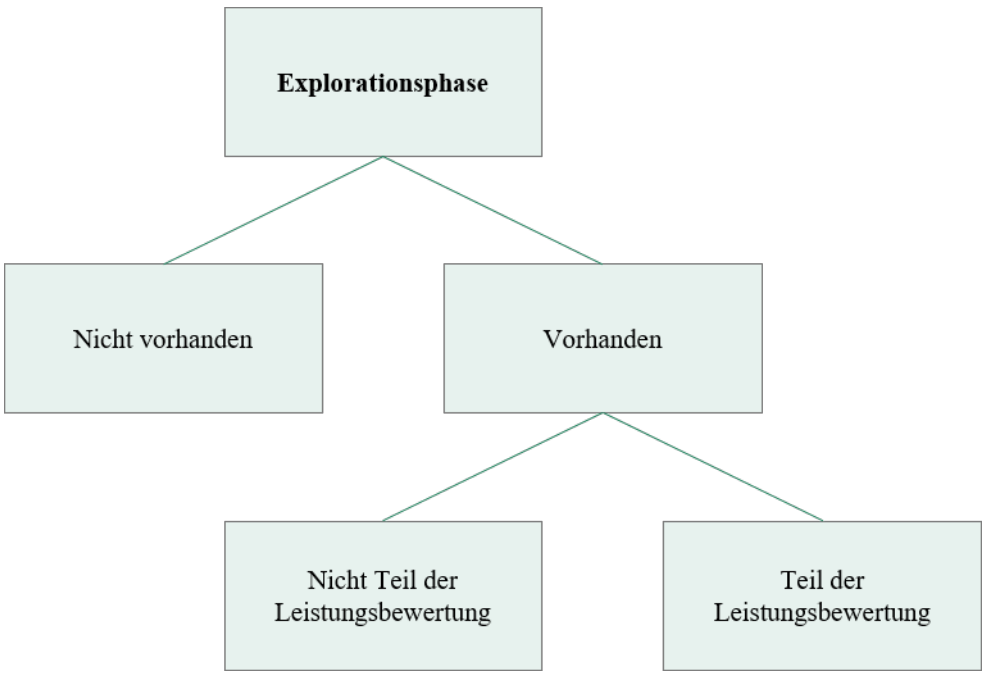


Abbildung 6.1: Gestaltungsmöglichkeiten der Explorationsphase. Quelle: Eigene Dastellung.

besteht eine Schwäche darin, dass willkürlich-exploratives Verhalten bei lernenden Gegnern zu irreversiblen und die noch erzielbare Auszahlung reduzierendem Antwortverhalten führt. Beispielsweise kann sich der ursprünglich kooperationsbereite Gegner als Reaktion auf augenscheinlich kausalitätsfreies Agieren des Gegners auf das suboptimale Nash-Gleichgewicht des Stufenspiels zurückziehen können. Das Explorieren verändert bei lernenden Gegnern das zu Lernende. Dies ist insbesondere bei menschlichen Gegnern zu beachten, wird aber auch durch das Spiel gegen einfache bedingende Strategien wie Gegner Grim Trigger illustriert. Wird eine Explorationsphase verwendet, ist außerdem zu entscheiden, ob diese in die Leistungsbewertung des Algorithmus einfließen soll. Geläufige MAL Agenten verzichten auf die Mitbewertung der Explorationsphase (vgl. z.B. Carmel & Markovitch, 1996; Powers & Shoham, 2005a, 2005b), beziehungsweise nehmen eine nur sehr reduzierte Validierung vor (vgl. z.B. Carmel & Markovitch, 1998). Diese Arbeit kommt dagegen zu dem Schluss, dass die Explorationsphase maßgeblich für die Bewertung der *Informationseffizienz* eines MAL Agenten ist.

Auch bei Mitbewertung der Exploration sei angemerkt, dass die Untersuchung von Spielverläufen von mehreren hundert bis zu $200,000$ Runden fragwürdig ist (vgl. z.B. Carmel & Markovitch, 1998; Powers & Shoham, 2005a). Denn klar ist, dass bei zunehmender Rundenzahl der relative Einfluss einer anfänglichen Lernphase abnimmt, sofern diese überhaupt berücksichtigt wird. Der Effizienzaspekt als Anforderung wurde von der MAL Literatur noch nicht aufgegriffen. Vor diesem Hintergrund kann AgentM als Ausgangspunkt für einen schlanken und effizienten MAL Agenten wert stiften, welcher selbst im Spiel gegen Menschen attraktive Leistungswerte erzielt. Aufgrund der Fähigkeit, Gegnerverhalten effizient modellieren zu können, kann AgentM unter Rückgriff seines beschränkten Aktions- und Fehlergedächtnisses selbst sich veränderndes Gegnerverhalten erfassen, was insbesondere bei Interaktionen mit Menschen von Bedeutung ist (vgl. Albrecht & Stone, 2018).

6.3.2 Kontextabhängigkeit des Leistungsbegriffes

Eine zentrale Herausforderung bei der Leistungsbewertung besteht darin, dass die mögliche erzielbare Leistung in *Abhängigkeit zur Gegnerpopulation* steht. Aufgrund dieser Kontextabhängigkeit existiert eine universell payoffmaximierende Strategie oder Lernalgorithmus je nach Informationslage nicht (vgl. Young, 2004). Formal tragen Powers und Shoham (2005b) dem Sachverhalt Rechnung, indem sie Leistungskriterien in Abhängigkeit einer Zielgegnerklasse formulieren, sodass Informationen zu übergreifenden Eigenschaften der Gegner vorliegen. Problematisch ist hierbei, dass durch die Eingrenzung der Gegnermenge bereits eine Informationsasymmetrie hinsichtlich des Verhaltens der Spieler geschaffen wurde, sodass die Gegenspieler von Beginn an mit einem Wettbewerbsnachteil antreten müssen. Die Gegnerspieler beziehungsweise der Entwickler der gegnerischen Algorithmen hatte keine Kenntnis von dem MAL Agenten, während bei der Entwicklung speziell auf das Spiel gegen bestimmte Spieler hin optimiert

wird (vgl. z.B. Carmel & Markovitch, 1996, 1998; Powers & Shoham, 2005a, 2005b). In diesem Sinne kommt durch die Eingrenzung der möglichen Gegnerspieler nicht nur eine Informationsasymmetrie zum Tragen. Viel mehr noch bestärkt die Eingrenzung der Gegnermenge die Entwicklung von Speziallösungen, die zwar für die eingegrenzte Spielermenge gut abschneiden können, aber andernfalls nicht zufriedenstellend performen.[64] Beim Transfer auf realweltliche Anwendungen besteht eine Herausforderung darin, dass Gegnerspieler nicht notwendigerweise nur aus der Menge von Zielgegnern stammen, dass die Klassen der Gegnern nicht bekannt sind oder gar, dass sich Spieler nicht klar einer Klasse von Gegnern zuordnen lassen. Vor diesem Hintergrund findet die experimentelle Validierung des AgentM primär gegen menschliche Spieler statt, die prinzipiell jede Strategie verfolgen können. Hierdurch wird angestrebt, dass die Informationsvorteile auf Seiten der Experimentleiter möglichst gering ausfallen.[65]
Hervorgehoben sei die Tatsache, dass AgentM in jeder der drei untersuchten Konfigurationen Mx1, M01 und M11 das Axelrod-Turnier mit dem ersten Platz abschließt, obwohl der Markovagent originär für das Spiel gegen Menschen entworfen wurde.[66]

6.3.3 Abgrenzung zu anderen MAL Lösungen

Die grundsätzliche vorhandenen Interdependenzen innerhalb einer Spielerpopulation wurden im Falle von MAL Spielern verstärkt, da auch die Gegner des Agenten im Begriff eines Lernprozesses sind und die Aktionen der anderen Spieler unmittelbar von den eigenen abhängen können (vgl. Shoham & Powers, 2014a).

Um vor diesem Hintergrund eine mit Menschen zielführende Interaktion abbilden zu können ist es demnach erforderlich, deren zugrundeliegendes bedingtes Strategieverständnis (vgl. Dal Bo & Frechette, 2018; Müller, 2018) mit einzubeziehen, ohne jedoch valide konzeptionelle Leistungskriterien zu vernachlässigen. Die meisten MAL Algorithmen, wie exemplarisch durch Fictitious Play (Brown, 1951), AWESOME (Conitzer & Sandholm, 2007) und GIGA-WoLF (Bowling & Veloso, 2002) repräsentiert, sind zwar adaptiv lernend, vernachlässigen jedoch durch einen stationären Fokus potentiell bedingende Spieldynamiken (Powers & Shoham, 2005a) und genügen nicht den ganzheitlichen Leistungskriterien von Powers und Shoham (2005b). MetaStrategy erfüllt diese Kriterien zwar, ist jedoch ebenfalls auf stationäre Interaktionen ausgelegt

[64] Bei Erfüllung des Sicherheitskriteriums von Powers und Shoham (2005a) wird über Zeit zumindest der Maximin-Payoff gegen jeden Spieler garantiert.

[65] Dies wird im Vergleich zwischen dem Tit-for-Tat Algorithmus deutlich. Obwohl Tit-for-Tat im Rahmen der empirischen Untersuchung gegen menschliche Spieler besser als AgentM abschneidet, tut dies der Leistungsfähigkeit des Markovagenten keinen Abtrag. Grund ist, dass Tit-for-Tat eine für das Prisoner's Dilemma entwickelte Speziallösing ist, während AgentM, wie diese Arbeit zeigt, generalisierbare Leistungseigenschaften über eine breite Palette von 2x2 aufweist. Interessanterweise kann AgentM Tit-for-Tat im Axelrodturnier sogar schlagen.

[66] Die Tatsache, dass Tit-for-Tat nicht länger der beste der ursprünglichen Algorithmen des Axelrod-Turniers ist, mag an der veränderten Grundgesamtheit der teilnehmenden Spieler liegen, welche sich durch die Hinzunahme des Markovagenten ergibt.

(Powers & Shoham, 2005b). IT-US-L* (Carmel & Markovitch, 1998) wiederum kann bedingte Interaktionslogiken erlernen und umsetzen, die Erfüllung der Kriterien Sicherheit und Kompatibilität ist jedoch nicht gewährleistet. Lediglich Manipulator erfüllt die Leistungskriterien und kann bedingte Interaktionsmuster erkennen (Powers & Shoham, 2005a). Es existieren jedoch diverse Kritikpunkte an der Methode, beziehungsweise deren experimenteller Untersuchung:

1. Verwendung der beliebigen Annahme, dass das Aktionsverhalten der Gegner höchstens auf den Aktionen des Agenten bedingt, nie jedoch auch von den historischen Aktionen des Gegners selbst abhängt. Dies steht insbesondere im Kontrast zu den empirischen Erkenntnissen von Müller (2018).

2. Vernachlässigung der Lernkosten durch die Betrachtung von langen Interaktionen mit $200,000$ Runden, wovon die ersten $180,000$ Runden im Sinne einer kostenlosen Explorationsphase zum Erforschen der Gegnerstrategie nicht gewertet werden.

3. Beschränkung der experimentellen Validierung auf Interaktionen nur gegen andere bekannte Algorithmen, welche die Manipulator Methode jedoch durch ihr zielgerichtetes reverse engineertes Design gezielt übertreffen kann.

Insbesondere die Validierung anhand von menschlichen Gegnern als Proxy für unbekannte Gegnerstrategien, die sich erschwerend dazu über den Spielverlauf ändern können, wurde auch für andere MAL Agenten aus der Literatur nicht festgestellt. Dies ist neben der zuvor beschriebenen erhöhten Problemkomplexität womöglich auch durch einen wesentlich größeren administrativen Aufwand in der Experimentabwicklung bedingt. Auch Erschwernisse in der Ableitung von generalisierbaren Erkenntnissen auf Basis einer spezifischen Spielerpopulation können eine Rolle spielen.

AgentM hingegen wurde gegen menschliche Spieler empirisch in einer Bandbreite von Spielen validiert. Der Markovagent kann dabei Interaktionen abbilden, welche der bisweilen bedingten Aktionslogik von Menschen gerecht wird (vgl. Dal Bo & Frechette, 2013; Müller, 2018). Eine experimentelle Untersuchung im Rahmen des Axelrod (1980) Turniers ergänzt diese Ergebnisse und deutet darauf hin, dass es sich nicht lediglich um eine Nischenlösung handelt. Weiterhin erfüllt AgentM das formale Leistungskriterium der gezielten Optimalität gegen Markovspieler. Auch das Kriterium der der Sicherheit kann durch eine geringfügige Ergänzung des Algorithmus gewährleistet werden.

6.3.4 Gestalterische Abwägungen

Sowohl der Modus der Explorationsphase als auch die Abhängigkeit zwischen möglicher Leistung und Gegnerpopulation deuten auf eine Abwägung zwischen *Offenheit der Lösung* und

Informationsanforderungen von MAL Agenten hin. Je mehr Eventualitäten abgefangen werden sollen oder je weniger die Lösung bereits vorab eingegrenzt werden soll, desto mehr Informationen müssen im Spielverlauf erhoben werden. IT-US-L* beispielsweise zieht DFA-Modelle mit bis zu 70 Knoten mit ein (vgl. Carmel & Markovitch, 1998), wodurch potentiell viele Informationen über den Gegner gesammelt werden müssen. Ob es sich hierbei noch um funktional valide Gegnermodelle handelt ist fraglich. Eine Einschränkung der Modellgröße könnte dem Sachverhalt entgegenwirken, gleicht aber auch einer vorweggenommenen Annahme über mögliches Gegnerverhalten. Der Manipulator Algorithmus versucht daneben, nicht nur herauszufinden, wie der Gegner für eine gegebene Modellstruktur agiert, sondern auch, welche Modellstruktur gegen den Gegner am besten geeignet ist (vgl. Powers & Shoham, 2005a). Insofern ergibt sich ein geschachteltes Lernproblem, welches ein offeneres Lösungsfeld zu Kosten eines höheren Informationsbedarfs bietet. AgentM ist hierbei mit der Beschränkung auf Markovgegner schlanker aufgestellt, sodass er zwar schneller lernen kann, dafür aber den Lösungsraum und somit die potentiell erzielbare Auszahlung bereits vorab weiter einschränkt.

Abschließend gestaltet sich somit die Problemstellung performanter MAL Algorithmen weiterhin als komplex, insbesondere, da für einen gegebenen Spieltyp und eine gegebene Informationsmenge Lernverfahren, das bestimmte Leistungs- und Konvergenzeigenschaften besitzt, möglicherweise überhaupt nicht existiert (vgl. Young, 2004). Gleichwohl stellt die Untersuchung vielversprechender Ansätze wie Manipulator (vgl. Powers & Shoham, 2005a), IT-US-L* (vgl. Carmel & Markovitch, 1998) und dem AgentM anhand einer menschlichen Gegnerpopulation und unter Berücksichtigung der Lernkosten einer Explorationsphase ein vielversprechendes Forschungsfeld dar.

7 Gesamtbetrachtung und Fazit

Das letzte Kapitel der Arbeit betrachtet im ersten Abschnitt zusammenfassend die erzielten Ergebnisse und diskutiert dessen Wertbeitrag. Anschließend geht der zweite Abschnitt auf Limitationen der Vorgehensweise ein, während Abschnitt drei einen Ausblick auf vielversprechende zukünftige Forschungsvorhaben bietet.

7.1 Zusammenfassung von Kernergebnissen und Wertbeitrag

Welche Leistung kann ein auf Markovketten basierender lernender Agent in wiederholten Spielen gegen Menschen erzielen? Diese präskriptive Frage wirft Müller (2018, S. 148) nach der erfolgreichen Entwicklung und Validierung einer deskriptiven Methode überlegener Vorhersagekraft zur Antizipation menschlichen Spielverhaltens anhand auf der Spielhistorie bedingender Übergangswahrscheinlichkeiten auf. Die Frage nach einem performanten lernenden Agenten schmiegt sich in die Strömung nichtkooperativ präskriptiver MAL Forschung. Zentrale Herausforderung dabei ist die nicht nur technisch sondern auch konzeptionell herausfordernde Natur des sich durch die Präsenz mehrerer lernender Agenten auszeichnenden Anwendungskontextes. Die interdependent verwobenen Lernprozesse der interagierenden Agenten beeinflussen sich gegenseitig, sodass der Lernprozess selbst das zu Lernende verändert (vgl. T. Sandholm, 2007; Shoham et al., 2007). Insbesondere stellen sich das unbegrenzte Möglichkeitenfeld zukünftiger gegnerischer und eigener Verhaltensweisen (vgl. Albrecht & Stone, 2018) in Kombination mit Lernkosten verursachender Exploration sowie der pfadabhängigen Irreversibilitätseigenschaft individueller Spielverläufe als zentrale Komplexitätstreiber dar.

Die entwickelte Lösung, *AgentM*, stellt einen lernenden bedingenden Markovagenten dar welcher sich in ein *Vorhersagemodul* und ein *Interaktionsmodul* gliedern lässt. Das Vorhersagemodul interpretiert Gegnerverhalten als Markovstrategien, indem er anhand der beobachtbaren Spielhistorie in Grundzügen basierend auf der Logik von Müller (2018) ein rundenweise aktualisiertes deskriptives Gegnermodell berechnet, welches eine häufigkeitsbasierte Schätzung der Strategie des Gegners im Sinne von Übergangswahrscheinlichkeiten von Aktionen auf Basis einer Partition der Spielhistorie zulässt. Das Interaktionsmodul kennt zwei Modi Operandi, welche durch die Erkenntnisse von Axelrod (1984) bezüglich dem entstehen von überlegenen Kooperationslösungen inspiriert sind:

1. Sofern möglich, strebt AgentM zu Spielbeginn eine Kooperationslösung an.

2. Lässt sich diese mit dem gegebenen Gegner nicht realisieren, wechselt der Algorithmus auf eine selbstoptimierende Logik, die auf Basis des Markov-Gegnermodells des Vorhersagemoduls in jeder Runde eine beste Markovantwort-Strategie wählt, welche den höchsten erwarteten durchschnittlichen Payoff für die als fix angenommene geschätzte Markov-Gegnerstrategie bietet. Im Falle eigener Indifferenz findet eine Maximierung der gegnerischen Auszahlung statt.

Die Vorhersagemethodik von Müller (2018) wurde an eine interaktive Verwendung angepasst. Erstens kann AgentM im Vorhersagemodul mehr als ein mögliches Markovmodell simultan führen. Aus der Menge der möglichen Gegnermodelle wählt eine Selektionsfunktion in jeder Runde das Modell mit der aktuell höchsten Glaubwürdigkeit. So lässt sich die Entscheidung zwischen verschiedenen Gedächtnistiefen anhand des individuellen Spielverlaufs bestimmen und muss nicht exogen vordefiniert werden. Zweitens wurde die Berechnung der gegnerischen Übergangswahrscheinlichkeiten um die optionale Verwendung eines Priors erweitert. Dieser kann sich aus empirischen Spielverläufen zwischen Menschen für repräsentative Spiele rekrutieren, aber auch im Sinne des Indifferenzprinzips als gleichverteilt angenommen werden. Bei der Berechnung des Posteriors kann der Einfluss des Priors nach Bedarf über einen Gewichtungsparameter eingestellt werden. Drittens kann AgentM sich auf ändernde Strategien einstellen und ist im Gegensatz zu anderen MAL Lösungen nicht durch die Annahme einer fixen Gegnerstrategie limitiert (S.34 Albrecht & Stone, 2018, vgl.). Während Müller (2018, S.146) eine stationäre Markov-Gegnerstrategie annimmt, trägt die entwickelte Lösung der Tatsache Rechnung, dass Menschen womöglich mehr als eine Strategie für verschiedene Phasen eines Spiels einsetzen (vgl. Ioannou & Romero, 2014). Die Umsetzung findet anhand eines das Gedächtnis des Agenten beschränkenden Verfallsparameters statt.

