How does one coordinate with others?

Using computational cognitive models to investigate the emergence of social norms.

Hendrik Nunner

Department of Experimental Psychology
Utrecht University

This thesis is submitted for the degree of
Master of Science

August 2017

Abstract

I investigate whether reinforcement learning models can explain the emergence of social norms in the Volunteer's Dilemma. A three player version of this coordination game showed that reward structures of the game have an immediate effect on the manifestation of behavioral patterns, precursors of social norms. It is, however, unknown whether and how cognitive mechanisms contribute to the emergence of social norms. I therefore describe how behavioral patterns in the Volunteer's Dilemma can be explained with simple cognitive mechanisms of individuals. Using two classes of computational cognitive models, based on reinforcement learning, I show that simple state-based learning does not suffice for pattern emergence. Reinforcement of successful future-oriented strategies, however, predicts the same behavioral patterns as found in the empirical data. Further, I show that certain characteristics of learning either support (e.g., realistic propensities), or suppress the emergence of patterns (e.g., altruism).

Table of contents

List of figures

List of tables

1. Introduction

Social norms shape human behavior in many situations of social interaction: Rules of etiquette we obey when we dine together, going to the back of the line when we queue up at the grocery store, and the ways we talk to family, friends, or superiors. All these examples are guided by behavioral rules enforced by our fellow human beings. And although social norms have been a long-term subject of sociological scholarship, it is often a big puzzle how they are formed and how they change.

Consider an example of the Volunteer's Dilemma (*VOD*) (Diekmann, 1985, 1993). Three friends, Jane, John, and Jean, share a flat. They also have a dog, Spot, that needs to be walked once a day. There are many solutions to prevent Spot from relieving himself in the kitchen. They could all go together or maybe John volunteers every day. Suppose all three friends are busy finishing an important scientific paper. Time becomes a precious resource and cooperation comes with a cost. A fair solution would be if only one friend walks Spot at a time. In order to coordinate, the friends could negotiate and draw up a timetable. However, coordination might also emerge tacitly (i.e. without communication) through repeated interactions in the same recurring situations: Maybe John takes over the task today, because Jane walked Spot yesterday, and Jean volunteered two days ago. Thus, a constant pattern of turn-taking emerged over time. In consequence, Jane and Jean expect John to walk Spot, because the last time John volunteered dates three days back. Recurring behavioral patterns therefore foster expectations. Wrong (1994) calls these expectations towards the behavior of others that result from repeated interactions *latent norms* (p.48), precursors of social norms. The stronger these expectations, the stronger the resentment in cases of non-compliance, and the more likely a behavior becomes a norm (Opp, 2004, p. 14).

In a controlled experiment, Diekmann and Przepiorka (2016) asked human participants to play computer based versions of the three player repeated VOD. Immediate rewards were given in the form of scores, and monetary rewards eventually. Over the course of time behavioral patterns emerged in the majority of games. Moreover, the study showed that the structural conditions (i.e. the reward structure) of the VOD had an immediate effect on the emerging patterns. That is, turn-taking when cooperation costs were the same for all

participants, and solitary volunteering when one participant had lower cooperation costs than the others. This research provides an early building block for a bottom-up theory on norm emergence.

However, what is unknown yet is how individual players make decisions in order to coordinate. More precisely: What are the cognitive mechanisms needed within each player to coordinate with others in the VOD? This is necessary to understand how people behave strategically in situations of social interaction. Especially a formalization of norm emergence is indispensable to predict the effects of social norms on group behavior. This study will advance a bottom-up theory on the emergence of social norms, thus adding a little piece to the puzzle of how norms emerge. Although some research has been conducted on cognitive mechanisms in strategic game playing (e.g., Camerer, 2003; Colman, 2003; Helbing et al., 2005; Juvina et al., 2015; Stevens et al., 2016), it usually neglects the connection with social norms. Further, it is mostly restricted to two player games. Thus, my study is a primer for further research on cognitive mechanisms: firstly, in context of social norms and, secondly, in groups with more than two participants.