Der entwickelte MAL Algorithmus entwickelt somit die vielversprechenden deskriptiven Methoden von Müller (2018) stringent zu einer interaktiven Lösung weiter. Diese zeichnet sich insbesondere durch die Fähigkeit aus, reichhaltige Interaktionen im Kontext gemischter auf der Aktionshistorie bedingender Strategien zu erkennen und auf diese zu reagieren. Die Ergebnisse der vorliegenden Arbeit tragen somit in Summe durch vier Hauptaspekte zum aktuellen Forschungsstand bei. Erstens wurde die Entwicklung des interaktiven Agenten ausführlich dokumentiert und formalisiert, sodass eine weiterführende Forschung möglichst gut auf die gewonnenen Erkenntnisse zurückgreifen kann. Zweitens und im Gegensatz zu existierenden bedingenden MAL Algorithmen (vgl. Carmel & Markovitch, 1998; Powers & Shoham, 2005a) liegt dabei ein Hauptaugenmerk auf Berücksichtigung von *Lernkosten*, insbesondere im Rahmen früher Interaktionsphasen, in denen das Gegnermodell noch mit unzureichend vielen Daten bestückt ist. Folgerichtig zeichnet sich der entwickelte AgentM durch seine *Informationseffizienz* aus,

sodass eine performante Leistung bereits nach wenigen Spielrunden möglich ist. Drittens wurde eine umfassende Validierung vorgenommen, die bisherige Untersuchungen zu bedingenden adaptiven MAL Spielern übersteigt. Letztere werden beispielsweise ausschließlich im Turnier gegen handverlesene algorithmische Spieler (vgl. Carmel & Markovitch, 1996, 1998), bisweilen unter Berücksichtigung formaler Kriterien (vgl. Powers & Shoham, 2005a, 2005b) untersucht. Diese Arbeit untersucht den Lösungsalgorithmus sowohl im Spiel gegen algorithmische Spieler unter Einbezug formaler Kriterien, aber auch im Spiel gegen menschliche Gegner. Letzteres findet in der MAL Literatur kaum Anwendung (S.34 Albrecht & Stone, 2018, vgl.), ist jedoch aus Sicht dieser Arbeit ein maßgeblicher Indikator für die Leistungsfähigkeit eines Lernalgorithmus, da Menschen den Archetypen eines natürlichen Lerners im MAL Kontext darstellen. Hervorzuheben ist dabei das sich potentiell über den Spielverlauf ändernde Verhalten auf Basis von Lernprozessen sowie die unbekannte Natur der individuellen menschlichen Strategie. Die vorliegende Arbeit untersucht die entwickelte Interaktionslogik anhand einer strukturell repräsentativen Teilmenge nichttrivialer symmetrischer Spiele der Topologie von D. Robinson und Goforth (2006) und liefert somit Anhaltspunkte für die Leistungsfähigkeit von AgentM gegen Menschen über alle Spieltypen. Dies deutet auf eine Generalisierbarkeit von AgentM auf wiederholte 2x2 Spiele jeder Auszahlungsfamilie hin. Es konnten klare empirische Indizien für eine über 2x2 Spiele und verschiedene Spielertypen generalisierbare Leistungsfähigkeit des AgentM gesammelt werden. Dies wurde erstens im Spiel gegen Menschen über Prisoner's Dilemma, Chicken Game und Hero Game, als auch im Axelrodturnier gegen algorithmische Spieler gezeigt.

7.2 Limitationen der Arbeit

Die zentralen Limitationen der Ergebnisse dieser Arbeit lassen sich auf (1) die Verwendung eines unpassenden Modellierungsansatzes, (2) die Beschränkung auf einen fachlichen Teilbereich und (3) den gewählten methodischen Zugang zurückführen.

Erstens stellen Markovstrategien nur eine *Näherung* tatsächlichen Verhaltens dar, welche inhärent unpräzise geartet sein kann. Markovmodelle zur Vorhersage menschlichen Spielverhaltens sind anderen Methoden überlegen (vgl. Müller, 2018) und auch der interaktive AgentM dieser Arbeit konnte vielversprechende Ergebnisse erzielen. Dennoch mag es individuelle Gegner geben, deren Verhalten sich nicht zielführend durch Markov-Übergangswahrscheinlichkeiten approximieren lässt oder für welche eine andere interaktive Lösung zielführender ist. Insbesondere die konkrete Parametrisierung des AgentM im Rahmen dieser Arbeit erhebt keinen generellen Optimalitätsanspruch, auch wenn dies für spezifische Parametrisierungen gegen spezifische Gegnertypen möglich ist. Aspekte wie beispielsweise Gedächtnistiefe, Priors und deren Aktualisierungsregel, die Wahl des Aktionsgedächtnisses, die Selektionsfunktion des glaub-

würdigsten Markovmodells, Selektionsfunktion der besten Antwort-Strategie oder die gewählte Eröffnungsstrategie sind durch eine heuristische Komponente motiviert.

Zweitens ist der präsentierte AgentM bisher nur für einen limitierten *Teilausschnitt der Spieltheorie* konzipiert beziehungsweise validiert. Dabei wurden lediglich wiederholte Spiele mit zwei Spielern, bekannter Auszahlungsmatrix und einer vollständig beobachtbaren Spielhistorie in Betracht gezogen. Insbesondere die Anwendung auf Spiele mit mehr als zwei Akteuren zieht für Markovstrategien einen Komplexitätsgewinn in der dem Gegner unterstellten Entscheidungslogik mit sich. Folglich wird die Angemessenheit der Methode für Interaktionen gegen Menschen mit mehr als zwei Spielern in Frage gestellt, da sich die zu modellierenden Menschen aus Gründen beschränkter Informationsverarbeitungskapazität auf einfachere Entscheidungsmuster zurückziehen können. Es sei jedoch angemerkt, dass AgentM selbst unbedingte stationäre Strategien abbilden vermag. Weiterhin kann AgentM nicht ohne Anpassung in stochastischen Spielen eingesetzt werden.

Drittens ist die *Methodik der Validierung* als limitierender Faktor anzuführen. Im Rahmen der formalen Untersuchung wurde zur Erfolgsmessung die erwartete Durchschnittsauszahlung herangezogen, wobei alternativ die abgezinste Summe zukünftiger Auszahlungen denkbar wäre. Im Rahmen der experimentellen Untersuchung wurde nur auf eine zwar strukturiert motivierte aber dennoch ausschnitthafte Teilmenge möglicher 2x2 Spiele zurückgegriffen, sodass die Ergebnisse keine vollumfängliche Bewertungsaussage zulassen. Weiterhin können die experimentellen Erhebungen durch die Struktur der Gegnerspieler verzerrt worden sein. Dies ist zum einen aufgrund einer systematisch beeinflussten Grundgesamtheit der menschlichen Probanden und zum anderen durch eine nicht repräsentative Auswahl der algorithmischen Gegner möglich.

7.3 Anknüpfungspunkte zukünftiger Forschung

Die Limitationen adressierend ergeben sich auf Basis des Untersuchungsgegenstandes methodische, inhaltliche und transferierende Anknüpfungspunkte für zukünftige Forschungsvorhaben. Die *methodischen* Gelegenheiten für weitere Forschung gliedern sich wie folgt:

- **Vergleich mit anderen MAL Algorithmen:** Vielversprechende Erkenntnisse kann die bisher nicht betrachtete vergleichende Leistungsbeurteilung von AgentM mit anderen MAL Algorithmen liefern. Dafür bieten sich aufgrund der Nähe des Vorgehens insbesondere bedingende adaptive Methoden an (vgl. Carmel & Markovitch, 1998; Powers & Shoham, 2005a). Die Leistungsbeurteilung kann auf zwei Arten erfolgen. Erstens im Spiel von verschiedenen MAL Agenten gegen eine menschliche Population und zweitens im Rahmen eines Turniers unter lernenden Algorithmen.

- **Betrachtung anderer Spiele:** Eine Ausweitung der empirischen Untersuchungen auf andere Spiele ist wünschenswert. Insbesondere wurden im Rahmen der Arbeit keine Konstantsummenspiele untersucht. Auch größere Spiele wie 2x3 oder 3x3 Spiele mögen neue Erkenntnisse ermöglichen. Für eine systematische weiterführende Betrachtung sei auf den Spielgenerator von Nudelman et al. (2004) verwiesen, welcher speziell für die Testung spieltheoretischer Algorithmen entwickelt wurde.

- **Längere Interaktion:** Der Schwerpunkt dieser Arbeit in der empirischen Untersuchung mit Menschen liegt mit 21 Runden auf vergleichbar kurzen Interaktionen. Hintergrund ist der Fokus auf Effizienzaspekte des MAL Algorithmus, welcher bereits nach wenigen Runden zufriedenstellende Ergebnisse liefern soll. Die Betrachtung von AgentM in längeren Interaktionen stellt somit einen interessanten Forschungsgegenstand dar, um die Eignung der entwickelten Lösung im vorgenannten Kontext zu beurteilen.

- **Eingeschränkte Informationsausstattung:** Auch die Einschränkung der Informationsausstattung der Agenten stellt eine Gelegenheit für die Gewährleistung methodischer Robustheit dar. Dementsprechend kommen Interaktionen, bei denen die Akteure die Auszahlungsmatrix oder die Aktionshistorie nicht (vollständig) beobachten können in Frage.

Inhaltliche Opportunität für weiterführende Forschung wird durch die iterative Weiterentwicklung von AgentM gewährleistet. Mögliche Betrachtungspunkte sind etwa:

- **Formale Defizite:** In Kapitel 6.1 wurde die nicht immer gewährleistete formale Validität von AgentM dargelegt. Die dort präsentierten Lösungsansätze geben Anlass für die empirische Untersuchung eines weiter verbesserten AgentM, der unter Verwendung eines Sicherheitsmoduls auch formal das Sicherheitskriterium erfüllt. Ebenso der Einfluss des Spezialfalls von nichtkonvergenten Spielzyklen im Selbstspiel auf das Kompatibilitätskriterium kann einen zukünftigen Forschungsgegenstand darstellen.

- **Verwendung von Exploration:** Der entwickelte Lernalgorithmus kommt vollständig ohne dezidierte Explorationsphase im Sinne eines aktiven Erforschens gegnerischer Verhaltensmuster durch bewusst suboptimale Handlungen aus. Dennoch gibt es Hinweise dafür, dass explorative Aspekte die Vorhersage eines Lernalgorithmus verbessern können (Carmel & Markovitch, 1997). Aufgrund von irreversiblen Pfadabhängigkeiten und Lernkosten ist der Mehrwert für einen interaktiven Agenten jedoch nicht immer gegeben, sodass hier eine gestalterische Abwägung vorgenommen werden muss. Die Untersuchung eines explorierenden AgentM kann Aufschlüsse darüber geben.

- **Verwendung von Exploitation:** Der entwickelte Lernalgorithmus nimmt die gegnerische Strategie stets als durch das eigene Verhalten unveränderlich an. Folglich optimiert

AgentM die eigene Strategie stets nur *gegeben* der gegnerischen Strategie. Die Integration eines ausnutzenden Moduls analog zu Powers und Shoham (2005a) welches, falls möglich proaktiv versucht, den Gegner zu einem bestimmten Verhalten zu drängen bietet sich als weiterer Anhaltspunkt für eine noch leistungsfähigere Lösung an.

- **Verwendung von Normstrategien:** Während AgentM jeden Schätzer gegnerischen Verhaltens als individuell beste Näherung des Gegnerverhaltens auf Basis der verfügbaren Informationen betrachtet, legen die Ergebnisse von Ioannou und Romero (2014), Müller (2018) nahe, dass sich die Strategien menschlicher Spieler zu Normstrategien clustern lassen. Dieser Aspekt findet in der hier erarbeiteten Lösung keinen Einfluss. Mögliche Stärke eines solchen Vorgehens ist, dass lediglich Antwort-Strategien für die Menge der Normstrategien, statt gegen die Menge aller Parametrisierungen berechnet werden müssen. Weiterhin kann so, falls die Normstrategien repräsentativ sind, die Konvergenz zur tatsächlichen Strategie des Gegners beschleunigt werden. Nachteilhaft dagegen ist, dass die Antwort-Strategien weniger geeignet für den individuellen Gegner sind, der entweder von Normstrategien abweicht oder sich durch diese nicht charakterisieren lässt. Die Untersuchung des Einflusses auf die Leistung eines interaktiven Algorithmus stellt somit eine interessante Fragestellung dar.

- **Aktualisierungsregel:** Alternativ zur relativen Häufigkeit als Schätzer für die bedingte Aktionswahrscheinlichkeit in einem Markovzustand kommen auch andere Aktualisierungsregeln wie beispielsweise eine Bayessches Updateregel in Frage (vgl. Rezek et al., 2008), deren Vorhersagekraft Gegenstand weiterführender Forschung sein kann.

Auch *transferierende* Forschung im Sinne von vollständig neuen aber verwandten Themen ist auf Basis von AgentM möglich:

- **Verwendung als Trainer von Neuronalen Netzen:** Auffällig ist die Abwesenheit von Neuronalen Netzen im Kontext der Forschung zu deskriptiven oder präskriptiven MAL Algorithmen. Ein Grund mag die Notwendigkeit einer enormen Datengrundlage zum Training der Netze sein. Insbesondere für präskriptive Lösungen auf Basis Neuronaler Netze verhärtet sich die Problematik dadurch, dass historische Spielverläufe aufgrund des aktiven Eingreifens des Netzes auf die Spielhistorie nicht in Frage kommen. Der entwickelte AgentM kann hier als Sparringspartner für das Training derartiger Lösungen in Betracht kommen und das Feld in diese Hinsicht erweitern.[67]

- **Verwendung als Ratgeber:** Die Empfehlung von Müller (2018, S. 148) auf die hiesige präskriptive Methodik übertragend, bieten sich empirische Untersuchungen zur Leistung

[67] Die kooperative Eröffnungsstrategie von AgentM muss zu Trainingszwecken eventuell angepasst werden.

von menschlichen Spielern an, die durch AgentM unterstützt werden. Der Algorithmus kann den Spielverlauf beobachten und Vorhersagen über mögliches Gegnerverhalten geben sowie Empfehlungen über mögliche Reaktionen des menschlichen Partners liefern.

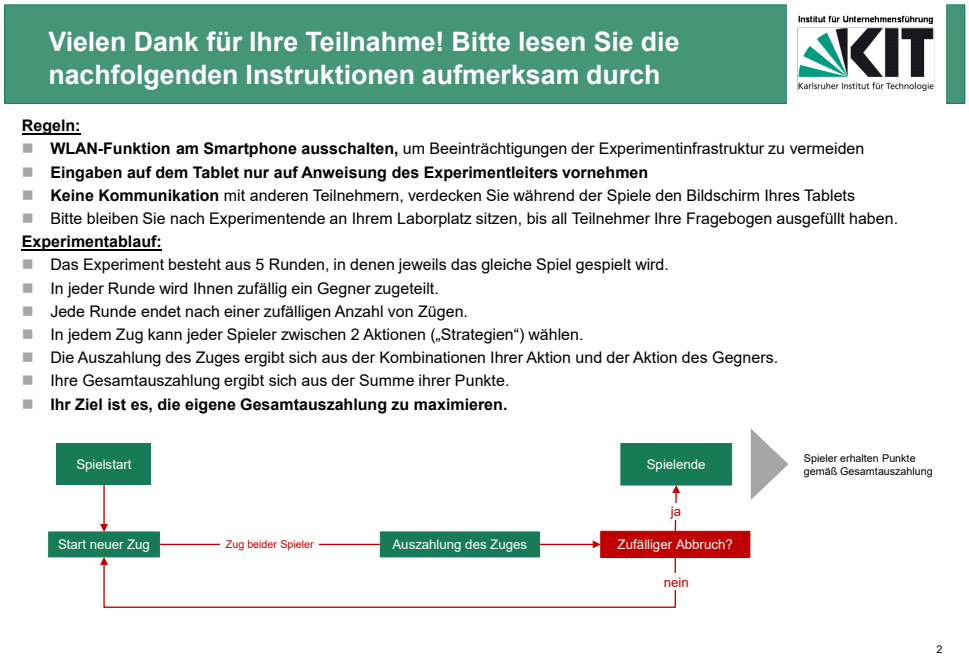
Zusammenfassend stellt diese Dissertation eine hoffentlich hilfreiche Grundlage für Forschung zu präskriptiv-interaktiven Lernalgorithmen im nichtkooperativen Kontext dar. Insbesondere durch die Fähigkeit, menschliches bedingendes adaptives Verhalten effizient abbilden zu können, ergeben sich interessante Ansätze für weiterführende empirische und formale Untersuchungen.

A Anhang

Die einfachste Darstellung der Übergangsmatrix von Tit-for-Tat ist mit $O^i = (0,1)$ möglich. Gleichwohl kann Tit-for-Tat durch jede Ordnung $O^i \geq (0,1)$ unter Verwendung von Redundanzen dargestellt werden. Beispielweise kann eine Darstellung von Tit-for-Tat mit $O^i = (1,1)$ gefunden werden, die strategisch äquivalent ist, indem sie den identischen Spielverlauf für jede gegnerische Strategie induzieren würde (siehe Gleichung A.1). Der allgemeine Fall ist analog in Gleichung A.2 dargestellt.

$$M^i_{(0,1)} = \begin{array}{c} \\ \emptyset a^j_{1,t-1} \\ \emptyset a^j_{2,t-1} \end{array} \begin{array}{c} \begin{array}{cc} a^i_{1,t} & a^i_{2,t} \end{array} \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \end{array} \Leftrightarrow \quad M^i_{(1,1)} = \begin{array}{c} \\ a^i_{1,t-1}a^j_{1,t-1} \\ a^i_{1,t-1}a^j_{2,t-1} \\ a^i_{2,t-1}a^j_{1,t-1} \\ a^i_{2,t-1}a^j_{2,t-1} \end{array} \begin{array}{c} \begin{array}{cc} a^i_{1,t} & a^i_{2,t} \end{array} \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \end{array} \tag{A.1}$$

$$\begin{aligned} M^i_{(0,1)} &= (m^i_{(\emptyset,a^j_1)}, \qquad\qquad m^i_{(\emptyset,a^j_2)} \qquad\qquad)^T \\ &\leftrightarrow \\ M^i_{(1,1)} &= (m^i_{(a^i_1,a^j_1)}, \quad m^i_{(a^i_1,a^j_2)}, m^i_{(a^i_2,a^j_1)}, \quad m^i_{(a^i_2,a^j_2)})^T \end{aligned} \tag{A.2}$$



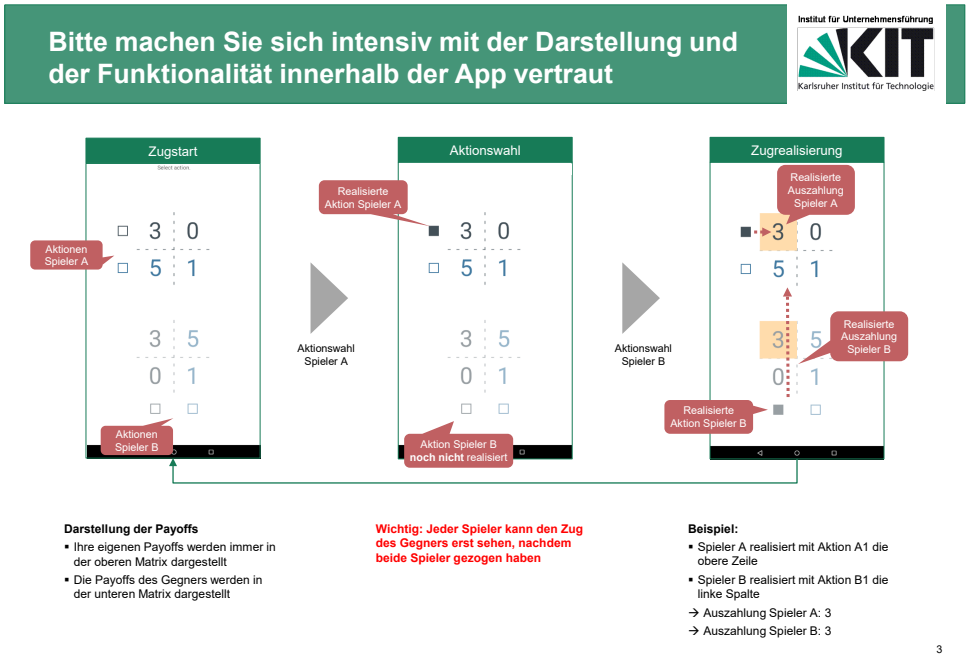


Abbildung A.1: Instruktionen der Teilnehmer. Exemplarische Version des Experiments am 24. Oktober 2019 zum wiederholten Prisoner's Dilemma. Quelle: Eigene Darstellung.

24. Oktober 2019 | Session 1 Platz:

Bitte beantworten Sie noch folgende Fragen. Dies hat keine Auswirkungen auf Ihre Bezahlung.

Was ist Ihre Sitzplatznummer?

_____

A Demographische Fragen

Wie alt sind Sie?

___ Jahre

Welches Geschlecht haben Sie?

☐ männlich ☐ weiblich

Bitte machen Sie Angaben zu Ihrem derzeitigen Bildungsstand

(höchste abgeschlossene Ausbildung)

☐ Promotion ☐ Master/Diplom ☐ Bachelor
☐ Kaufmännische Ausbildung ☐ Abitur ☐ Keine Antwort zutreffend

In welcher Fachrichtung liegt/lag Ihr Studienschwerpunkt?

☐ Wirtschaftswissenschaften ☐ Ingenieurswissenschaften ☐ Naturwissenschaften
☐ Geisteswissenschaften ☐ Andere: _____

B Fragen zu Ihren Erfahrungen mit ökonomischen Entscheidungsproblemen

Bitte beschreiben Sie, ob und auf welche Art und Weise Sie sich Kenntnisse in den Bereichen Spieltheorie und Verhandlungstheorie angeeignet haben

(Mehrfachauswahl möglich)

☐ Keine Vorkenntnisse ☐ Im Rahmen einer Ausbildung angeeignete Kenntnisse
☐ Privat angeeignete Kenntnisse ☐ Kenntnisse aus beruflichen Tätigkeiten

Wie gut schätzen Sie Ihre Kenntnisse in den Bereichen Spieltheorie und Verhandlungstheorie ein?

Kreuzen Sie die am ehesten zutreffende Aussage an

☐ Keine ☐ Gering ☐ Grundlegend ☐ Erweitert ☐ Sehr gut

→
Bitte wenden →
→

Bitte Fragen auf Vorder- und Rückseite beantworten!

Abbildung A.2: Seite 1 des Fragebogens. Exemplarische Version der ersten Session am 24. Oktober 2019 zum wiederholten Prisoner's Dilemma. Quelle: Eigene Darstellung.

24. Oktober 2019 | Session 1 Platz:

C Fragen zu Ihren Erfahrungen mit Laborexperimenten

An wie vielen ökonomischen/soziologischen/psychologischen Laborexperimenten haben Sie vor diesem Experiment teilgenommen?

☐ Keine ☐ 1 ☐ 2 ☐ 3 ☐ 3 – 9 ☐ > 10

Wie viele dieser Laborexperimente bestanden aus spieltheoretischen Entscheidungsproblemen, Verhandlungen, Auktionen oder Marktsimulationen?

☐ Keine ☐ 1 ☐ 2 ☐ 3 ☐ 3 – 9 ☐ > 10

D Fragen zum Experimentablauf

Bitte antworten Sie ehrlich, die Antworten werden nur anonymisiert ausgewertet

Was war Ihr wesentliches Ziel während des Experiments?

- ☐ Ich habe versucht, meine eigene Punktzahl zu maximieren
- ☐ Ich habe versucht, die Gesamtpunktzahl aller Spieler zu maximieren
- ☐ Ich habe versucht, die Punktzahl meiner Mitspieler zu minimieren
- ☐ Ich habe ein anderes Ziel verfolgt: _____

Haben Sie teilweise abweichend von Ihrer wesentlichen Zielsetzung agiert?

(Mehrfachauswahl möglich)

- ☐ Ich habe einmal oder öfter auf Punkte verzichtet, um eine Spielrunde schneller beenden zu können
- ☐ Ich habe einmal oder öfter auf Punkte verzichtet, um die Punkte der anderen zu minimieren
- ☐ Ich habe einmal oder öfter auf Punkte verzichtet, um _____

Bitte markieren Sie alle zutreffenden Aussagen

- ☐ Ich habe versucht, bei meiner Strategiewahl mögliche Reaktionen der Gegner zu berücksichtigen
- ☐ Ich habe in mindestens einer Runde versucht, durch häufige Strategiewechsel den Spielverlauf zu behindern bzw. meinen Gegner zu zermürben

Bitte schildern Sie Ihre Erfahrung mit diesem Experiment

(Mehrfachauswahl möglich)

- ☐ Ich habe den Ablauf des Experiments auch nach der Erklärung nicht verstanden
- ☐ Ich habe den Ablauf des Experiments über die gesamte Dauer nicht verstanden
- ☐ Ich hatte Schwierigkeiten mit der Bedienung der Experimentplattform
- ☐ Ich stand während des gesamten Experiments unter großem Zeitdruck
- ☐ Ich hatte Spaß an diesem Experiment

→
Bitte wenden →
→

Bitte Fragen auf Vorder- und Rückseite beantworten!

Abbildung A.3: Seite 2 des Fragebogens. Exemplarische Version der ersten Session am 24. Oktober 2019 zum wiederholten Prisoner's Dilemma. Quelle: Eigene Darstellung.

Abbildung A.4: Laboraufbau im Rahmen der experimentellen Erhebungen. Quelle: Eigene Darstellung.

Tabelle A.1: Ergebnisse des zwei-Stichproben t-Tests für unabhängige Daten (Student, 1908) zu Prestudy II auf H_0, dass die Mittelwerte beider Spielertypen identisch sind (grüne Kennzeichnung für positiv, rote für negativ unterschiedliche Mittelwerte; indikativ: † $p < 10\%$; signifikant: * $p < 5\%$, ** $p < 1\%$, *** $p < 0.1\%$) unter Berücksichtigung der Erfüllung der zugrundeliegenden Unabhängigkeitsannahme (UA). Quelle: Eigene Darstellung.

Vergleich								
Spielertyp 1	Spielertyp 2	$\bar{x}_1$	$\bar{x}_2$	t	p		H_0	UA
AgentMx1a	Mensch	45.52	45.79	−0.78	93.8%		✓	×
AgentMx1b	Mensch	51.53	45.79	1.72	8.9%	†	✓	×
AgentM01	Mensch	49.69	45.79	1.15	25.6%		✓	×

Tabelle A.2: Ergebnisse des gepaarten t-Tests (Student, 1908) für Daten aus Prestudy II auf H_0, dass die Mittelwerte beider Spielertypen identisch sind (grüne Kennzeichnung für positiv unterschiedliche Mittelwerte; indikativ: † $p < 10\%$; signifikant: * $p < 5\%$, ** $p < 1\%$, *** $p < 0.1\%$) unter Berücksichtigung der Erfüllung der zugrundeliegenden Normalverteilungsannahme (NvA) der paarweisen Differenzen aus Tabelle 5.5. Quelle: Eigene Darstellung.

Vergleich								
Spielertyp 1	Spielertyp 2	$\bar{x}_1$	$\bar{x}_2$	t	p		H_0	NvA
AgentMx1a	Mensch	45.52	45.79	−0.10	92.5%		✓	✓
AgentMx1b	Mensch	51.53	45.79	2.464	1.9%	*	×	×
AgentM01	Mensch	49.69	45.79	1.52	13.6%		✓	×

Tabelle A.3: Ergebnisse des Mann-Whitney-U-Tests (Mann & Whitney, 1947) für Prestudy II auf H_0, dass zwei unabhängige Stichproben aus der gleichen Verteilung gezogen wurden und demnach den gleichen Median aufweisen (grüne Kennzeichnung für positiv unterschiedliche Mediane; indikativ: † $p < 10\%$; signifikant: * $p < 5\%$, ** $p < 1\%$, *** $p < 0.1\%$) unter Berücksichtigung der Erfüllung der zugrundeliegenden Unabhängigkeitsannahme (UA). Quelle: Eigene Darstellung.

Vergleich							
Spielertyp 1	Spielertyp 2	$\tilde{x}_1$	$\tilde{x}_2$	z	p	H_0	UA
AgentMx1a	Mensch	46.67	50.48	−0.26	79.2%	✓	× ↓
AgentMx1b	Mensch	60.00	50.48	1.47	14.1%	✓	× ↓
AgentM01	Mensch	57.14	50.48	0.97	33.1%	✓	× ↓

Tabelle A.4: Ergebnisse des Random Effects Panelregressionsmodells (vgl. Das, 2019, S. 494) zu Prestudy II im wiederholten Prisoner's Dilemma (indikativ: † $p < 10\%$; signifikant: * $p < 5\%$, ** $p < 1\%$, *** $p < 0.1\%$). Quelle: Eigene Darstellung.

	Koeffizient	Standardfehler	z	p	
Spielertyp					
- Mensch (Referenz)	–				
- AgentMx1	2.57	1.86	1.38	16.7%	
- AgentM01	3.83	2.15	1.78	7.5%	†
Lerneffekte: Anzahl Spiele (Gegner)	3.15	0.55	5.71	0.0%	***
Konstante	36.41	2.79	13.06	0.0%	***
Beobachtungen	156				
Gruppen	39				
R^2_{within}	0.25				
$R^2_{between}$	0.00				
$R^2_{overall}$	0.09				
$\chi^2(3)$	36.14				
Signifikanz ($\mathbf{P} > \chi^2$)	0.0%				***

Tabelle A.5: Ergebnisse des OLS Regressionsmodells zu Prestudy II im wiederholten Prisoner's Dilemma (indikativ: † $p < 10\%$; signifikant: * $p < 5\%$, ** $p < 1\%$, *** $p < 0.1\%$). Quelle: Eigene Darstellung.

	Koeffizient	Standardfehler	t	p	
Spielertyp					
- Mensch (Referenz)	–				
- AgentMx1	2.59	2.76	0.94	34.9%	
- AgentM01	3.83	3.18	1.20	230%	
Lerneffekte: Anzahl Spiele (Gegner)	2.90	0.80	3.64	0.0%	***
Konstante	37.17	3.27	11.38	0.0%	***
Beobachtungen	156				
R^2	0.09				
$R^2_{adjusted}$	0.07				
$F(3, 152)$	4.95				
Signifikanz ($\mathbf{P} > F$)	0.3%				**

Tabelle A.6: Ergebnisse des Shapiro-Wilk-Tests (Shapiro & Wilk, 1965) auf H_0, dass bei der Differenz der Auszahlungswerte eine Normalverteilung vorliegt für Experimente I bis III zum Chicken Game (CG), Hero Game (HG) und Prisoner's Dilemma (PD) (rote Kennzeichnung für verworfene Normalverteilungsannahme (NvA); indikativ: † $p < 10\%$; signifikant: * $p < 5\%$, ** $p < 1\%$, *** $p < 0.1\%$). Quelle: Eigene Darstellung.

	Vergleich			
Spiel	Spielertyp 1	Spielertyp 2	p	H_0
CG	AgentMx1	Mensch	44.4%	✓
	AgentM01	Mensch	25.2%	✓
	AgentM11	Mensch	66.7%	✓
HG	AgentMx1	Mensch	49.0%	✓
	AgentM01	Mensch	29.8%	✓
	AgentM11	Mensch	69.1%	✓
PD	AgentMx1	Mensch	0.1% ***	×
	AgentM01	Mensch	18.3%	✓
	AgentM11	Mensch	0.3% **	×
	Tit-for-Tat	Mensch	0.0% ***	×

Tabelle A.7: Ergebnisse des Tests auf Symmetrie der Differenzen (D'Agostino et al., 1990; Royston, 1991) mit H_0, dass Auszahlungsdifferenzen $D = X_1 - X_2$ für Experimente I bis III zum Chicken Game (CG), Hero Game (HG) und Prisoner's Dilemma (PD) symmetrisch sind (rote Kennzeichnung für verworfene Symmetrieannahme (SA); indikativ: † $p < 10\%$; signifikant: * $p < 5\%$, ** $p < 1\%$, *** $p < 0.1\%$). Quelle: Eigene Darstellung.

	Vergleich			
Spiel	Spielertyp 1	Spielertyp 2	p	H_0
CG	AgentMx1	Mensch	85.0%	✓
	AgentM01	Mensch	27.9%	✓
	AgentM11	Mensch	51.0%	✓
HG	AgentMx1	Mensch	85.1%	✓
	AgentM01	Mensch	84.3%	✓
	AgentM11	Mensch	99.0%	✓
PD	AgentMx1	Mensch	26.1%	✓
	AgentM01	Mensch	79.6%	✓
	AgentM11	Mensch	87.7%	✓
	Tit-for-Tat	Mensch	9.8% †	✓

Tabelle A.8: Ergebnisse des zwei-Stichproben t-Tests für unabhängige Daten (Student, 1908) aus Experimenten I bis III zum Chicken Game (CG), Hero Game (HG) und Prisoner's Dilemma (PD) auf H_0, dass die Mittelwerte beider Spielertypen identisch sind (grüne Kennzeichnung für positiv unterschiedliche Mittelwerte; indikativ: † $p < 10\%$; signifikant: * $p < 5\%$, ** $p < 1\%$, *** $p < 0.1\%$) unter Berücksichtigung der Erfüllung der zugrundeliegenden Unabhängigkeitsannahme (UA). Quelle: Eigene Darstellung.

	Vergleich								
Spiel	Spielertyp 1	Spielertyp 2	$\bar{x}_1$	$\bar{x}_2$	t	p		H_0	UA
CG	AgentMx1	Mensch	37.98	33.99	1.00	32.1%		✓	×
	AgentM01	Mensch	41.22	33.99	1.81	7.3%	†	✓	×
	AgentM11	Mensch	42.11	33.99	2.09	4.0%	*	×	×
HG	AgentMx1	Mensch	58.05	51.81	1.71	9.1%	†	✓	×
	AgentM01	Mensch	56.02	51.81	1.17	24.5%		✓	×
	AgentM11	Mensch	56.19	51.81	1.23	22.1%		✓	×
PD	AgentMx1	Mensch	47.73	44.94	0.81	42.0%		✓	×
	AgentM01	Mensch	47.49	44.94	0.75	45.5%		✓	×
	AgentM11	Mensch	45.98	44.94	0.29	77.3%		✓	×
	Tit-for-Tat	Mensch	51.65	44.94	2.09	3.9%	*	×	×

Tabelle A.9: Ergebnisse des gepaarten t-Tests (Student, 1908) für Daten aus Experimenten I bis III zum Chicken Game (CG), Hero Game (HG) und Prisoner's Dilemma (PD) auf H_0, dass die Mittelwerte beider Spielertypen identisch sind (grüne Kennzeichnung für positiv, rote für negativ unterschiedliche Mittelwerte; indikativ: † $p < 10\%$; signifikant: * $p < 5\%$, ** $p < 1\%$, *** $p < 0.1\%$) unter Berücksichtigung der Erfüllung der zugrundeliegenden Normalverteilungsannahme (NvA) der paarweisen Differenzen aus Tabelle A.6. Quelle: Eigene Darstellung.

	Vergleich								
Spiel	Spielertyp 1	Spielertyp 2	$\bar{x}_1$	$\bar{x}_2$	t	p		H_0	NvA
CG	AgentMx1	Mensch	37.98	33.99	1.07	29.1%		✓	✓
	AgentM01	Mensch	41.22	33.99	2.33	2.5%	*	×	✓
	AgentM11	Mensch	42.11	33.99	2.82	0.7%	**	×	✓
HG	AgentMx1	Mensch	58.05	51.81	1.80	7.8%	†	✓	✓
	AgentM01	Mensch	56.02	51.81	1.22	23.1%		✓	✓
	AgentM11	Mensch	56.19	51.81	1.21	23.1%		✓	✓
PD	AgentMx1	Mensch	47.73	44.94	1.06	29.7%		✓	×
	AgentM01	Mensch	47.49	44.94	0.88	38.6%		✓	✓
	AgentM11	Mensch	45.98	44.94	0.40	69.3%		✓	×
	Tit-for-Tat	Mensch	51.65	44.94	3.09	0.4%	**	×	×

Tabelle A.10: Ergebnisse des Mann-Whitney-U-Tests (Mann & Whitney, 1947) für Experimente I bis III zum Chicken Game (CG), Hero Game (HG) und Prisoner's Dilemma (PD) auf H_0, dass zwei unabhängige Stichproben aus der gleichen Verteilung gezogen wurden und demnach den gleichen Median aufweisen (grüne Kennzeichnung für positiv unterschiedliche Mediane; indikativ: † $p < 10\%$; signifikant: * $p < 5\%$, ** $p < 1\%$, *** $p < 0.1\%$) unter Berücksichtigung der Erfüllung der zugrundeliegenden Unabhängigkeitsannahme (UA). Quelle: Eigene Darstellung.

	Vergleich								
Spiel	Spielertyp 1	Spielertyp 2	$\tilde{x}_1$	$\tilde{x}_2$	z	p		H_0	UA
CG	AgentMx1	Mensch	39.05	38.10	1.18	23.9%		✓	×
	AgentM01	Mensch	43.81	38.10	1.86	6.3%	†	✓	×
	AgentM11	Mensch	45.71	38.10	2.14	3.2%	*	×	×
HG	AgentMx1	Mensch	56.19	54.29	1.31	19.1%		✓	×
	AgentM01	Mensch	56.19	54.29	0.88	38.0%		✓	×
	AgentM11	Mensch	56.19	54.29	0.93	35.4%		✓	×
PD	AgentMx1	Mensch	60.00	52.38	1.01	31.3%		✓	×
	AgentM01	Mensch	53.33	52.38	0.76	44.6%		✓	×
	AgentM11	Mensch	57.14	52.38	0.44	66.1%		✓	×
	Tit-for-Tat	Mensch	60.00	52.38	2.04	4.1%	*	×	×

Tabelle A.11: Ergebnisse des Vorzeichentests (Arbuthnott, 1710; Snedecor & Cochran, 1991, vgl.) für Experimente I bis III zum Chicken Game (CG), Hero Game (HG) und Prisoner's Dilemma (PD) auf H_0, dass der Median der Auszahlungsdifferenzen $D = X_1 - X_2$ Null ist (grüne Kennzeichnung für positiv, rote für negativ unterschiedliche Mediane; indikativ: † $p < 10\%$; signifikant: * $p < 5\%$, ** $p < 1\%$, *** $p < 0.1\%$). Quelle: Eigene Darstellung.

	Vergleich						
Spiel	Spielertyp 1	Spielertyp 2	$\bar{D}$	$\tilde{D}$	p		H_0
CG	AgentMx1	Mensch	3.28	1.90	18.8%		✓
	AgentM01	Mensch	7.23	4.76	8.1%	†	✓
	AgentM11	Mensch	8.12	6.67	5.3%	†	✓
HG	AgentMx1	Mensch	6.24	1.90	76.6%		✓
	AgentM01	Mensch	4.21	4.76	37.1%		✓
	AgentM11	Mensch	4.38	4.76	55.2%		✓
PD	AgentMx1	Mensch	2.79	0.00	6.1%	†	✓
	AgentM01	Mensch	2.55	0.00	73.6%		✓
	AgentM11	Mensch	1.04	0.00	48.7%		✓
	Tit-for-Tat	Mensch	6.71	3.81	2.4%	*	×

Tabelle A.12: Integrierte Ergebnisse des Fixed Effects Panelregressionsmodells ohne Interaktionseffekte mit clusterrobusten Standardfehlern zu Prestudy II und Experiment I bis III im wiederholten Chicken Game, Hero Game und Prisoner's Dilemma (indikativ: † $p < 10\%$; signifikant: * $p < 5\%$, ** $p < 1\%$, *** $p < 0.1\%$). Quelle: Eigene Darstellung.