The scenario of a VOD contains two relevant characteristics: repeated game play and immediate feedback in the form of scores. This combination allows each player to reinforce successful and demote unsuccessful actions. In other words, players can learn over time which actions give the highest reward in different situations. Using computational cognitive models based on reinforcement learning, I investigate whether simple models of learning can explain the behavioral patterns in the VOD. This approach follows a prominent idea of cognitive science that assumes simple underlying mechanisms as the reason for complex phenomena (e.g., Anderson, 2002; Pfeifer and Scheier, 2001; Simon, 1969).

Before I introduce the model in chapter 5, I first discuss literature on norm emergence in chapter 2, focusing on why the volunteers dilemma is a suitable paradigm for studying it, and what the human data from the experiment by Diekmann and Przepiorka (2016) revealed. This is followed by an overview of learning as a key cognitive mechanism in repeated interactions and relevant methodological concepts of reinforcement learning in chapter 3; and potential values of my research for the field of artificial intelligence in chapter 4. In chapter 6 I describe the most important results. And finally, in chapter 7, I put the results back into a broader context, discuss implications and limitations of my work, and outline possible paths for future research.

2. Norm emergence in the Volunteer's Dilemma

2.1 The Volunteer's Dilemma as a driving force for norm emergence

The Volunteer's Dilemma (VOD) describes a social scenario in which cooperation of a single agent is necessary and sufficient for the benefit of the whole group (Diekmann, 1985, 1993). A single agent, for example, provides a public good that can be used by the other group members (Allison and Kerr, 1994). However, in the VOD, cooperation also comes with a cost; and in case nobody cooperates, nobody benefits.

Let's now tie these general characteristics of the VOD in with our example from the introduction about who needs to walk the dog in a shared flat. The public good is taking Spot for a walk. Everybody is very busy and volunteering requires time.[1] Each friend might decide individually not to volunteer. This however, results in a bad mutual outcome: a messy kitchen. In contrast, they might come up with a better solution collectively. Together they make a plan that increases welfare for the group as a whole. This discrepancy in reasoning between individual and collective welfare constitutes a *social dilemma* (Rapoport, 1974).

More precisely, the VOD constitutes a *coordination problem* (Lewis, 1969). All group members have the same interest: to maintain a clean kitchen. However, nobody prefers to invest the costs for cooperation, as they are busy working on their papers. To resolve the dilemma between individual preference (working on the paper) and, as a result, negative effects for the group as a whole (messy kitchen) requires the friends to coordinate.

Coordination can, for example, be achieved through behavioral conventions that emerge over time (e.g., turn-taking or solitary volunteering). As already mentioned, these conventions result from patterns in repeated interactions and lead to expectations towards collective

[1] Of course the friends enjoy walking Spot. Thus, cooperation comes also with a benefit. That is why walking the dog is still preferred over starving the dog.

behavior – *latent norms* (Wrong, 1994, p. 48). Non-compliance leads to resentment. And the stronger the resentment, the more likely a latent norm becomes a *social norm* (Opp, 2004, p. 14). Based on a broad range of literature, Diekmann and Przepiorka (2016, p. 1310) define social norm: "A rule guiding social behavior, the deviation from (adherence to) which is negatively (positively) sanctioned."

Voss (2001, p. 110) argues that a reason to develop social norms is "to improve the aggregate welfare of the norm beneficiaries". For example, taking turns to walk Spot keeps the kitchen clean (and brings joy to the volunteer), while everyone minimizes cooperation costs. As a result, the inherent need for coordination in the VOD generates a need for social norms. For that reason, the VOD constitutes an ideal model to investigate the conditions of norm emergence.

One-shot dilemmas, however, do not suffice to investigate the dynamics of norm emergence. The VOD itself defines only the stage game. It merely defines the ways actors can interact and the outcomes depending on collective behavior. In order for behavioral patterns to emerge and actors to form expectations about each other's actions, they need to encounter the same situation repeatedly. As a result, only repeated interactions in the VOD provide sufficient conditions for norms to emerge (e.g., Opp, 2004; Thibaut and Kelley, 1978b; Wrong, 1994).