	Koeffizient	Standardfehler	t	p	
Spielertyp					
- Mensch (Referenz)	–				
- AgentMx1	3.96	1.42	2.80	0.6%	**
- AgentM01	4.63	1.48	3.12	0.2%	**
- AgentM11	4.29	1.56	2.74	0.7%	**
- Tit-for-Tat (nur PD)	8.11	1.54	5.25	0.0%	***
Spieltyp					
- Chicken Game (Referenz)	–				
- Hero Game	invariant				
- Prisoner's Dilemma	invariant				
Lerneffekte: Anzahl Spiele (Gegner)	2.04	0.34	6.04	0.0%	***
Konstante	37.75	1.41	26.83	0.0%	***
Beobachtungen	718				
Gruppen	169				
R^2_{within}	0.12				
$R^2_{between}$	0.00				
$R^2_{overall}$	0.05				
$F(5, 168)$	12.78				
Signifikanz ($\mathbf{P} > F$)	0.0%				***

Tabelle A.13: Integrierte Ergebnisse des Fixed Effects Panelregressionsmodells mit Interaktionseffekten zu Prestudy II und Experiment I bis III (indikativ: † $p < 10\%$; signifikant: * $p < 5\%$, ** $p < 1\%$, *** $p < 0.1\%$). Quelle: Eigene Darstellung.

	Koeffizient	Standardfehler	t	p	
Spielertyp im Chicken Game					
- Mensch (Referenz)	–				
- AgentMx1	3.88	3.11	1.25	21.4%	
- AgentM01	7.19	3.10	2.32	2.1%	*
- AgentM11	8.10	2.89	2.80	0.6%	**
Spielertyp im Hero Game					
- Mensch (Referenz)	–				
- AgentMx1	6.21	3.51	1.77	7.8%	†
- AgentM01	4.12	3.44	1.20	23.2%	
- AgentM11	4.35	3.58	1.22	22.5%	
Spielertyp im Prisoner's Dilemma					
- Mensch (Referenz)	–				
- AgentMx1	2.72	1.57	1.73	86%	†
- AgentM01	3.42	1.67	2.05	4.2%	*
- AgentM11	1.14	1.97	0.58	56.4%	
- Tit-for-Tat	6.60	1.63	4.06	0.0%	***
Spieltyp					
- Chicken Game (Referenz)	–				
- Hero Game	invariant				
- Prisoner's Dilemma	invariant				
Lerneffekte: Anzahl Spiele (Gegner)					
- Chicken Game	−0.16	0.71	−0.22	82.3%	
- Hero Game	1.31	0.77	1.71	9.0%	†
- Prisoner's Dilemma	3.08	0.40	7.67	0.0%	***
Konstante	38.19	1.37	27.85	0.0%	***

Tabelle A.14: Ergänzende Informationen zum integrierten Fixed Effects Panelregressionsmodell mit Interaktionseffekten zu Prestudy II und Experiment I bis III in Tabelle A.13 (indikativ: † $p < 10\%$; signifikant: * $p < 5\%$, ** $p < 1\%$, *** $p < 0.1\%$). Quelle: Eigene Darstellung.

Beobachtungen	718	
Gruppen	169	
R^2_{within}	0.17	
$R^2_{between}$	0.03	
$R^2_{overall}$	0.08	
$F(13,168)$	7.78	
Signifikanz ($\mathbf{P} > F$)	0.0%	***

Tabelle A.15: Integrierte Ergebnisse des OLS Regressionsmodells ohne Interaktionseffekte zu Prestudy II und Experiment I bis III im wiederholten Chicken Game, Hero Game und Prisoner's Dilemma (indikativ: † $p < 10\%$; signifikant: * $p < 5\%$, ** $p < 1\%$, *** $p < 0.1\%$). Quelle: Eigene Darstellung.

	Koeffizient	Standardfehler	t	p	
Spielertyp					
- Mensch (Referenz)	–				
- AgentMx1	4.23	1.58	2.68	0.8%	**
- AgentM01	4.62	1.65	2.79	0.5%	**
- AgentM11	3.87	1.78	2.17	3.0%	*
- Tit-for-Tat (nur PD)	6.93	2.67	2.60	1.0%	**
Spieltyp					
- Chicken Game (Referenz)	–				
- Hero Game	16.58	1.63	10.15	0.0%	***
- Prisoner's Dilemma	7.83	1.46	5.35	0.0%	***
Lerneffekte: Anzahl Spiele (Gegner)	1.96	0.38	5.21	0.0%	***
Konstante	29.89	1.91	15.61	0.0%	***
Beobachtungen	718				
R^2	0.17				
$R^2_{adjusted}$	0.16				
$F(7,710)$	20.63				
Signifikanz ($\mathbf{P} > F$)	0.0%				***

Tabelle A.16: Integrierte Ergebnisse des OLS Regressionsmodells mit Interaktionseffekten zu Prestudy II und Experiment I bis III (indikativ: † $p < 10\%$; signifikant: * $p < 5\%$, ** $p < 1\%$, *** $p < 0.1\%$). Quelle: Eigene Darstellung.

	Koeffizient	Standardfehler	t	p	
Spielertyp im Chicken Game					
- Mensch (Referenz)	–				
- AgentMx1	3.94	3.32	1.19	23.6%	
- AgentM01	7.15	3.30	2.17	3.1%	*
- AgentM11	8.08	3.30	2.45	1.4%	*
Spielertyp im Hero Game					
- Mensch (Referenz)	–				
- AgentMx1	6.21	3.18	1.95	5.1%	†
- AgentM01	4.12	3.18	1.29	19.6%	
- AgentM11	4.35	3.18	1.37	17.2%	
Spielertyp im Prisoner's Dilemma					
- Mensch (Referenz)	–				
- AgentMx1	3.16	2.16	1.46	14.4%	
- AgentM01	3.41	2.36	1.45	14.8%	
- AgentM11	−0.10	2.84	−0.03	97.3%	
- Tit-for-Tat	5.36	2.85	1.88	6.0%	†
Spieltyp					
- Chicken Game (Referenz)	–				
- Hero Game	12.60	4.76	2.65	0.8%	**
- Prisoner's Dilemma	0.40	4.17	0.10	92.3%	
Lerneffekte: Anzahl Spiele (Gegner)					
- Chicken Game	−0.36	0.84	−0.43	66.7%	
- Hero Game	1.38	0.80	1.74	8.3%	†
- Prisoner's Dilemma	3.01	0.49	6.14	0.0%	***
Konstante	35.10	3.47	10.13	0.0%	***

Tabelle A.17: Ergänzende Informationen zum integrierten OLS Regressionsmodell mit Interaktionseffekten zu Prestudy II und Experiment I bis III in Tabelle A.16 (indikativ: † $p < 10\%$; signifikant: * $p < 5\%$, ** $p < 1\%$, *** $p < 0.1\%$). Quelle: Eigene Darstellung.

Beobachtungen	718	
R^2	0.19	
$R^2_{adjusted}$	0.17	
$F(15,702)$	10.96	
Signifikanz ($\mathbf{P} > F$)	0.0%	***

Tabelle A.18: Interaktionsverlauf zwischen AgentM00 mit $O = (0,0)$, Prior $\hat{M}_0 = (\frac{\sqrt{2}}{1+\sqrt{2}})$ und Priorgewicht $\gamma_0 = 0$ und einem deterministisch alternierenden Gegner im wiederholten Matching Pennies Spiel.

	Aktionen		AgentM00 $i=1$			Auszahlung	
t	$\mathbf{P}[a_2^i]$	a^j	$\hat{M}_{(0,0)}$	z	$m^*(z_t)$	$\mathbf{E}[r^i]$	$\bar{\mathbf{E}}[r^j]$
1	100%	a_1^j	$(\frac{\sqrt{2}}{1+\sqrt{2}})$	–	1.00	−1.00	−1.00
2	0%	a_2^j	(0.00)	–	0.00	−1.00	−1.00
3	50%	a_1^j	(0.50)	–	0.50	0.00	−0.67
4	0%	a_2^j	(0.33)	–	0.00	−1.00	−0.75
5	50%	a_1^j	(0.50)	–	0.50	0.00	−0.60
...	...	...	...	...	...	...	...
$T-1$	50%	a_1^j	$(0.50-\varepsilon_t)$	–	0.00	0.00	−0.50
T	0%	a_2^j	(0.50)	–	0.50	−1.00	−0.50

Tabelle A.19: Interaktionsverlauf zwischen AgentM01 mit $O = (0,1)$, Prior $\hat{M}_0 = (\frac{\sqrt{2}}{1+\sqrt{2}}, \frac{\sqrt{2}}{1+\sqrt{2}})$, Eröffnungszug a_2^i und Priorgewicht $\gamma_0 = 0$ und einem deterministischen Gegner im wiederholten Matching Pennies Spiel.

	Aktionen		AgentM00 $i=1$			Auszahlung	
t	$\mathbf{P}[a_2^i]$	a^j	$\hat{M}_{(0,1)}$	z	$m^*(z_t)$	$\mathbf{E}[r^i]$	$\bar{\mathbf{E}}[r^j]$
1	a_2^i	a_1^j	$(\frac{\sqrt{2}}{1+\sqrt{2}}, \frac{\sqrt{2}}{1+\sqrt{2}})$	–	–	-1.00	-1.00
2	100%	a_1^j	$(\frac{\sqrt{2}}{1+\sqrt{2}}, \frac{\sqrt{2}}{1+\sqrt{2}})$	a_1^j	1.00	-1.00	-1.00
3	0%	a_2^j	$(\frac{\sqrt{2}}{1+\sqrt{2}}, 0.00)$	a_1^j	0.00	-1.00	-1.00
4	100%	a_1^j	$(\frac{\sqrt{2}}{1+\sqrt{2}}, 0.50)$	a_2^j	1.00	-1.00	-1.00
5	30%	a_2^j	$(0.00, 0.50)$	a_1^j	0.30	-0.40	-0.88
6	0%	a_2^j	$(0.00, 0.67)$	a_2^j	0.00	-1.00	-0.90
7	0%	a_2^j	$(0.41, 0.70)$	a_2^j	0.00	-1.00	-0.91
8	100%	a_1^j	$(0.63, 0.70)$	a_2^j	1.00	-1.00	-0.93
9	100%	a_1^j	$(0.46, 0.70)$	a_1^j	1.00	-1.00	-0.93
10	100%	a_1^j	$(0.46, 0.53)$	a_1^j	1.00	-1.00	-0.94
11	0%	a_2^j	$(0.46, 0.43)$	a_1^j	0.00	-1.00	-0.95
12	0%	a_2^j	$(0.46, 0.52)$	a_2^j	0.00	-1.00	-0.95
13	100%	a_1^j	$(0.57, 0.52)$	a_2^j	1.00	-1.00	-0.95
14	100%	a_1^j	$(0.47, 0.52)$	a_1^j	1.00	-1.00	-0.96
...	...	...	...	...	...	...	...
$T-3$	100%	a_1^j	$(0.50-\varepsilon_{t,1}, 0.50+\varepsilon_{t,2})$	a_1^j	1.00	-1.00	-1.00
$T-2$	100%	a_1^j	$(0.50-\varepsilon_{t,1}, 0.50-\varepsilon_{t,2})$	a_1^j	0.00	-1.00	-1.00
$T-1$	0%	a_2^j	$(0.50-\varepsilon_{t,1}, 0.50+\varepsilon_{t,2})$	a_2^j	0.00	-1.00	-1.00
T	0%	a_2^j	$(0.50+\varepsilon_{t,1}, 0.50+\varepsilon_{t,2})$	a_2^j	1.00	-1.00	-1.00

Tabelle A.20: Deskriptive Auswertung der durchschnittlichen normierten Auszahlungen $\bar{\rho}^i$ von AgentM11 im nachgestellten Round Robin Turnier von Axelrod (1980) zum wiederholten Prisoner's Dilemma. Quelle: Eigene Darstellung.

			Perzentil				
Algorithmus	n	$\bar{x}$	25%	$\tilde{x}$	75%	s	$\frac{s}{\bar{x}}$
AgentM11	160	52.70	44.38	60.00	60.00	12.93	0.25
Grofman	160	52.45	45.28	60.00	60.00	13.22	0.25
Stein & Rapoport	160	52.27	50.50	60.00	60.00	13.37	0.26
Tit-for-Tat	160	51.76	45.78	60.00	60.00	13.01	0.25
Shubik	160	51.65	48.10	60.00	60.00	15.05	0.29
Tideman & Chieruzzi	160	50.72	47.23	60.00	60.00	15.63	0.31
Nydegger	160	50.72	50.55	60.00	60.00	16.48	0.33
Grim Trigger	160	49.76	37.65	60.00	60.00	16.32	0.33
Davis	160	49.65	38.35	60.00	60.00	16.05	0.32
Graaskamp	160	44.82	30.00	52.05	55.60	13.64	0.30
Downing	160	42.09	21.63	35.35	59.25	22.75	0.54
Feld	160	39.65	25.65	30.70	54.10	17.05	0.43
Joss	160	36.67	22.50	27.80	50.63	16.21	0.44
Tullock	160	35.15	25.50	29.85	45.45	13.23	0.38
Anonymous	160	31.98	13.68	34.90	44.38	19.40	0.61
Random	160	31.82	13.68	35.30	44.20	19.35	0.61

Tabelle A.21: Deskriptive Auswertung der durchschnittlichen normierten Auszahlungen $\bar{\rho}^i$ von AgentMx1 im nachgestellten Round Robin Turnier von Axelrod (1980) zum wiederholten Prisoner's Dilemma. Quelle: Eigene Darstellung.

			Perzentil				
Algorithmus	n	$\bar{x}$	25%	$\tilde{x}$	75%	s	$\frac{s}{\bar{x}}$
AgentMx1	160	53.72	54.83	60.00	60.00	12.46	0.23
Stein & Rapoport	160	52.66	52.00	60.00	60.00	13.00	0.25
Grofman	160	52.18	44.28	60.00	60.00	13.68	0.26
Tit-for-Tat	160	51.44	45.30	60.00	60.00	13.48	0.26
Shubik	160	51.41	44.23	60.00	60.00	15.00	0.29
Tideman & Chieruzzi	160	50.85	46.93	60.00	60.00	15.69	0.31
Nydegger	160	50.74	50.63	60.00	60.00	16.51	0.33
Grim Trigger	160	49.77	38.45	60.00	60.00	16.16	0.32
Davis	160	49.58	38.25	60.00	60.00	16.07	0.32
Graaskamp	160	46.12	29.78	52.50	56.30	14.14	0.31
Downing	160	41.76	21.58	29.05	59.20	22.81	0.55
Feld	160	39.80	25.58	29.40	55.50	17.58	0.44
Joss	160	36.64	22.90	27.70	49.53	16.20	0.44
Tullock	160	35.05	25.08	28.45	46.03	13.47	0.38
Anonymous	160	32.53	13.78	34.80	45.28	19.60	0.60
Random	160	32.31	13.73	34.20	44.53	19.82	0.61

Tabelle A.22: Übersicht der algorithmischen Strategien des Turniers von Axelrod (1980) zum wiederholten Prisoner's Dilemma. Quelle: Eigene Darstellung.

Algorithmus	Autor
Anonymous	Unbekannt
Davis	Morton Davis
Downing	Leslie Downing
Feld	Scott Feld
Graaskamp	Jim Graaskamp
Grim Trigger	James W. Friedman
Grofman	Bernard Grofman
Joss	Johann Joss
Nydegger	Rudy Nydegger
Random	Unbekannt
Shubik	Martin Shubik
Stein & Rapoport	Stein & Anatol Rapoport
Tideman & Chieruzzi	T. Nicolaus Tideman & Paula Chieruzzi
Tit-for-Tat	Anatol Rapoport
Tullock	Gordon Tullock

B Gestaltungsmaßnahmen von der Beta- zur Vollversion

Auf Basis der Ergebnisse von Prestudy I, welcher der Erprobung der Leistungsfähigkeit und inneren Logik der Betaversion des Markovagenten diente, wurden hin zur finalen Vollversion gestalterische Veränderungen vorgenommen. Diese rekrutieren sich aus mündlichem Feedback durch die Probanden, der Analyse der Spieldaten sowie konzeptionellen und heuristischen Überlegungen. Das nachfolgende Kapitel beschriebt die Vorgenommenen Änderungen und motiviert diese kurz.

B.1 Anpassung des Fehlerlimits

Die Betaversion des Markovagenten enthielt das *Fehlerlimit* als einen Parameter, der die Anzahl der Runden, welche die Funktion Q^i zur Auswahl des passendsten Gegnermodells zurückblickt. Dabei wurde zwischen einem Fehlerlimit von 5 (AgentM-vβ I) Runden und einem Fehlerlimit von 10 (AgentM-vβ II) Runden unterschieden. Die Ergebnisse in Tabelle C.2 legen nahe, dass die Parametrisierungsalternativen keinen Einfluss auf die erreichten Spielzustände zu haben scheinen. Die zugrundeliegende Überlegung war, dass alte Vorhersagefehler die Modellauswahl verzerren könnten und daher nur hinreichend aktuelle Werte zu berücksichtigen sind. Aus Gründen der Komplexitätsreduktion des Markovagenten wurde dieser Parameter entfernt, mitunter weil in Prestudy I keine eindeutige Differenz zwischen AgentM-vβ I und II festgestellt werden konnte. Dies entspricht einem Wert von ∞ für das Fehlerlimit, da Q^i alle Runden der Interaktion zur Modellauswahl heranzieht.

B.2 Anpassung der geführten Gegnermodelle

Die Betaversion des Markovagenten führte mehre Gegnermodelle parallel. Wie Abbildung B.1 dargestellt, wird die Menge der parallel von AgentM-vβ geführten Gegnermodelle durch eine Kombinatorik an Parametrisierungen der Dimensionen *Gedächtnistiefe* und *Aktionsspeicherlimit* bestimmt. In diesem Kapitel wird die Überarbeitung dieser Gestaltungsvariablen hin zur Vollversion von AgentM vorgestellt.

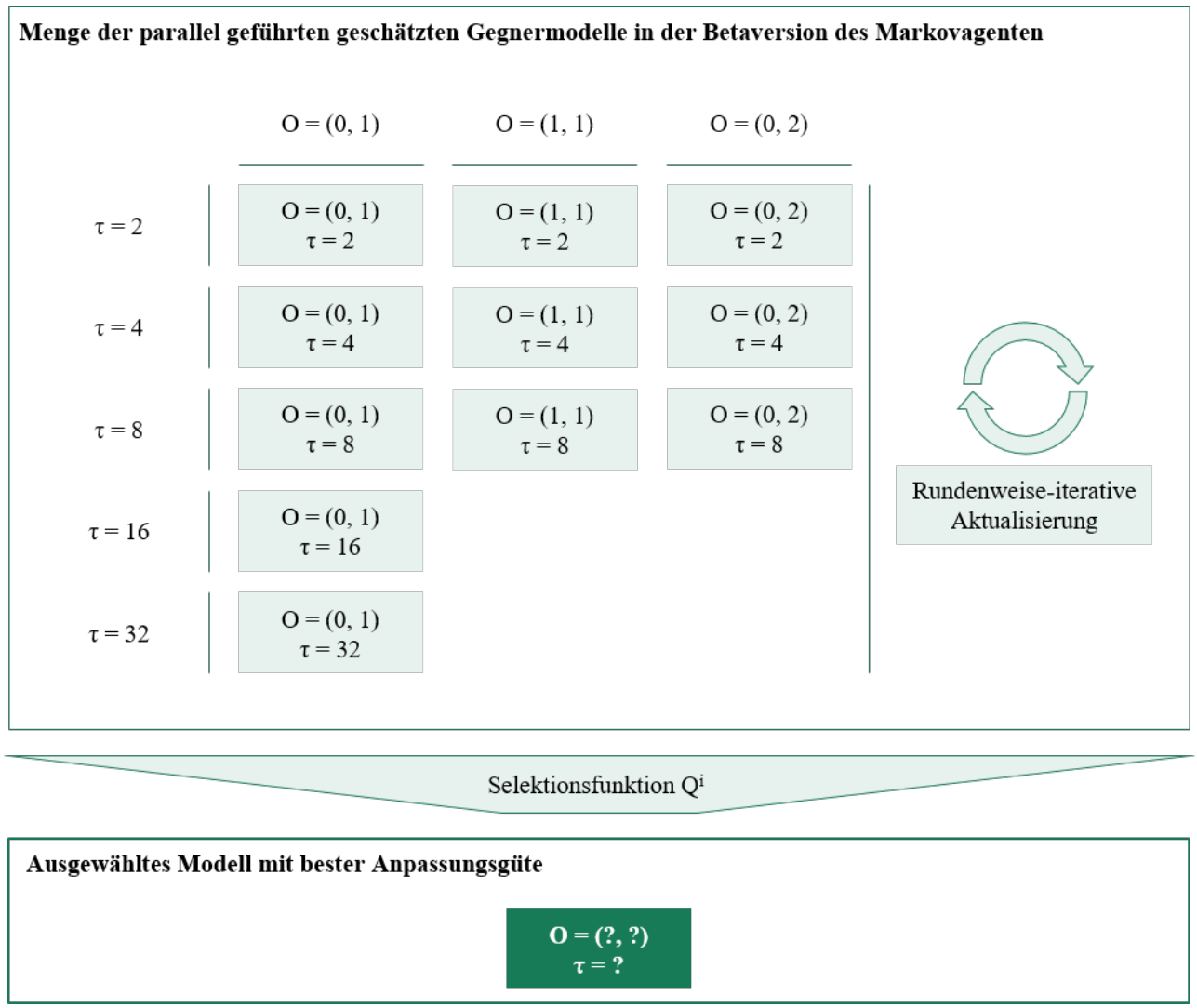


Abbildung B.1: Kombinatorik der von der Betaversion des Markovagenten parallel geführten Gegnermodelle. Quelle: Eigene Darstellung.

B.2.1 Gedächtnistiefe

AgentMxx-vβ führt simultan ein $\hat{M}_{(0,1)}$ sowie ein $\hat{M}_{(1,1)}$ und auch ein $\hat{M}_{(0,2)}$ Gegnermodell. Dementsprechend kann die Betaversion als AgentMxx-vβ bezeichnet werden. Auf Basis der Menge an potentiellen Gegnermodelle wird rundenweise anhand der besten Anpassungsgüte (siehe Kapitel 3.3.2.2) das vielversprechendste Modell auswählt und auf Basis dessen die eigene Antwort-Strategie bestimmt. Für die Vollversion des Markovagenten ist bezüglich der Gedächtnistiefe eine trennschärfere Ausgestaltung wünschenswert. Diese wird nachfolgend motiviert.

Bezüglich der von AgentMxx-vβ rundenweise ausgewählten Gegnermodelle konnte im Rahmen von Prestudy I, wie Tabelle B.1 darstellt, kein wesentlicher Unterschied in der Anpassungs-

Tabelle B.1: Verteilung der ausgewählten Gegnermodelle mit bester Anpassungsgüte für alle Runden nach der ersten gegnerischen Abweichung in Prestudy I (Zeilensumme rundungsbedingt von 100% abweichend). Quelle: Eigene Darstellung.

	Ausgewähltes Gegnermodell		
Spielertyp	$\hat{M}_{(0,1)}$	$\hat{M}_{(1,1)}$	$\hat{M}_{(0,2)}$
AgentMxx-vβ I	75%	12%	13%
AgentMxx-vβ II	76%	12%	11%

güte zwischen den beiden Varianten festgestellt werden.[68] Die deutlich häufigere Auswahl des Gegnermodells mit $O = (0,1)$ kann auf verschiedene Gründe zurückgeführt werden. Erstens wählt AgentMxx-vβ bei identischer Anpassungsgüte stets das Gegnermodell mit der geringeren Markovordnung, um so eine Überanpassung zu vermeiden. Zweitens müssen für $O = (0,1)$ lediglich $|Z_{(0,1)}| = 2$ Werte geschätzt werden, während für $O = (1,1)$ und $O = (0,2)$ jeweils ein Modell mit $|Z_{(1,1)}| = |Z_{(0,2)}| = 4$ Werten geschätzt werden muss. Infolgedessen sind für letztere Modelle mehr Beobachtungen des Gegnerverhaltens für eine belastbare Vorhersage notwendig; sie sind in Bezug auf ihren Informationsbedarf ineffizienter. Hierbei ist anzumerken, dass weniger die Gesamtzahl der Beobachtungen, sondern die Zahl der Beobachtungen *je Markovzustand* maßgeblich ist. Kommt ein Spiel nur vereinzelt in spezifische Zustände kann sich dies in einer schlechteren Anpassungsgüte niederschlagen, welche die Auswahlchancen des Gesamtmodells durch den Markovagenten senkt.

Das Führen von und der Wechsel zwischen multiplen Gegnermodellen birgt darüber hinaus Risiken. Zum einen ist bei einer größeren Auswahl an Gegnermodellen potentiell ein häufigerer Wechsel zwischen diesen möglich. In Konsequenz der sich dadurch ändernden Strategie der Betaversion des Markovagenten kann dies von menschlichen Gegnern als erratisch wahrgenommen werden, wodurch ein kausal motivierter Interaktionsmodus beeinträchtigt werden kann. Weiterhin ist eine differenzierte Aussage über den Einfluss der einzelnen Markovordnungen auf die Gesamtleistungsfähigkeit der Betaversion des Markovagenten aufgrund der bis zu rundenweise wechselnden Auswahl des Gegnermodells schwerer möglich. Infolgedessen wird für die weitere Untersuchung des Markovagenten eine klarere Ausdifferenzierung in folgende Varianten vorgenommen:

[68] Es werden lediglich Spielrunden nach einer ersten Abweichung a_2^j durch den Gegenspieler betrachtet. Alle dem vorangegangenen Runden sind aufgrund der kontinuierlichen Kooperation beider Spieler trivial in dem Sinne, dass der Schätzer einen stets kooperierenden Gegner aufgrund der fehlenden Varianz in der Aktionswahl korrekt als solchen einstuft. Erst nach der ersten gegnerischen Abweichung findet im wiederholten Prisoner's Dilemma die Markov for Tat Logik als Gegnermodell-basierte Aktionswahl durch den Markovagenten statt.

- **AgentMx1** führt parallel ein $\hat{M}^1_{(0,1)}$ sowie ein $\hat{M}^1_{(1,1)}$ Gegnermodell, aus welchen er anhand der Anpassungsgüte rundenweise das vielversprechendste Modell auswählt und auf Basis dessen die eigene Antwort-Strategie bestimmt.
- **AgentM01** führt lediglich ein $\hat{M}^1_{(0,1)}$ Gegnermodell, auf Basis dessen er rundenweise die eigene Antwort-Strategie bestimmt.
- **AgentM11** führt lediglich ein $\hat{M}^1_{(1,1)}$ Gegnermodell, auf Basis dessen er rundenweise die eigene Antwort-Strategie bestimmt.

Auf die Verwendung von Modellen mit $O = (0,2)$ wird verzichtet. Grund ist, dass die Anzahl der testbaren Varianten des Markovagenten ressourcenbedingt begrenzt ist. Zusätzlich wird die in Kapitel 3.3.1.1 beschriebene Diskrepanz zwischen Modell und tatsächlichem Entscheidungsverhalten menschlicher Spieler unter Berücksichtigung kognitiver Kapazitäten aufgrund des weiten Rückblicks der Markovkette auf 2 vorangegangene Spielrunden als unter den diskutierten Varianten am größten bewertet. Durch die Reduktion auf die zwei am plausibelsten erscheinenden Gedächtnistiefen sowie die Aufspaltung in Varianten des Markovagenten mit nur einer Gedächtnistiefe wird besseres Verständnis über einzelnen Modelltypen erwartet.

B.2.2 Aktionsspeicherlimit

Wie in Abbildung B.1 dargestellt, wurden in der Betaversion des Markovagenten neben der Gedächtnistiefe simultan diverse Gegnermodelle mit unterschiedlichen Werten für das Aktionsspeicherlimit τ Parametrisierungen geführt (siehe Kapitel 3.3.2.1). Durch die Nutzung multipler Parametrisierungen gekoppelt mit einer Auswahl der Parametrisierung mit der besten Anpassungsgüte sollte das Risiko einer mögliche Fehleinstellung von AgentMxx-vβ vermieden werden. Gleichwohl führt die Mannigfaltigkeit parallel geführter Modelle wie auch bei der Markovordnung im vorigen Kapitel zu einer erschwerten Nachvollziehbarkeit von Einflussfaktoren auf die Leistung des Markovagenten sowie zu einer potentiellen Beeinträchtigung der Nachvollziehbarkeit des Agentenverhaltens durch menschliche Gegenspieler.

Tabelle B.2: Verteilung der ausgewählten Aktionsspeicherlimits mit bester Anpassungsgüte für alle Runden nach der ersten gegnerischen Abweichung in Prestudy I (Zeilensumme rundungsbedingt von 100% abweichend). Quelle: Eigene Darstellung.

	Ausgewähltes Aktions-Limit τ				
Spielertyp	2	4	8	16	32
AgentMxx-vβ I	33%	23%	20%	0%	25%
AgentMxx-vβ II	30%	16%	16%	0%	37%

Tabelle B.2 zeigt die Verteilung der Aktionsspeicherlimits der ausgewählten Gegner-Modelle in Runden nach der ersten gegnerischen Abweichung. Obwohl der Agent Mxx-vβ bei identischer Anpassungsgüte stets das Modell mit dem größeren Aktionslimit vorzieht, um ein möglichst auflösendes Gegnermodell abzubilden, ist zusätzlich eine Tendenz zur Auswahl von Modellen mit vergleichsweise geringen Werten von τ festzustellen. Modelle mit $\tau = 16$ werden nicht ausgewählt, da sie aufgrund der Spiellänge Schätzer ähnlich zu Modellen mit $\tau = 32$ produzieren, wobei auch hier letztere Variante von AgentMxx-vβ bei gleicher Anpassungsgüte stets bevorzugt wird.

Eine genauere Analyse in Tabelle B.3 zeigt, dass hohe Aktionsspeicherlimits im Mittel eher in frühen Runden des Spiels verwendet werden. Dies kann zum einen darin begründet sein, dass sich die Modelle in frühen Runden des Spiels weniger stark unterscheiden und so bei gleicher Anpassungsgüte noch vermehrt höher auflösende Modelle ausgewählt werden.

Alles in allem kann für das Aktionsspeicherlimit analog zur Gedächtnistiefe festgestellt werden, dass eine Modellheterogenität die Nachvollziehbarkeit des Verhaltens des Markovagenten für menschliche Gegenspieler wie auch die Auswertung der generierten Spieldaten durch zusätzliche Komplexität beeinträchtigt. Für den weiterentwickelten Markovagent soll daher nur ein fester Wert für das Aktionsspeicherlimit festgelegt werden. Grundsätzlich sind hierbei zwei Tendenzen zu unterscheiden:

1. Ein zu geringer Wert kann sich schnell auf Veränderungen im Gegnerverhalten einstellen, weil wie im Beispiel von $\tau = 2$ nur zwei Runden benötigt werden um einen bestehenden Schätzwert für einen Markovzustand komplett zu erneuern.

2. Jedoch besteht hierbei die Gefahr, dass exploratives Verhalten oder Fehleingaben durch den Gegenspieler in Bezug auf dessen strategischen Verhaltens überbewertet werden und das ursprüngliche Modell tatsächlich besser nicht zu verwerfen wäre.

Unter Berücksichtigung dieser Faktoren wird für den weiterentwickelten Markovagent heuristisch ein endliches Aktionslimit mit $\tau = 10$ festgesetzt, das einerseits robust gegen Rauschen in

Tabelle B.3: Mittelwert der Spielrunde nach ausgewähltem Aktionslimit mit bester Anpassungsgüte für alle Runden nach der ersten gegnerischen Abweichung in Prestudy I. Quelle: Eigene Darstellung.

	Ausgewähltes Aktions-Limit τ				
Spielertyp	2	4	8	16	32
AgentMxx-vβ I	12.3	16.1	8.3	–	5.6
AgentMxx-vβ II	11.2	14.7	8.6	–	8.4

der Gegnerstrategie ist, aber andererseits Altlasten aus zum Beispiel einer initialen Explorationsphase des Spiels in spätere Runden schleppt.

Zusammenfassend führt also AgentMx1 zwei Gegnermodelle mit $O = (0,1)$ und $O = (1,1)$ jeweils mit $\tau = 10$, aus denen er anhand der besten Anpassungsgüte rundenweise das passendste Gegnermodell selektiert. AgentM01 führt lediglich ein Gegnermodell mit $O = (0,1)$ und $\tau = 10$, während AgentM11 lediglich ein Gegnermodell mit und $O = (1,1)$ und $\tau = 10$ führt. AgentM01 und AgentM11 wählen folglich nicht aus einer Menge von Gegnermodellen, sondern treffen a priori eine feste Annahme über das strategische Verhalten des menschlichen Gegenspielers.

B.2.3 Prior

Die Betaversion des Markovagenten verwendete nach dem Indifferenzprinzip einen Prior von $\hat{m}_0 = \frac{1}{2}$ und ein Priorgewicht von $\gamma_0 = 0$. Ersteres entspricht einer gleich gewichteten Erwartung bezüglich der Aktionen des Gegners. Zweiteres entspricht einem sofortigem Update der einzelnen Werte bei Ankunft der ersten Information je Zustand. Während Prestudy I erhärtete sich der Verdacht, dass eine derartige Parametrisierung regelmäßig dazu führen kann, dass der Markovagent sich in einer Abweichungsspirale mit dem Gegner begibt. Unglücklicherweise ist es durch diesen Login bisweilen nicht mehr möglich, den Markovspieler in andere Zustände zu bringen, die ihm helfen, das noch stark auf Priorwerten basierende Gegnermodell adäquat zu aktualisieren. Infolgedessen wurde für die Vollversion die Verwendung eines empirischen Priors je Spieltyp, mit einem graduellem Update im Sinne von $\gamma_0 = 1$ gewählt. Kapitel 3.3.2.1 beschreibt dies ausführlicher.

B.3 Zusammenfassung

Zusammenfassend wurden auf Basis der Erkenntnisse von Prestudy I folgende Anpassungen an der Betaversion des Markovagenten vorgenommen:

- **Gegnermodelle:** Mit dem Ziel der Komplexitätsreduktion weist der endgültige Markovagent in Bezug auf mögliche Gegnermodelle im Vergleich zu Betaversion eine wesentlich klarere Struktur auf. Dafür wurde die Parametrisierung der Gedächtnistiefe und des Aktionsspeicherlimits vereinfacht. Jeder Markovagent führt demnach nur noch zwei Gegnermodelle (AgentMx1), beziehungsweise ein einziges Gegnermodell (AgentM01 und AgentM11).
  - **Gedächtnistiefe:** Der weiterentwickelte Markovagent verwendet lediglich die Markov-Ordnungen $O = (0,1)$ oder $O = (1,1)$ sowie eine Kombination der beiden, sodass

die Markovzustände ausschließlich auf den Zuginformationen der Runde bedingen (siehe Kapitel 3.3.1.1.

  - **Aktionsspeicherlimit:** Zur Komplexitätsreduktion verwendet der weiterentwickelte Markovagent einzig das Aktionsspeicherlimit von $\tau = 10$, sodass je Markovzustand die bis zu zehn letzten Zuginformationen zur Berechnung der Wahrscheinlichkeitsschätzer verwendet werden (siehe Kapitel 3.3.2.1).

- **Strategiedatenbank:** Die Berechnung der Strategiedatenbank wurde an die reduzierte Gedächtnistiefe angepasst. Weiterhin orientieren sich die gesampelten Wahrscheinlichkeiten nun an den real beobachtbaren Werten der Wahrscheinlichkeitsschätzer bei einem Aktionsspeicherlimit von $\tau = 10$ (siehe Kapitel 3.3.2.3).[69]

- **Fehlerlimit:** Die Berechnung der Anpassungsgüte zur Wahl des passenden Gegnermodells findet im weiterentwickelten Markovagent auf der gesamten Vorhersagehistorie statt. Dies entspricht einem Fehlerlimit von ∞ und ist analog zu dem Wegfall des Parameters.

- **Prior:** Der weiterentwickelte Markovagent verwendet empirisch erhobene Prior aus menschlichen Interaktionen je Spieltyp, die als gleichwertige Zuginformation graduell statt sofortig aktualisiert werden (siehe Kapitel 3.3.2.1).

[69] Zuvor wurden die Werte von m der Markovstrategien für die Bestimmung der Strategiedatenbank anhand eines in τ äquidistant aufgeteilten Wahrscheinlichkeitsintervalls $[0, 1]$ bestimmt.

C Validierung der Betaversion des Markovagenten in Prestudy I

Im folgenden Kapitel werden die Ergebnisse der initialen Validierung der Betaversion des Markovagenten AgentMxx-vβ vorgestellt. Die Untersuchung findet im Rahmen einer Prestudy statt, die ressourcenschonend eine erste Indikation über die Leistung des Markovagenten geben soll, bevor sich ein vollumfängliche Untersuchung über mehrere Spiele anschließt. Eine Übersicht der verschiedenen Versionierungen und deren empirische Untersuchung findet sich in Tabelle 4.5. Ziel ist es dabei insbesondere, ressourceneffizient eine erste Indikation über die grundsätzliche Leistungsfähigkeit des Markovagenten zu generieren und mögliche Anpassungsbedarfe zu identifizieren.

C.1 Experimenthergang

Das Experiment zu *Prestudy I* zur initialen Validierung der Betaversion der Markovagenten im wiederholten Prisoner's Dilemma fand am 11. und 14. Januar 2019 als Teil der Lehrveranstaltung *Organisationsmanagement*[70] am Institut für Unternehmensführung des Karlsruher Instituts für Technologie statt. Die 71 registrierten Teilnehmer wurden über die beiden Tage auf

[70] Kapitel 4.5 zeigt Besonderheiten hinsichtlich des organisatorischen Kontextes auf.

11. Januar 2019

09:00 – 10:15 Sitzung 1 Registriert: 7 Teilgenommen: 7	**11:30 – 12:45** Sitzung 2 Registriert: 9 Teilgenommen: 8	**14:30 – 15:45** Sitzung 3 Registriert: 10 Teilgenommen: 10	**17:00 – 15:15** Sitzung 4 Registriert: 7 Teilgenommen: 6

14. Januar 2019

09:00 – 10:15 Sitzung 5 Registriert: 9 Teilgenommen: 6	**11:30 – 12:45** Sitzung 6 Registriert: 9 Teilgenommen: 9	**14:30 – 15:45** Sitzung 7 Registriert: 10 Teilgenommen: 8	**17:00 – 15:15** Sitzung 8 Registriert: 10 Teilgenommen: 9

Abbildung C.1: Organisatorische Übersicht des Experiments zu Prestudy I. Quelle: Eigene Dastellung.

acht Sitzungen von je 75 Minuten mit bis zu 10 Teilnehmern verteilt. Insgesamt erschienen 63 Teilnehmer zum Experiment. Abbildung C.1 stellt den Sachverhalt dar.[51] Die Aufteilung und Abstände zwischen den Sitzungen erfolgt der Logik von Chinczewski (2019) und hat zum Ziel, die Kommunikation zwischen den Probanden zu minieren und den Versuchsablauf robuster gegen eventuelle technische oder prozessuale Komplikationen zu machen.

Innerhalb der Sitzungen spielt jeder Teilnehmer gegen fünf Gegenspieler. Die Paarungslogik erfolgt nach Kapitel 4.1.2. Die daraus resultierende Paarung gliedert sich wie in Tabelle C.1 dargestellt. Jeder Proband spielt in zufälliger Reihenfolge nach folgendem Aufbau:

- In der Rolle des Sparringspielers gegen einen anderen Menschen.
- In der Rolle des Sparringspielers gegen den AgentMxx-vβ I.
- In der Rolle des Sparringspielers gegen den AgentMxx-vβ II
- In der Rolle des Sparringspielers gegen einen Tit-for-Tat Algorithmus.
- In der Rolle als beobachteter Spieler gegen einen menschlichen Sparringspieler, welcher nicht dem vorigen menschlichen Partner entspricht.