Another important characteristic of the VOD is that no salient solution exists (Diekmann, 1985). There are many ways to maintain a clean kitchen. In fact, there are several solutions that guarantee minimal costs. That is, only one friend needs to go at a time. Technically speaking, the VOD offers multiple (Nash) equilibria. That is, a combination of strategies in which a single player cannot increase her own benefit by changing her strategy, while all other players stick to their strategies (Osborne and Rubinstein, 1994). Consider, John volunteers to walk Spot, while Jane and Jean keep working on their papers. Nobody can increase benefit by being the only person to change the strategy: If John decides to also work on his paper, Spot creates a mess in the kitchen. If Jane or Jean walk Spot together with John, they have less time to finish their work. But the question "Who is the one that cooperates?" is unsolved. There is no guideline for a single player that guarantees an outcome with minimal costs on the group-level. As a result, top-down analysis, the typical approach of classical game theory, cannot provide a satisfying answer.

The above example describes a *symmetric VOD*. That is, all players have the same costs to cooperate. Consider, Jean is done early with her paper. Lack of time is not a big issue for her anymore. Cooperation still comes with a cost, however, not as high as for her two friends. Thus, she might want to support her friends and decides to volunteer solitarily from now on (see Fig. 2.1). A VOD with different cooperation costs for at least one player is classified as *asymmetric VOD*.

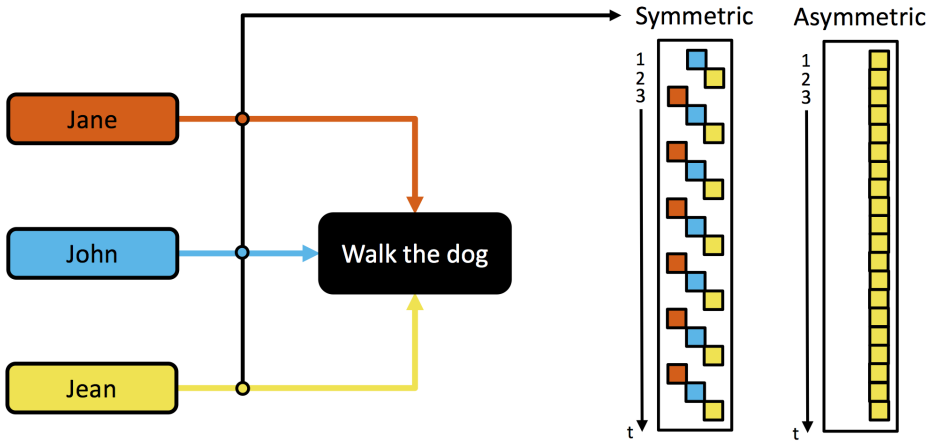

Fig. 2.1 **Example of a Volunteer's Dilemma.** Consider three friends sharing a house and a dog. The dog needs to be walked every day. Walking the dog requires time (and brings joy to the volunteer). If all three are equally busy (symmetric), they take turns. If one friend is less busy (asymmetric), she volunteers solitarily.

2.2 Human data

A recent experiment on the three-player repeated VOD provided a first building block for a bottom-up theory on norm emergence (Diekmann and Przepiorka, 2016). The goal of the experiment was to investigate the structural conditions of latent norm emergence in the VOD. More precisely, to determine the effects of three conditions with different payoff structures on the emergence of behavioral patterns.

Prior to the experiment participants were randomly assigned to a group of three players. Groups then played repeated versions of the VOD. In each round participants were asked to select between cooperation and defection. This was done using a computer program and with no means of communication. Participants clicked on either a button for cooperation or a button for defection on the computer screen. A round was finished, after all players made a choice. At the end of each round, participants saw what choices they and their co-players made. In addition, they received a score to formalize the outcome.