Die Betaversion des Markovagenten ist gemäß Kapitel B parametrisiert, wobei AgentMxx-vβ I mit bei der Auswahl des passenden Gegnermodells die Anpassungsgüte über die letzten 5 Runden berechnet, während AgentMxx-vβ II dafür die letzten 10 Runden miteinbezieht. Die Unterscheidung wurde zur Evaluation möglicher Leistungsimplikationen des Parameters getroffen. Der Tit-for-Tat-Spieler wurde als zusätzlicher Benchmark zum Spielertyp Mensch aufgrund seiner auf Axelrod (1984) zurückgehenden Verbreitung im wiederholten Prisoner's Dilemma gewählt.

Aus der Menge der 63 Teilnehmer konnten die Daten der 8 Teilnehmer aus Sitzung 2 nicht verwendet werden. Während der Sitzung kam es zu einem nicht sofort korrigierbaren Absturz

Tabelle C.1: Spielerpaarungen der Prestudy I gemäß Kapitel 4.1.2. Quelle: Eigene Darstellung.

Spiele für Proband A	Beobachteter Spieler	Sparringspieler	Kohorte
1	Proband B	Proband A	Sparringspieler A
2	AgentMxx-vβ I	Proband A	
3	AgentMxx-vβ II	Proband A	
4	Tit-for-Tat	Proband A	
5	Proband A	Proband C	Sparringspieler C
	...	...	

des Experiment-Servers, sodass die Sitzung vorzeitig abgebrochen werden musste. In Folge konnten für insgesamt 55 Spieler entsprechend 55 Kohorten-Datensätze mit jeweils Daten zu den 4 beobachteten Spielertypen erzeugt werden.

C.2 Deskriptive Analyse

Erste Anhaltspunkte über das Spielverhalten der verschiedenen Spielertypen in Prestudy I soll eine deskriptive Analyse liefern. Dabei werden Aktionsverhalten der Spieler,sowie die mit dem Gegner realisierten Zustände betrachtet, um Aufschluss über deren Sentiment zu geben. Im zweiten Schritt findet eine Untersuchung statistischer Momente der normierten durchschnittlichen Auszahlung statt.

C.2.1 Aktionsverhalten und realisierte Zustände

Die in Tabelle C.2 veranschaulichte Analyse der von den verschiedenen Spielertypen erreichten Spielzustände deutet auf Unterschiede im jeweiligen Spielverhalten hin. Die Markovagenten fallen mit dem höheren Anteil an a_2^i Aktionen hierbei durch eine im Vergleich zu den Spielertypen Mensch und Tit-for-Tat aggressivere Spielweise auf. Infolgedessen werden sie in $a_1^i a_2^j$

Tabelle C.2: Verteilung der Aktionswahl und der erreichten Spielzustände der Spielertypen nach aufsteigend sortierter normierter Auszahlung in Prestudy I zum Prisoner's Dilemma; mit Nash-Gleichgewicht des Stufenspiels (N), paretoeffizientem Zustand (grüne Farbe) und paretodominiertem Zustand (rote Farbe); exemplarisch um Kooperation (C) und Abweichung (D) ergänzt. Quelle: Eigene Darstellung.

	Aktionswahl				Realisierter Spielzustand			
	Spieler		Gegner (Mensch)					
	a_1^i	a_2^i	a_1^j	a_2^j	$a_1^i a_2^j$	$a_2^i a_2^j$	$a_1^i a_1^j$	$a_2^i a_1^j$
	C	D	C	D	CD	DD	CC	DC
Spielertyp								
- AgentMxx-vβ I	71%	29%	72%	28%	4%	24%	66%	6%
- AgentMxx-vβ II	71%	29%	72%	28%	5%	23%	66%	6%
- Mensch	76%	24%	74%	26%	8%	18%	67%	7%
- Tit-for-Tat	80%	20%	79%	21%	7%	14%	73%	6%
Normierte Auszahlung								
- Spieler					0	20	60	100
- Gegner (Mensch)					100	20	60	0
Eigenschaft						N		

seltener ausgebeutet und erreichen stattdessen häufiger das attraktivere Nash-Gleichgewicht $a_2^i a_2^j$ des one-shot Spiels. Die aggressivere Spielweise resultiert jedoch gleichzeitig in einem selteneren Erreichen der beidseitigen Kooperationslösung $a_1^i a_1^j$ als paretooptimaler Zustand des one-shot Spiels. Der potentielle Zugewinn einer womöglich häufigeren Ausbeutung des Gegners in $a_2^i a_1^j$ konnte jedoch nicht häufiger realisiert werden als von den kooperativen Spielern. Die Auswirkungen der Differenzen auf den erzielten normierten Payoff werden im nächsten Kapitel betrachtet.

C.2.2 Deskriptive Betrachtung der normierten Auszahlung

Die durchschnittlichen Auszahlungen der Spielertypen in Tabelle C.3 liegen vergleichsweise nah beieinander, wobei sich allein der Tit-for-Tat-Spieler etwas von den anderen Spielertypen leicht abhebt. Ursache ist die streng reziproke Spielweise der Strategie, welche wie die in Tabelle C.2 dargestellt in einer wesentlich öfteren Erreichung der attraktiveren Kooperationslösung bei seltenerem Erreichen der Nash-Lösung des one-shot Spiels resultiert (vgl. Axelrod, 1984). Der Tit-for-Tat-Spieler scheint das leicht bessere Ergebnis mit einer geringeren Standardabweichung auch zuverlässiger zu erzielen als die anderen Spielertypen. Gleichwohl ist anzumerken, dass Tit-for-Tat speziell für das wiederholte Prisoner's Dilemma konzipiert wurde, während sich die anderen Spielertypen durch eine spielübergreifende Adaptionsfähigkeit auszeichnen. Insofern ist die Diskrepanz der anderen Spieler zu Tit-for-Tat intuitiv als vergleichsweise gering einzuschätzen. Weiterhin fällt auf, dass der AgentMxx-vβ nur äußerst knapp hinter der Leistung der menschlichen Spieler liegt. Insofern gibt die deskriptive Analyse bereits Hinweise auf eine valide und vergleichsweise performante Konzeption der Betaversion.

Tabelle C.3: Deskriptive Auswertung der durchschnittlichen normierten Auszahlungen $\bar{\rho}^i$ der beobachteten Spielertypen für Prestudy I. Quelle: Eigene Darstellung.

			Perzentile				
Spielertyp	n	$\bar{x}$	25%	$\tilde{x}$	75%	s	$\frac{s}{\bar{x}}$
AgentMxx-vβ I	55	50.20	34.29	60.00	60.00	14.19	0.28
AgentMxx-vβ II	55	50.13	40.95	60.00	60.00	13.96	0.28
Mensch	55	50.55	44.76	60.00	60.00	14.73	0.29
Tit-for-Tat	55	52.61	48.57	59.05	60.00	11.96	0.23

C.3 Hypothesentests

Im Folgenden werden die Ergebnisse von Prestudy I zur initialen Validierung der Betaversion des Markovagenten im wiederholten Prisoner's Dilemma anhand von Hypothesentests präsen-

tiert. Der zu zeigende Sachverhalt ist eine signifikante positive Abweichung der Leistung des Markovagenten im Vergleich der von menschlichen Spielern. Ausgangsbasis für sämtliche Tests ist die Datenlage aus Tabelle C.3.

C.3.1 Zentrale Analysen

Aus den parametrischen Verfahren wird ein zwei-Stichproben t-Tests weiterhin wegen unzureichender Unabhängigkeit der Beobachtungen innerhalb der Spielerkohorten als unpassend eingestuft (vgl. Toutenburg & Heumann, 2008, S. 142-145). Ein gepaarter t-Test zur Betrachtung der Payoffdifferenz zwischen Spielertypen ist aufgrund eingeschränkter Normalverteilung der Differenzwerte nur eingeschränkt möglich (vgl. Toutenburg & Heumann, 2008, S. 145-147). Tabelle C.4 stellt diesen Sachverhalt auf Basis des Shapiro-Wilk-Tests (vgl. Royston, 1992) mit Anpassung durch Shapiro und Wilk (1965) dar. Die Nullhypothese, dass bei der Differenz der Auszahlungswerte eine Normalverteilung vorliegt konnte bei einem Signifikanzniveau von 5% für alle paarweisen Vergleiche verworfen werden.

Aus den nicht-parametrischen Verfahren wird der Mann-Whitney-U-Test gleich dem zwei-Stichproben t-Test aufgrund der fehlenden Unabhängigkeit der Beobachtungen ausgeschlossen (Mann & Whitney, 1947).

Gleichwohl kommt der Wilcoxon-Vorzeichen-Rang-Test (Wilcoxon, 1945, 1947) in Frage, welcher jedoch statt Unabhängigkeit der Beobachtungen eine *symmetrische Verteilung der Differenz* der Beobachtungen erfordert (vgl. Toutenburg & Heumann, 2008, S. 182). Die Nullhypothese bezüglich der Symmetrie der Differenz der paarweise nach Sparringspielern gruppierten Beobachtungen kann anhand des Tests von D'Agostino et al. (1990) mit Anpassung von Royston (1991) nicht verworfen werden (siehe Tabelle C.5). Folglich wird keine Verletzung der Symmetrie-Annahme angenommen und der Einsatz des Wilcoxon-Vorzeichen-Rang-Tests ist möglich.

Tabelle C.4: Ergebnisse des Shapiro-Wilk-Tests (Shapiro & Wilk, 1965) auf H_0, dass bei der Differenz der Auszahlungswerte $D = X_1 - X_2$ für Prestudy I eine Normalverteilung vorliegt (rote Kennzeichnung für verworfene Normalverteilungsannahme (NvA); indikativ: † $p < 10\%$; signifikant: * $p < 5\%$, ** $p < 1\%$, *** $p < 0.1\%$). Quelle: Eigene Darstellung.

Vergleich				
Spielertyp 1	Spielertyp 2	p		H_0
AgentMxx-vβ I	Mensch	0.0%	***	×
AgentMxx-vβ II	Mensch	0.0%	***	×
Tit-for-Tat	Mensch	0.0%	***	×

Tabelle C.5: Ergebnisse des Tests auf Symmetrie der Differenzen (D'Agostino et al., 1990; Royston, 1991) mit H_0, dass Auszahlungsdifferenzen $D = X_1 - X_2$ für Prestudy I symmetrisch sind (rote Kennzeichnung für verworfene Symmetrieannahme (SA); indikativ: † $p < 10\%$; signifikant: * $p < 5\%$, ** $p < 1\%$, *** $p < 0.1\%$). Quelle: Eigene Darstellung.

Vergleich			
Spielertyp 1	Spielertyp 2	p	H_0
AgentMxx-vβ I	Mensch	13.6%	✓
AgentMxx-vβ II	Mensch	93.2%	✓
Tit-for-Tat	Mensch	28.2%	✓

Tabelle C.6: Ergebnisse des Wilcoxon-Vorzeichen-Rang-Tests (Wilcoxon, 1945) für gepaarte Daten aus Prestudy I auf H_0, dass zwei Stichproben aus der gleichen Verteilung gezogen wurden und demnach der Median der Auszahlungsdifferenzen $D = X_1 - X_2$ Null ist (grüne Kennzeichnung für positive Auszahlungsdifferenzen; indikativ: † $p < 10\%$; signifikant: * $p < 5\%$, ** $p < 1\%$, *** $p < 0.1\%$) unter Berücksichtigung der Erfüllung der zugrundeliegenden Symmetrieannahme (SA) aus Tabelle C.5. Quelle: Eigene Darstellung.

Vergleich							
Spielertyp 1	Spielertyp 2	$\bar{D}$	$\tilde{D}$	z	p	H_0	SA
AgentMxx-vβ I	Mensch	−0.35	0.00	0.19	85.0%	✓	✓
AgentMxx-vβ II	Mensch	−0.42	0.00	0.76	44.5%	✓	✓
Tit-for-Tat	Mensch	2.06	0.00	0.49	62.2%	✓	✓

Tabelle C.6 zeigt die Ergebnisse des Wilcoxon-Vorzeichen-Rang-Tests mit Signifikanzniveau von 5%. Die Nullhypothese H_0, dass der Median der Auszahlungsdifferenzen im paarweisen Vergleich Null konnte für keinen der paarweisen Vergleiche verworfen werden.

Zusammenfassend lässt festhalten, dass die Leistung von AgentMxx-vβ auf Basis der Hypothesentests nicht signifikant von der Leistung der anderen Spielertypen abweicht. Dies kann als Indiz für die Validität des Konzeptes interaktiver Markovspieler gewertet werden. Erstens wurde bereits im Prototypenstadium die Leistungsfähigkeit von Menschen erreicht. Zweitens, liegt die Leistungsfähigkeit ebenfalls nicht signifikant hinter einer Tit-for-Tat-Strategie. Hierbei ist verstärkend anzumerken, dass es sich bei Tit-for-Tat um eine Speziallösung für das wiederholte Prisoner's Dilemma handelt, während die Betaversion des Markovagenten und die menschlichen Spieler sich individuell auf die Gegebenheiten eines jeden 2x2 Spiels anpassen können. Nach der Präsentation zusätzlicher Robustheitsanalysen anhand der ausgeschlossenen Hypothesentests findet eine Regressionsanalyse statt.

C.3.2 Zusätzliche Robustheitsanalysen

Die Ergebnisse der Testhypothese werden durch Anwendung der aus in Kapitel C.3.1 genannten Verletzungen der Annahmen verworfenen Tests bestätigt. Der zwei-Stichproben t-Test wird in Tabelle C.7 dargestellt, während der gepaarte t-Test in Tabelle C.8 aufgeführt ist. Der Mann-Whitney-U-Test wiederum findet sich in Tabelle C.9 wieder. Die Ergebnisse der Robustheitsanalysen decken sich weitestgehend mit den Ergebnissen der Testhypothese und unterstreichen somit deren Validität.

Tabelle C.7: Ergebnisse des zwei-Stichproben t-Tests für unabhängige Daten (Student, 1908) zu Prestudy I auf H_0, dass die Mittelwerte beider Spielertypen identisch sind (grüne Kennzeichnung für positiv, rote für negativ unterschiedliche Mittelwerte; indikativ: † $p < 10\%$; signifikant: * $p < 5\%$, ** $p < 1\%$, *** $p < 0.1\%$) unter Berücksichtigung der Erfüllung der zugrundeliegenden Unabhängigkeitsannahme (UA). Quelle: Eigene Darstellung.

Vergleich							
Spielertyp 1	Spielertyp 2	$\bar{x}_1$	$\bar{x}_2$	t	p	H_0	UA
AgentMxx-vβ I	Mensch	50.20	50.55	−0.13	90.0%	✓	×
AgentMxx-vβ II	Mensch	50.13	50.55	−0.15	88.0%	✓	×
Tit-for-Tat	Mensch	52.61	50.55	0.81	42.2%	✓	×

Tabelle C.8: Ergebnisse des gepaarten t-Tests (Student, 1908) für Daten aus Prestudy I auf H_0, dass die Mittelwerte beider Spielertypen identisch sind (grüne Kennzeichnung für positiv unterschiedliche Mittelwerte; indikativ: † $p < 10\%$; signifikant: * $p < 5\%$, ** $p < 1\%$, *** $p < 0.1\%$) unter Berücksichtigung der Erfüllung der zugrundeliegenden Normalverteilungsannahme (NvA) der paarweisen Differenzen aus Tabelle C.4. Quelle: Eigene Darstellung.

Vergleich							
Spielertyp 1	Spielertyp 2	$\bar{x}_1$	$\bar{x}_2$	t	p	H_0	NvA
AgentMxx-vβ I	Mensch	50.20	50.55	−0.13	89.6%	✓	×
AgentMxx-vβ II	Mensch	50.13	50.55	−0.18	85.5%	✓	×
Tit-for-Tat	Mensch	52.61	50.55	0.81	41.9%	✓	×

Tabelle C.9: Ergebnisse des Mann-Whitney-U-Tests (Mann & Whitney, 1947) für Prestudy I auf H_0, dass zwei unabhängige Stichproben aus der gleichen Verteilung gezogen wurden und demnach den gleichen Median aufweisen (grüne Kennzeichnung für positiv unterschiedliche Mediane; indikativ: † $p < 10\%$; signifikant: * $p < 5\%$, ** $p < 1\%$, *** $p < 0.1\%$) unter Berücksichtigung der Erfüllung der zugrundeliegenden Unabhängigkeitsannahme (UA). Quelle: Eigene Darstellung.

Vergleich							
Spielertyp 1	Spielertyp 2	$\tilde{x}_1$	$\tilde{x}_2$	z	p	H_0	UA
AgentMxx-vβ I	Mensch	60.00	60.00	0.67	50.5%	✓	×
AgentMxx-vβ II	Mensch	60.00	60.00	−0.02	98.7%	✓	×
Tit-for-Tat	Mensch	59.05	60.00	0.17	86.5%	✓	×

C.4 Regressionsanalyse

Im Folgenden werden die Ergebnisse von Prestudy I zur initialen Validierung der Betaversion des Markovagenten im wiederholten Prisoner's Dilemma anhand von Regressionsanalysen präsentiert. Der zu zeigende Sachverhalt ist eine signifikante positive Abweichung der Leistung des Markovagenten im Vergleich der von menschlichen Spielern unter Berücksichtigung von Einflüssen wie Lerneffekten auf Seiten der menschlichen Spieler sowie die Abhängigkeit der Leistung vom Spielverhalten des Gegenspielers.

C.4.1 Zentrale Analysen

Die Daten der Prestudy I werden im Rahmen einer Panelregression analysiert, welche sich an der Modellierungslogik von Prestudy II orientiert (siehe Kapitel 5.1.4). Zusammenfassend ergibt sich das folgende Design:

- **Abhängige Variable:** Erklärt werden soll der normierte Payoff (siehe Kapitel 4.4) der betrachteten Spieler als Maß für dessen Fähigkeit in den entsprechenden Spiele zielführend zu agieren.
- **Panelvariable:** Der menschliche Gegenspieler wird in Einklang mit dem Experimenthergang in Tabelle C.1 (analog zu Abbildung 4.2) zur Bildung von Panelkohorten genutzt. So kann die Abhängigkeit zwischen Payoffs der beobachteten Spielers und der Strategie des Gegenspielers (siehe Kapitel 4.1.2.1) berücksichtigt werden.
- **Kontrollvariablen:** Das verwendete Regressionsmodell kontrolliert für folgende erklärende Variablen:
  - **Wie beeinflusst der Spielertyp des beobachteten Spielers dessen Leistung?** Durch die Verwendung von Dummyvariablen wird berücksichtigt, ob es sich bei

dem beobachteten Spieler um den Typ *Mensch*, *AgentMxx-vβ I*, *AgentMxx-vβ I* oder Tit-for-Tat handelt. Der beobachtete Mensch dient als Vergleichskategorie, sodass die Regressionskoeffizienten der anderen Spielertypen als relative Leistungsdifferenz zu diesem zu verstehen ist.

  - **Wie beeinflusst die Erfahrung des menschlichen Gegenspielers die Leistung des beobachteten Spielers?** Um temporale Lerneffekte der Gegenspieler zu berücksichtigen, findet darüber hinaus die Zählvariable *Anzahl Spiele (Gegner)* Einfluss in das Modell. Sie gibt an, auf das wievielte Spiel für den beobachteten Menschen sich die Spiel-Performance sich bezieht.

- **Modelltyp:** Zur Steigerung einer Varianzaufklärung innerhalb der Vergleichsgruppe wird ein Fixed Effects Modell unter Berücksichtigung clusterrobuster Standardfehler verwendet (vgl. Das, 2019, S. 481). Die Verwendung eines Fixe Effects Modells wird durch die vorangegangenen qualitativen Überlegungen zur Abhängigkeit des Payoffs von der panelstrukturierenden Gegnerstrategie unterstützt.[71]

Zusammenfassend wird die nachfolgende Regression wird mit clusterrobusten Standardfehlern als Fixed Effects Modell und Dummyvariablen für die verschiedenen Spielertypen unter Kontrolle für Lerneffekte durch eine Zählvariable für die Anzahl der Spiele der menschlichen Gegenspieler gerechnet. Als Panelvariable wurde jeweils der gegnerische menschliche Spieler verwendet. Tabelle C.10 stellt die Ergebnisse der Auswertung dar.