Score calculation depended on the payoff condition. There were four different conditions, of which three are relevant for this study: *Symmetric*, *Asymmetric 1* and *Asymmetric 2*. Scores were composed of two components: base utility (U) and cooperation costs (K). Base utility was the same for every player in all conditions ($U_{1,2,3} = 80$). In case no player cooperated in a round, none of the players received a reward. In case a player cooperated, the defecting players received the full 80 points, while the cooperating player had to pay the cooperation

costs ($U - K$). The amount of costs was defined by the payoff condition. In the symmetric condition, all players had the same cooperation costs ($K_{1,2,3} = 50$). In the asymmetric conditions, one of the players had lower cooperation costs than the other two (Asymmetric 1: $K_1 = 30$, $K_{2,3} = 50$; Asymmetric 2: $K_1 = 10$, $K_{2,3} = 50$). The player with lower cooperation costs is also referred to as the *strong player*. Table 2.1 shows the payoff structure for the different conditions. Participants had full knowledge about the payoff structures.

	Symmetric	**Asymmetric 1**		**Asymmetric 2**	
	$P_{1,2,3}$	P_1	$P_{2,3}$	P_1	$P_{2,3}$
Cooperate	30	50	30	70	30
Defect, while ≥ 1 co-players cooperate	80	80	80	80	80
Defect, while 0 co-players cooperate	0	0	0	0	0

Table 2.1 **Payoff structures of the tested Volunteer's Dilemmas.** P_x denotes players 1, 2, and 3. P_1 in the asymmetric conditions denotes the player with the lowest cooperation costs.

The results showed that different patterns emerged for the different conditions. The most prominent pattern in the symmetric condition was turn-taking between all three participants (49.5% of all rounds). Solitary volunteering by the participant with the lowest cost was dominant in the asymmetric conditions. Moreover, the degree of asymmetry had an immediate effect on the ratio of solitary volunteering (Asymmetric 1: 34.9% of all rounds; Asymmetric 2: 61.7% of all rounds). This result was not anticipated, as the degree of asymmetry does not make a difference from a theoretical standpoint. That is, continuous volunteering by the actor with the lowest cost guarantees lowest collective costs.

There are multiple explanations possible for the unanticipated result that the degree of asymmetry affects the ratio of solitary volunteering. Diekmann and Przepiorka (2016) argue that the effect that asymmetry has on the ratio of solitary volunteering may be a result of social preferences. For example, actors who have a high aversion towards inequity might not be content with earning less than others. Further, willingness to volunteer solitarily might be higher, when costs are low. As a result, low cost cooperation causes only minor differences between the actors.

Another noteworthy argument on willingness to volunteer solitarily stems from critical mass theory Oliver et al. (1985): Strong players might be willing to pay initial costs in order

to produce a public good. This willingness, however, might fade over time. Consequently, players might want to switch to turn-taking on the long run.

I will formalize the cognitive processes underlying acting in the VOD using cognitive models, to see whether these results can also be explained based on human learning. In the following chapter I will outline the theoretical foundations for my work, before I describe the specific models in detail in chapter 5.

3. A cognitive psychological perspective on social norms

The work by Diekmann and Przepiorka (2016) provided a first building block for a bottom-up theory on norm emergence. They showed that structural conditions of social interactions affect the formation of latent norms. However, the decision-making processes of individual players in order to coordinate are yet unknown. Therefore, the question driving my current research is: What are the cognitive mechanisms needed within each player to coordinate with others in the VOD? And consequently, what are appropriate ways to investigate those?

3.1 Learning as key cognitive mechanism

To an outside observer, behavioral patterns such as turn-taking in the Volunteer's Dilemma might come across as the result of complicated cognitive mechanisms, such as high-order reasoning. However, Simon (1969) argues that the actual mechanisms within an agent might be considerably simpler than the emergent behavior might suggest (see also Pfeifer and Scheier, 2001, p. 81). In fact, emerging complex behavior might be just a result of simple, more generic mechanisms within an agent acting in a complex environment. Consequently, there are two crucial elements for the emergence of complex behavior: an environment and the agent acting within it.