Das insgesamt signifikante Modell zeigt, dass die Leistung beider AgentMxx-vβ für das wiederholte Prisoner's Dilemma in Prestudy I nicht signifikant von der Leistung der menschlichen Referenzspieler abweicht. Gleichwohl wurde ein hochsignifikanter positiver Lerneffekt auf seiten der menschlichen Gegenspieler als Einflussfaktor auf die Leistung der beobachteten Spieler festgestellt. Im Schnitt steigt der durch Menschen erzielte normierte Payoff mit jedem Spiel einer Session um 4.56.

Zwischenfazit

Weder Tit-for-Tat, noch AgentMxx-vβ I und II können in Prestudy I zum wiederholten Prisoner's Dilemma signifikant bessere Ergebnisse als menschliche Referenzspieler erzielen.

Vor der Weiterentwicklung des Markovagenten durch Identifikation un Implementierung von Verbesserungspotentialen soll zunächst die Robustheit anhand alternativer Modelle in Kapitel C.4.2 sichergestellt werden. So wird den anderslautenden Empfehlungen quantitativer Spezifikationstests wie dem gewöhnlichen Hausman Test (Hausman, 1978), dem robusten Hausman

[71] Die Ergebnisse quantitativer Spezifikationstests werden in Tabelle C.11 dargestellt.

Tabelle C.10: Ergebnisse des Fixed Effects Panelregressionsmodells mit clusterrobusten Standardfehlern (vgl. Das, 2019, S. 482) zu Prestudy I im wiederholten Prisoner's Dilemma (indikativ: † $p < 10\%$; signifikant: * $p < 5\%$, ** $p < 1\%$, *** $p < 0.1\%$). Quelle: Eigene Darstellung.

	Koeffizient	Standardfehler	t	p	
Spielertyp					
- Mensch (Referenz)	–				
- AgentMxx-vβ I	-0.26	2.15	-0.12	90.3%	
- AgentMxx-vβ II	-075	1.92	-0.39	69.8%	
- Tit-for-Tat	2.72	2.16	1.26	21.2%	
Lerneffekte: Anzahl Spiele (Gegner)	4.56	0.60	7.60	0.0%	***
Konstante	36.78	2.58	14.27	0.0%	***
Beobachtungen	220				
Gruppen	55				
R^2_{within}	0.34				
$R^2_{between}$	0.00				
$R^2_{overall}$	0.21				
$F(4,54)$	19.23				
Signifikanz ($\mathbf{P} > F$)	0.0%				***

Test (Greene, 2008; Hoechle, 2007; Wooldridge, 2009) und dem (robusten) Sargan-Hansen Test (Arellano, 1993; Wooldridge, 2002) Rechnung getragen. Es finden diverse Random Effects Analysen zur Berücksichtigung zeitinvarianter unbeobachteter Heterogenität Eingang. Außerdem wird eine Ordinary Least Squares (OLS) Regression durchgeführt, welche den Einfluss des des Gegenspielers als Panelvariable vernachlässigt.

C.4.2 Zusätzliche Robustheitsanalysen

Quantitative Tests für Empfehlungen zur Gestaltung des Modells legen ein Random Effects Modell nahe (siehe Tabelle C.11). Die in Kapitel C.4.1 festgestellte, aus qualitativer Sicht bessere Eignung eines Fixed Effects Modells ist dies jedoch nicht abträglich. Insbesondere die zur Gewährleistung der Modellrobustheit in Tabelle C.12 dargestellten Ergebnisse einer Random Effects Regression bestärken die Validität des gewählten Fixed Effects Ansatzes anhand konsistenter Ergebnisse. Tabelle C.13 zeigt die Ergebnisse einer OLS Regression, welche keine Paneleffekte berücksichtigt.

Tabelle C.11: Ergebnisse von Tests zur Modellspezifikation der Panelregression zu Prestudy I (* $p < 0.1$, *** $p < 5\%$, *** $p < 1\%$). Quelle: Eigene Darstellung.

Test	Statistik	p	Empfohlenes Modell
Hausman Test	$\chi^2(4) = 1.00$	91.0%	Random Effects
Robuster Hausman Test	$F(4,54) = 1.66$	17.2%	Random Effects
Sargan-Hansen Test	$\chi^2(1) = 1.00$	31.8%	Random Effects
Robuster Sargan-Hansen Test	$\chi^2(1) = 1.01$	29.9%	Random Effects

Tabelle C.12: Ergebnisse des Random Effects Panelregressionsmodells (vgl. Das, 2019, S. 494) zu Prestudy I im wiederholten Prisoner's Dilemma (indikativ: † $p < 10\%$; signifikant: * $p < 5\%$, ** $p < 1\%$, *** $p < 0.1\%$). Quelle: Eigene Darstellung.

	Koeffizient	Standardfehler	z	p	
Spielertyp					
- Mensch (Referenz)	–				
- AgentMxx-vβ I	−0.26	1.94	−0.14	89.2%	
- AgentMxx-vβ II	−0.74	1.94	−0.38	70.3%	
- Tit-for-Tat	2.71	1.94	1.40	16.3%	
Lerneffekte: Anzahl Spiele (Gegner)	4.48	0.50	8.99	0.0%	***
Konstante	37.01	2.24	16.52	0.0%	***
Beobachtungen	220				
Gruppen	55				
R^2_{within}	0.34				
$R^2_{between}$	0.00				
$R^2_{overall}$	0.21				
$\chi^2(4)$	83.01				
Signifikanz ($\mathbf{P} > F$)	0.0%				***

Tabelle C.13: Ergebnisse des OLS Regressionsmodells zu Prestudy I im wiederholten Prisoner's Dilemma (indikativ: † $p < 10\%$; signifikant: * $p < 5\%$, ** $p < 1\%$, *** $p < 0.1\%$). Quelle: Eigene Darstellung.

	Koeffizient	Standardfehler	t	p	
Spielertyp					
- Mensch (Referenz)	–				
- AgentMxx-vβ I	-0.27	2.35	-0.11	90.9%	
- AgentMxx-vβ II	-0.73	2.35	-0.31	75.6%	
- Tit-for-Tat	2.69	2.35	1.15	25.3%	
Lerneffekte: Anzahl Spiele (Gegner)	4.35	0.59	7.37	0.0%	***
Konstante	37.42	2.43	15.37	0.0%	***
Beobachtungen	220				
R^2	0.21				
$R^2_{adjusted}$	0.19				
$F(4,215)$	13.97				
Signifikanz ($\mathbf{P} > F$)	0.0%				***

Literatur

Agrawal, A. & Jaiswal, D. (2012). *When Machine Learning Meets AI and Game Theory*. Stanford University. Stanford, CA.

Aiken, L. S. & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA, Sage Publications.

Albrecht, S. V. & Stone, P. (2018). Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, *258*, 66–95.

Allais, M. (1979). The Foundations of a Positive Theory of Choice Involving Risk and a Criticism of the Postulates and Axioms of the American School (1952). In M. Allais & O. Hagen (Hrsg.), *Expected Utility Hypotheses and the Allais Paradox* (S. 27–145). Dordrecht, Springer Netherlands.

Andreoni, J. (1988). Why free ride? *Journal of Public Economics*, *37*(3), 291–304.

Andreoni, J. & Croson, R. (2008). Partners versus Strangers: Random Rematching in Public Goods Experiments: Chapter 82. In C. R. Plott & V. L. Smith (Hrsg.), *Handbook of experimental economics results* (S. 776–783). Amsterdam, North-Holland.

Aoyagi, M. & Fréchette, G. (2009). Collusion as public monitoring becomes noisy: Experimental evidence. *Journal of Economic Theory*, *144*(3), 1135–1165.

Arbuthnott, J. (1710). An argument for divine providence, taken from the constant regularity observed in the births of both sexes. *Philosophical Transaction of the Royal Society of London*, *27*, 186–190.

Arellano, M. (1993). On the testing of correlated effects with panel data. *Journal of Econometrics*, *59*(1-2), 87–97.

Arrow, K. J. (1986). Rationality of Self and Others in an Economic System. *The Journal of Business*, *59*(4), S385–S399.

Arrow, K. J. (1976). *Essays in the theory of risk-bearing* (3. Aufl.). Amsterdam, North-Holland Publishing.

Arthur, W. B. (1994). Inductive Reasoning and Bounded Rationality. *The American Economic Review*, *84*(2), 406–411.

Auer, P., Cesa-Bianchi, N., Freund, Y. & Schapire, R. E. (2002). The Nonstochastic Multiarmed Bandit Problem. *SIAM Journal on Computing*, *32*(1), 48–77.

Axelrod, R. (1980). Effective Choice in the Prisoner's Dilemma. *The Journal of Conflict Resolution*, *24*(1), 3–25.

Axelrod, R. (1984). *The Evolution of Cooperation*. New York, NY, Basic Books.

Axelrod, R. (1997). The evolution of strategies in the iterated prisoner's dilemma. In C. Bicchieri, R. Jeffrey & B. Skyrms (Hrsg.), *The dynamics of norms* (S. 1–16). Cambridge, MA, Cambridge University Press.

Banerjee, B. & Peng, J. (2003). Efficient no-regret multiagent learning. In T. Fawcett & N. Mishra (Hrsg.), *Proceedings of the Twentieth International Conference on Machine Learning*. Twentieth International Conference on Machine Learning, Menlo Park, CA, AAAI Press. American Association for Artificial Intelligence.

Bardsley, N., Cubitt, R. & Loomes, G. (2020). *Experimental economics: Rethinking the rules*. Princeton, NJ, Princeton University Press.

Belot, M., Duch, R. & Miller, L. (2015). A comprehensive comparison of students and non-students in classic experimental games. *Journal of Economic Behavior & Organization*, *113*, 26–33.

Betsch, T., Haberstroh, S., Glöckner, A., Haar, T. & Fiedler, K. (2001). The effects of routine strength on adaptation and information search in recurrent decision making. *Organizational behavior and human decision processes*, *84*(1), 23–53.

Betsch, T. & Haberstroh, S. (2001). Financial incentives do not pave the road to good experimentation. *Behavioral and Brain Sciences*, *24*(3), 404.

Betsch, T., Plessner, H., Schwieren, C. & Gütig, R. (2001). I Like It But I Don't Know Why: A Value-Account Approach to Implicit Attitude Formation. *Personality and Social Psychology Bulletin*, *27*(2), 242–253.

Bock, O., Baetge, I. & Nicklisch, A. (2014). hroot: Hamburg Registration and Organization Online Tool. *European Economic Review*, *71*, 117–120.

Bodner, R. & Prelec, D. (2008). Self-Signaling and Diagnostic Utility in Everyday Decision-Making. In I. Brocas & J. D. Carrillo (Hrsg.), *Rationality and Well-Being* (S. 105–124). Oxford, Oxford University Press.

Bohm, P. (2002). Pitfalls in Experimental Economics. In F. Andersson & H. Holm (Hrsg.), *Experimental Economics: Financial Markets, Auctions, and Decision Making* (S. 117–126). Boston, MA, Springer.

Bowling, M. (2005). Convergence and No-Regret in Multiagent Learning. In L. K. Saul & Weiss , Yair Bottou, Léon (Hrsg.), *Advances in Neural Information Processing Systems 17: Proceedings of the 2004 Conference*, Cambridge, MA, MIT Press.

Bowling, M. & Veloso, M. (2001). Rational and Convergent Learning in Stochastic Games, In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*. Seventeenth International Joint Conference on Artificial Intelligence.

Bowling, M. & Veloso, M. (2002). Multiagent learning using a variable learning rate. *Artificial Intelligence*, *136*(2), 215–250.

Breitmoser, Y. (2015). Cooperation, But No Reciprocity: Individual Strategies in the Repeated Prisoner's Dilemma. *American Economic Review*, *105*(9), 2882–2910.

Brown, G. W. (1951). Iterative solution of games by Fictitious Play. In T. C. Koopmans (Hrsg.), *Activity Analysis of Production and Allocation* (S. 374–376). New York, NY, Wiley.

Bruns, B. (2015). Names for Games: Locating 2×2 Games. *Games*, *6*(4), 495–520.

Cabral, L., Ozbay, E. Y. & Schotter, A. (2014). Intrinsic and instrumental reciprocity: An experimental study. *Games and Economic Behavior*, *87*, 100–121.

Cachon, G. P. & Camerer, C. F. (1996). Loss-Avoidance and Forward Induction in Experimental Coordination Games. *The Quarterly Journal of Economics*, *111*(1), 165–194.

Camera, G. & Casari, M. (2009). Cooperation among Strangers under the Shadow of the Future. *American Economic Review*, *99*(3), 979–1005.

Camerer, C. F., Ho, T.-H. & Chong, J.-K. (2004). A Cognitive Hierarchy Model of Games. *The Quarterly Journal of Economics*, *119*(3), 861–898.

Camerer, C. F. (1997). Progress in Behavioral Game Theory. *Journal of Economic Perspectives*, *11*(4), 167–188.

Camerer, C. F. (2004). Behavioral Game Theory: Predicting Human Behavior in Strategic Situations. In C. Camerer, G. F. Loewenstein & M. Rabin (Hrsg.), *Advances in behavioral economics: ca* (S. 374–392). New York, NY, Princeton University Press.

Camerer, C. F. (2011). *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton, NJ, Princeton University Press.

Camerer, C. F., Ho, T.-H. & Chong, J.-K. (2002). Sophisticated Experience-Weighted Attraction Learning and Strategic Teaching in Repeated Games. *Journal of Economic Theory*, *104*(1), 137–188.

Camerer, C. F. & Hogarth, R. M. (1999). The Effects of Financial Incentives in Experiments:A Review and Capital-Labor-Production Framework. *Journal of Risk and Uncertainty*, *19*(1-3), 7–42.

Cameron, A. C. & Trivedi, P. K. (2010). *Microeconometrics using Stata* (überarbeitet). College Station, TX, Stata Press.

Carmel, D. & Markovitch, S. (1996). Learning models of intelligent agents, In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*. Thirteenth International Joint Conference on Artificial Intelligence, AAAI Press.

Carmel, D. & Markovitch, S. (1997). Exploration and Adaptation in Multiagent Systems: A Model-based Approach. *Technical Report CIS9704*.

Carmel, D. & Markovitch, S. (1998). Model-based learning of interaction strategies in multiagent systems. *Journal of Experimental & Theoretical Artificial Intelligence*, *10*(3), 309–332.

Chai, T. & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250.

Chan, C., Landry, S. P. & Troy, C. (2011). Examining External Validity Criticisms in the Choice of Students as Subjects in Accounting Experiment Studies. *The Journal of Theoretical Accounting Research*, *7*(1), 53–78.

Chang, Y.-H., Ho, T. & Kaelbling, L. P. (2004). Mobilized ad-hoc networks: A reinforcement learning approach, In *Proceedings of the International Conference on Autonomic Computing*. International Conference on Autonomic Computing.

Chang, Y.-H. & Kaelbling, L. P. (2002). Playing is believing: The role of beliefs in multi-agent learning. In T. G. Dietterich, S. Becker & Z. Ghahramani. (Hrsg.), *Advances in Neural Information Processing Systems 14*. Neural Information Processing Systems.

Charness, G., Gneezy, U. & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, *81*(1), 1–8.

Chinczewski, J. (2019). *Strategische Verschlechterung in dynamischen Konflikten: Eine empirische Untersuchung im Rahmen der Konfliktanalyse* (Dissertation) [Institut für Unternehmensführung]. Karlsruher Institu für Technologie. Karlsruhe.

Claus, C. & Boutilier, C. (1998). The Dynamics of Reinforcement Learning in Cooperative Multiagent Systems. In J. Mostow, C. Rich & B. Buchanan (Hrsg.), *Fifteenth National Conference on Artificial Intelligence*, Menlo Park, CA, AAAI Press. American Association for Artificial Intelligence.

Cleff, T. (2019). *Angewandte Induktive Statistik und Statistische Testverfahren*. Wiesbaden, Springer Fachmedien Wiesbaden.

Conitzer, V. & Sandholm, T. (2007). AWESOME: A general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents. *Machine Learning*, *67*(1-2), 23–43.

Croson, R. (2005). The Method of Experimental Economics. *International Negotiation*, *10*(1), 131–148.

Croson, R. T. (1996). Partners and strangers revisited. *Economics Letters*, *53*(1), 25–32.

D'Agostino, R. B., Belanger, A. & D'Agostino, R. B., JR. (1990). A Suggestion for Using Powerful and Informative Tests of Normal. *The American Statistician*, *44*(4), 316–321.

Dal Bo, P. & Frechette, G. R. (2013). Strategy Choice in the Infinitely Repeated Prisoners' Dilemma. *SSRN Electronic Journal*.

Dal Bo, P. & Frechette, G. R. (2018). On the Determinants of Cooperation in Infinitely Repeated Games: A Survey. *Journal of Economic Literature*, *56*(1), 60–114.

Dal Bo, P. & Frechette, G. R. (2019). Strategy Choice in the Infinitely Repeated Prisoner's Dilemma. *American Economic Review*, *109*(11), 3929–3952.

Dal Bó, P. (2005). Cooperation under the Shadow of the Future: Experimental Evidence from Infinitely Repeated Games. *American Economic Review*, *95*(5), 1591–1604.

Das, P. (2019). *Econometrics in Theory and Practice: Analysis of Cross Section, Time Series and Panel Data with Stata 15.1*. Singapur, Springer Nature.

Davis, D. D. & Holt, C. A. (1993). *Experimental economics*. Princeton, NJ, Princeton University Press.

Dawes, R. M. (1988). *Rational choice in an uncertain world*. San Diego, CA, Harcourt Brace Jovanovich.

Dingelstedt, A. (2015). *Die Wirkung von Incentives auf die Antwortqualität in Umfragen* (Dissertation) [Quantitative Methoden und Statistik]. Georg-August-Universität Göttingen. Göttingen.

Duffy, J. & Ochs, J. (2009). Cooperative behavior and the frequency of social interaction. *Games and Economic Behavior*, *66*(2), 785–812.

Ellsberg, D. (1965). Risk, Ambiguity, and the Savage Axioms. *The Quarterly Journal of Economics*, *61*(4), 643–669.

Erev, I. & Haruvy, E. (2016). Learning and the Economics of Small Decisions. In J. H. Kagel & A. E. Roth (Hrsg.), *The Handbook of Experimental Economics: Volume Two* (S. 638–716). Princeton, NJ, Princeton University Press.

Erev, I. & Roth, A. E. (1998). Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria. *The American Economic Review*, *88*(4), 848–881.

Erev, I. & Roth, A. E. (2007). Multi-agent learning and the descriptive value of simple models. *Artificial Intelligence*, *171*(7), 423–428.

Feltovich, N. (2000). Reinforcement-based vs. Belief-based Learning Models in Experimental Asymmetric-information Games. *Econometrica*, *68*(3), 605–641.

Feltovich, N., Iwasaki, A. & ODA, S. H. (2012). Payoff levels, loss avoidance, and equilibrium selection in games with multiple equilibria: An experimental study. *Economic Inquiry*, *50*(4), 932–952.

Foster, D. P. & Vohra, R. (1999). Regret in the On-Line Decision Problem. *Games and Economic Behavior*, *29*(1-2), 7–35.

Frechette, G. R. (2015). Laboratory Experiments: Professionals Versus Students. In G. R. Fréchette & A. Schotter (Hrsg.), *Handbook of experimental economic methodology* (S. 360–390). Oxford, Oxford University Press.

Fréchette, G. R. & Yuksel, S. (2017). Infinitely repeated games in the laboratory: four perspectives on discounting and random termination. *Experimental Economics*, *20*(2), 279–308.

Freund, Y. & Schapire, R. E. (1999). Adaptive Game Playing Using Multiplicative Weights. *Games and Economic Behavior*, *29*(1-2), 79–103.

Friedman, D. & Sunder, S. (1994). *Experimental methods: A primer for economists*. Cambridge, MA, Cambridge University Press.

Friedman, M. & Savage, L. J. (1948). The Utility Analysis of Choices Involving Risk. *Journal of Political Economy*, *56*(4), 279–304.

Fudenberg, D. & Levine, D. K. (1995). Consistency and cautious fictitious play. *Journal of Economic Dynamics and Control*, *19*(5-7), 1065–1089.

Fudenberg, D. & Levine, D. K. (1998). *The theory of learning in games* (Bd. 2). Cambridge, MA, MIT Press.

Fudenberg, D. & Maskin, E. (1986). The folk theorem in repeated games with discounting or with incomplete information. *Econometrica*, *54*(3), 533–554.

Fudenberg, D., Rand, D. G. & Dreber, A. (2012). Slow to Anger and Fast to Forgive: Cooperation in an Uncertain World. *American Economic Review*, *102*(2), 720–749.

Graf, A. (2018). *Reinforcement Learning in der Spieltheorie: Entwicklung einer Q-Learning-basierten KI zur optimalen Strategiefindung im iterativen Gefangenendilemma* (Masterarbeit) [Insitut für Unternehmensführung]. Karlsruher Institu für Technologie. Karlsruhe.

Greene, W. H. (2008). *Econometric analysis* (6. Aufl.). Upper Saddle River, NJ, Pearson Prentice Hall.

Greenwald, A. & Hall, K. (2003). Correlated-Q Learning. In T. Fawcett & N. Mishra (Hrsg.), *Proceedings of the Twentieth International Conference on Machine Learning*. Twentieth International Conference on Machine Learning, Menlo Park, CA, AAAI Press. American Association for Artificial Intelligence.

Greenwald, A. & Jafari, A. (2003). A General Class of No-Regret Learning Algorithms and Game-Theoretic Equilibria, In *Learning Theory and Kernel Machines*. 16th Annual Conference on Learning Theory and 7th Kernel Workshop, Berlin, Springer. Washington, DC.

Greenwald, A., Jafari, A., Ercal, G. & Gondek, D. (2001). On No-Regret Learning, Fictitious Play, and Nash Equilibrium. In C. E. Brodley & A. P. Danyluk (Hrsg.), *Proceedings of the Eighteenth International Conference on Machine Learning*. Eighteenth International Conference on Machine Learning, San Francisco, CA, Morgan Kaufmann Publishers Inc.

Greenwald, A. & Littman, M. L. (2007a). Introduction to the special issue on learning and computational game theory. *Machine Learning*, *67*(1-2), 3–6.

Greenwald, A. & Littman, M. L. (2007b). Special Issue on Learning & Computational Game Theory. *Machine Learning*, *67*(1-2).

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association*, *1*(1), 114–125.

Guestrin, C., Koller, D. & Parr, R. (2002). Multiagent Planning with Factored MDPs. In T. G. Dietterich, S. Becker & Z. Ghahramani. (Hrsg.), *Advances in Neural Information Processing Systems 14*. Neural Information Processing Systems.

Häggström, O. (2002). *Finite Markov chains and algorithmic applications* (5. Aufl., Bd. 52). Cambridge, MA, Cambridge University Press.

Harrison, G. W. & List, J. A. (2004). Field Experiments. *Journal of Economic Literature*, *42*(4), 1009–1055.

Hart, S. & Mas-Colell, A. (2000). A Simple Adaptive Procedure Leading to Correlated Equilibrium. *Econometrica*, *68*(5), 1127–1150.

Hausman, J. A. (1978). Specification Tests in Econometrics. *Econometrica*, *46*(6), 1251–1271.

Hertwig, R. & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences*, *24*(3), 383–403.

Hoechle, D. (2007). Robust Standard Errors for Panel Regressions with Cross-Sectional Dependence. *Stata Journal*, *7*(3), 281–312.

Hogarth, R. M., Gibbs, B. J., McKenzie, C. R. & Marquis, M. A. (1991). Learning from feedback: Exactingness and incentives. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(4), 734–752.

Holt, C. A. (1995). Industrial Organization: A Survey of Laboratory Research. In J. H. Kagel & A. E. Roth (Hrsg.), *The handbook of experimental economics* (S. 402–403). Princeton, NJ, Princeton University Press.

Hu, J. & Wellman, M. P. (1998). Multiagent reinforcement learning: theoretical framework and an algorithm. In J. W. Shavlik (Hrsg.), *Proceedings of the Fifteenth International Conference on Machine Learning*. Fifteenth International Joint Conference on Artificial Intelligence, San Francisco, CA, Morgan Kaufmann Publishers.

Hu, J. & Wellman, M. P. (2003). Nash Q-Learning for General-Sum Stochastic Games. *Journal of Machine Learning Research*, *4*, 1039–1069.

Ioannou, C. A. & Romero, J. (2014). A generalized approach to belief learning in repeated games. *Games and Economic Behavior*, *87*, 178–203.

Jenkins, G. D., JR., Mitra, A., Gupta, N. & Shaw, J. D. (1998). Are financial incentives related to performance? A meta-analytic review of empirical research. *Journal of Applied Psychology*, *83*(5), 777–787.

Kabakcha, F. (2017). *Konzeption und Entwicklung eines innovativen Softwaresystems zur experimentellen Untersuchung spieltheoretischer Fragestellungen* (Masterarbeit) [Institut für Unternehmensführung]. Karlsruher Institu für Technologie. Karlsruhe.

Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291.

Kalai, E. & Lehrer, E. (1993). Rational Learning Leads to Nash Equilibrium. *Econometrica*, *61*(5), 1019–1045.

Kapetanakis, S. & Kudenko, D. (2005). Reinforcement Learning of Coordination in Heterogeneous Cooperative Multi-agent Systems. In D. Kudenko, D. Kazakov & E. Alonso (Hrsg.), *Adaptive Agents and Multi-Agent Systems II: Adaptation and Multi-Agent Learning* (S. 119–131). Berlin, Springer.

Keser, C. & van Winden, F. (2000). Conditional Cooperation and Voluntary Contributions to Public Goods. *Scandinavian Journal of Economics*, *102*(1), 23–39.

Klopfer, A. (2018). *Koalitionäre Lösungskonzepte für dynamische Konfliktsituationen - eine empirische Untersuchung* (Dissertation) [Institut für Unternehmensführung]. Karlsruher Institu für Technologie. Karlsruhe.

Kohler, U. & Kreuter, F. (2017). *Datenanalyse mit Stata: Allgemeine Konzepte der Datenanalyse und ihre praktische Anwendung* (5. Aufl.). Berlin, De Gruyter Oldenbourg.

LaPlace, P. S. (1812). *Théorie analytique des probabilités* (Bde. 2). Paris, Courcier.

Leyton-Brown, K. & Shoham, Y. (2008). *Essentials of Game Theory: A Concise Multidisciplinary Introduction* (Bd. 3). Morgan & Claypool Publishers.

Leyton-Brown, K. & Tennenholtz, M. (2003). Local-Effect Games, In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*. Eighteenth International Joint Conference on Artificial Intelligence. Acapulco, Mexico.

Lindstädt, H. (2006). *Beschränkte Rationalität: Entscheidungsverhalten und Organisationsgestaltung bei beschränkter Informationsverarbeitungskapazität* (Bd. 7). München, Hampp.

Littlestone, N. & Warmuth, M. K. (1994). The Weighted Majority Algorithm. *Information and Computation*, *108*(2), 212–261.

Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. In W. W. Cohen & H. Hirsh (Hrsg.), *Proceedings of the Eleventh International Conference on Machine Learning*. Eleventh International Conference on Machine Learning, San Francisco, CA, Morgan Kaufmann Publishers Inc.

Littman, M. L. (1996). A Generalized Reinforcement-Learning Model: Convergence and Applications. In L. Saitta (Hrsg.), *Proceedings of the 13th International Conference on Machine Learning*. 13th International Conference on Machine Learning, San Francisco, CA, Morgan Kaufmann Publishers.

Littman, M. L. (2001). Friend-or-Foe Q-learning in General-Sum Games. In C. E. Brodley & A. P. Danyluk (Hrsg.), *Proceedings of the Eighteenth International Conference on Machine Learning*. Eighteenth International Conference on Machine Learning, San Francisco, CA, Morgan Kaufmann Publishers Inc.

Littman, M. L. & Stone, P. (2002). Implicit Negotiation in Repeated Games (J.-J. C. Meyer & M. Tambe, Hrsg.). In J.-J. C. Meyer & M. Tambe (Hrsg.), *Intelligent Agents VIII: Agent Theories, Architectures, and Languages*. 8th International Workshop on Agent Theories, Architectures, and Languages, Berlin, Springer. Seattle,WA.

Loewenstein, G. (1999). Experimental Economics from the Vantage-Point of Behavioural Economics. *The Economic Journal*, *109*(453), F25–F34.

Mann, H. B. & Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, *18*(1), 50–60.

Mannor, S. & Shimkin, N. (2003). The Empirical Bayes Envelope and Regret Minimization in Competitive Markov Decision Processes. *Mathematics of Operations Research*, *28*(2), 327–345.

McKelvey, R. D. & Palfrey, T. R. (1992). An Experimental Study of the Centipede Game. *Econometrica*, *60*(4), 803.

Milgrom, P. & Roberts, J. (1991). Adaptive and sophisticated learning in normal form games. *Games and Economic Behavior*, *3*(1), 82–100.

Mitchell, T. M. (1997). *Machine Learning*. New York, NY, McGraw-Hill.

Miyasawa, K. (1961). *On the Convergence of Learning Processes in a 2x2 Non-Zero-Person Game* (Bd. 33). Princeton, NJ.

Müller, F. (2018). *Predicting and classifying individual behavior in repeated games with Markov strategies* (Dissertation) [Institut für Unternehmensführung]. Karlsruher Institu für Technologie. Karlsruhe.

Murnighan, J. K. & Roth, A. E. (1983). Expecting Continued Play in Prisoner's Dilemma Games. *Journal of Conflict Resolution*, *27*(2), 279–300.

Nachbar, J. H. (1990). "Evolutionary" selection dynamics in games: Convergence and limit properties. *International Journal of Game Theory*, *19*(1), 59–89.

Nash, J. F. (1950). Equilibrium Points in n-Person Games. *Proceedings of the National Academy of Sciences of the United States of America*, *36*(1), 48–49.

Neumann, J. V. & Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton, NJ, Princeton University Press.

Normann, H.-T. & Wallace, B. (2012). The impact of the termination rule on cooperation in a prisoner's dilemma experiment. *International Journal of Game Theory*, *41*(3), 707–718.

Nudelman, E., Wortman, J., Shoham, Y. & Leyton-Brown, K. (2004). Run the GAMUT: A Comprehensive Approach to Evaluating Game-Theoretic Algorithms. Third International Joint Conference on Autonomous Agents and Multiagent Systems. New York, NY.

Ortega, P. A. & Legg, S. (2018). Modeling Friends and Foes. *Computing Research Repository*.

Palfrey, T. R. & Prisbrey, J. E. (1996). Altuism, reputation and noise in linear public goods experiments. *Journal of Public Economics*, *61*(3), 409–427.

Papadimitriou, C. H. & Tsitsiklis, J. N. (1987). The Complexity of Markov Decision Processes. *Mathematics of Operations Research*, *12*(3), 441–450.

Powers, R. & Shoham, Y. (2005a). Learning against opponents with bounded memory, In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*. Nineteenth International Joint Conference on Artificial Intelligence.

Powers, R. & Shoham, Y. (2005b). New criteria and a new algorithm for learning in multi-agent systems. In L. K. Saul & Weiss , Yair Bottou, Léon (Hrsg.), *Advances in Neural Information Processing Systems 17: Proceedings of the 2004 Conference*, Cambridge, MA, MIT Press.

Press, W. H. & Dyson, F. J. (2012). Iterated Prisoner's Dilemma contains strategies that dominate any evolutionary opponent. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(26), 10409–10413.

Read, D. (2005). Monetary incentives, what are they good for? *Journal of Economic Methodology*, *12*(2), 265–276.

Rezek, I., Leslie, D. S., Reece, S., Roberts, S. J., Rogers, A., Dash, R. K. & Jennings, N. R. (2008). On Similarities between Inference in Game Theory and Machine Learning. *Journal of Artificial Intelligence Research*, *33*, 259–283.

Richter, M. K. (1966). Revealed Preference Theory. *Econometrica*, *34*(3), 635–641.

Robinson, D. & Goforth, D. (2006). *The topology of the 2x2 games: A new periodic table*. London, Routledge.

Robinson, J. (1951). An Iterative Method of Solving a Game. *Annals of Mathematics*, *54*(2), 296–301.

Romero, J. & Rosokha, Y. (2018). Constructing strategies in the indefinitely repeated prisoner's dilemma game. *European Economic Review*, *104*, 185–219.

Romero, J. & Rosokha, Y. (2019). Mixed Strategies in the Indefinitely Repeated Prisoner's Dilemma. *SSRN Electronic Journal*.

Roth, A. E. (1996). Individual Rationality as a Useful Approximation: Comments on Tversky's Rational Theory and Constructive Choice. In K. J. Arrow, E. Colombatto, M. Perlman & C. Schmidt (Hrsg.), *The Rational Foundations of Economic Behavior* (S. 198–202). London, Macmillan.

Roth, A. E. & Murnighan, J. (1978). Equilibrium behavior and repeated play of the prisoner's dilemma. *Journal of Mathematical Psychology*, *17*(2), 189–198.

Royston, P. (1991). sg3.5: Comment on sg3.4 and an improved D'Agostino test. In Stata Corporation (Hrsg.), *Stata Technical Bulletin* (S. 23–24).

Royston, P. (1992). Approximating the Shapiro-Wilk W-test for non-normality. *Statistics and Computing*, *2*(3), 117–119.

Rubinstein, A. (1979). Equilibrium in supergames with the overtaking criterion. *Journal of Economic Theory*, *21*(1), 1–9.

Sabater-Grande, G. & Georgantzis, N. (2002). Accounting for risk aversion in repeated prisoners' dilemma games: an experimental test. *Journal of Economic Behavior & Organization*, *48*(1), 37–50.

Sandholm, T. (2007). Perspectives on multiagent learning. *Artificial Intelligence*, *171*(7), 382–391.

Sandholm, T. W. & Crites, R. H. (1996). Multiagent reinforcement learning in the Iterated Prisoner's Dilemma. *Biosystems*, *37*(1-2), 147–166.

Selten, R. (1965). Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit - Teil I: Bestimmung des dynamischen Preisgleichgewichts. *Zeitschrift für die gesamte Staatswissenschaft*, *121*(2), 301–324.

Selten, R. (1998). Axiomatic Characterization of the Quadratic Scoring Rule. *Experimental Economics*, *1*(1), 43–61.

Selten, R. & Stoecker, R. (1986). End behavior in sequences of finite Prisoner's Dilemma supergames A learning theory approach. *Journal of Economic Behavior & Organization*, *7*(1), 47–70.

Shapiro, S. S. & Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, *52*(3), 591–611.

Sherstyuk, K., Tarui, N. & Saijo, T. (2013). Payment schemes in infinite-horizon experimental games. *Experimental Economics*, *16*(1), 125–153.

Shoham, Y. (2008). Computer science and game theory. *Communications of the ACM*, *51*(8), 74–79.

Shoham, Y. & Leyton-Brown, K. (2008). *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge, MA, Cambridge University Press.

Shoham, Y. & Powers, R. (2014a). Multi-agent Learning I: Problem Definition. In C. Sammut & G. I. Webb (Hrsg.), *Encyclopedia of Machine Learning and Data Mining* (S. 1–4). Boston, MA, Springer US.

Shoham, Y. & Powers, R. (2014b). Multi-agent Learning II: Algorithms. In C. Sammut & G. I. Webb (Hrsg.), *Encyclopedia of Machine Learning and Data Mining* (S. 1–5). Boston, MA, Springer US.

Shoham, Y., Powers, R. & Grenager, T. (2007). If multi-agent learning is the answer, what is the question? *Artificial Intelligence*, *171*(7), 365–377.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, *69*(1), 99–118.

Simon, H. A. (1990). Bounded Rationality. In J. Eatwell, M. Milgate & P. Newman (Hrsg.), *Utility and Probability* (S. 15–18). London, Palgrave Macmillan.

Singh, S. P., Kearns, M. J. & Mansour, Y. (2000). Nash Convergence of Gradient Dynamics in General-Sum Games. In B. Craig & M. Goldszmidt (Hrsg.), *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*. 16th Conference on Uncertainty in Artificial Intelligence, San Francisco, CA, Morgan Kaufmann Publishers Inc.

Smith, V. L. (1976). Experimental Economics: Induced Value Theory. *The American Economic Review*, *66*(2), 274–279.

Snedecor, G. W. & Cochran, W. G. (1991). *Statistical methods* (8. Aufl., Bd. 276). Wiley.

Stewart, W. J. (2009). *Probability, Markov Chains, Queues, and Simulation: The Mathematical Basis of Performance Modeling*. Princeton, NJ, Princeton University Press.

Stone, D. N. & Ziebart, D. A. (1995). A Model of Financial Incentive Effects in Decision Making. *Organizational behavior and human decision processes*, *61*(3), 250–261.

Student. (1908). The Probable Error of a Mean. *Biometrika*, *6*(1), 1.

Suryadi, D. & Gmytrasiewicz, P. J. (1999). Learning Models of Other Agents Using Influence Diagrams. In J. Kay (Hrsg.), *UM99 User Modeling: Proceedings of the Seventh International Conference*. Seventh International Conferences on User Modeling, Vienna, Springer Vienna.

Sutton, R. S. & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA, MIT Press.

Tennenholtz, M. (2002). Game Theory and Artificial Intelligence. In M. d'Inverno, M. Luck, M. Fisher & C. Preist (Hrsg.), *Foundations and Applications of Multi-Agent Systems: UKMAS Workshops 1996-2000 Selected Papers* (S. 49–58). Berlin, Heidelberg, Springer.

Toutenburg, H. & Heumann, C. (2008). *Induktive Statistik*. Berlin, Springer.

Tversky, A. (1996). Rational theory and constructive choice. In K. J. Arrow, E. Colombatto, M. Perlman & C. Schmidt (Hrsg.), *The Rational Foundations of Economic Behavior* (S. 185–197). London, Macmillan.

Vespa, E. (2011). Cooperation in Dynamic Games: An Experimental Investigation. *SSRN Electronic Journal*.

Vince Knight, Owen Campbell, Marc, T.J. Gaffney, Eric Shaw, VSN Reddy Janga, James Campbell, Karol M. Langner, Nikoleta Glynatsi, Sourav Singh, Julie Rymer, Thomas Campbell, Jason Young, MHakem, Geraint Palmer, Kristian Glass, edouardArgenson, Daniel Mancia, Martin Jones, … Adam Pohl. (2020). *Axelrod-Python/Axelrod: v4.9.1*. Github.

Vohra, R. & Wellman, M. (2007). Foundations of Multi-Agent Learning. *Artificial Intelligence*, *171*(7), 363–452.

Vohra, R. V. & Wellman, M. P. (2007). Foundations of multi-agent learning: Introduction to the special issue. *Artificial Intelligence*, *171*(7), 363–364.

Vu, T., Powers, R. & Shoham, Y. (2006). Learning against multiple opponents. In H. Nakashima, M. Wellman, G. Weiss & P. Stone (Hrsg.), *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*. Fifth international joint conference on Autonomous agents and multiagent systems, New York, NY, Association for Computing Machinery.

Wang, X. & Sandholm, T. (2002). Reinforcement learning to play an optimal Nash equilibrium in team Markov games (S. Becker, T. Sebastian & K. Obermayer, Hrsg.). In S. Becker, T. Sebastian & K. Obermayer (Hrsg.), *Proceedings of the 15th International Conference on Neural Information Processing Systems*. 15th International Conference on Neural Information Processing Systems, Cambridge, MA, MIT Press.

Wang, X. & Sandholm, T. (2003). Reinforcement learning to play an optimal Nash equilibrium. In S. Becker, S. Thrun & K. Obermayer (Hrsg.), *Advances in neural information processing systems 15*. Neural Information Processing Systems, Cambridge, MA, MIT Press.

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, *1*(6), 80–83.

Wilcoxon, F. (1947). Probability Tables for Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, *3*(3), 119–122.

Willer, D. & Walker, H. A. (2007). *Building experiments: Testing social theory*. Stanford, CA, Stanford University Press.

Willmott, C. J., Matsuura, K. & Robeson, S. M. (2009). Ambiguities inherent in sums-of-squares-based error statistics. *Atmospheric Environment*, *43*(3), 749–752.