Simon (1969, p. 51 ff.) exemplifies this with a simple example. Consider an ant on a beach. It roams around, halting here and there resulting in a path that appears to be arbitrary. However, the ant has a simple goal. That is, to get back to its anthill. The obstacles along the way and the fact that the ant cannot foresee all events it encounters require the ant to adapt to the environment constantly. Simon (1969) argues that a designer could create an artificial ant, based on simple rules, that shows the same complex behavior as the real ant on the beach. Thus, he argues, "the complexity of its behavior over time is largely a reflection of the complexity of the environment in which it finds itself" (p. 52).

In the above example the beach is the environment and the ant is the agent acting within it. I argue that latent norms in the VOD constitute complex behavior that is also explainable with simple underlying mechanisms: The game represents the environment (available actions, payoff structures, repeated interactions, co-players, etc.) and a player is the agent acting within it.

The idea of simple mechanisms giving rise to complex phenomena has also been discussed in cognitive science. Newell (1990) introduced the concept of different *bands of cognition*. He differentiated between four bands on four different time scales: the *biological* (milliseconds), *cognitive* (milliseconds to seconds), *rational* (minutes to hours) and *social* bands (days to months). Anderson (2002) picked up this idea and formulated the *Decomposition Thesis* (p. 86). He suggested that bigger scale phenomena on the social band may be grounded in smaller scale events on the cognitive band. Especially that the long-term effects of learning can be broken down into smaller bits of cognition.

With the idea that learning on the cognitive band may inform about phenomena on the social band, Anderson (2002) provides a sensible starting point for my study. I would in particular like to stress a main feature of the environment: repetition. It has been argued that repeated interactions are indispensable for latent norms to emerge (Opp, 2004; Wrong, 1994). The empirical results further show that over time humans are able to adapt to the scenario and produce coordinated behavior tacitly. Thus, experiences made in the previous rounds is one of the crucial factors in this scenario. Further, Juvina et al. (2015) and Helbing et al. (2005) showed that learning is key for successful strategic interaction in different game theoretic scenarios (i.e. *Chicken* and the *Prisoner's Dilemma*). Thus, I hypothesize that learning from experience is the key cognitive mechanism that gives rise to latent norms in the VOD.

So what is the form of learning in the VOD? Note that the environment (the game) provides immediate feedback after each round. This happens in the form of points received for each action. I therefore hypothesize that a type of learning occurs within each single agent which evaluates the success of actions based on the immediate rewards received from the environment. This feedback is then used to promote or, in other words, to reinforce actions that maximize reward positively and reinforce actions that minimize reward negatively.

3.1.1 Reinforcement Learning

Reinforcement learning is a formal theory that describes how an agent learns what to do in order to maximize a reward (Sutton and Barto, 1998). A reinforcement model must meet three requirements. First, an agent must be able to sense its environment. More precisely, it must recognize the state of the environment the agent is in. Second, the agent must have a

goal relating to the state of the environment. Third, the agent must have a set of actions that allow her to change her state within the environment.

Consider the ant on the beach: The ant's *goal* is to get back to the anthill. However, it *recognizes* a big rock in an upright standing position on its way. Three *actions* are available: walk left, walk right, or walk over it. In order to find the solution with the highest reward (e.g., the shortest path) the ant must experience the same situation repeatedly and try all available actions. This trial and error approach allows an agent to experience all the states of the environment, the *problem state*, and find the solution with the highest reward.

However, things might change in a dynamic environment. Over time the ant learns that walking left around the rock is the shortest path to the anthill and might exploit this knowledge. After a while the rock falls over to the left. Now, walking around it on the right is much shorter. Thus, the ant must also explore alternative actions from time to time to realize changes within its environment. This balancing between *exploration* and *exploitation* is crucial for the success of reinforcement learning, as it prevents that agents get stuck in local or temporal optima.

Reinforcement learning is usually modeled in the form of a Markov decision process (*MDP*). And since this is not the appropriate place to discuss all details, I limit myself to the aspects important for my current work. For a detailed description, please refer to Sutton and Barto (1998, ch. 3.6).

MDPs are mathematical representations of decision-making when outcomes are not fully controlled by the decision maker. An MDP is a 5-tuple, consisting of:

1. a set of states, $\mathbf{S}$
2. a set of actions available to the agent, $\mathbf{A}$
3. a function that describes the probability to reach a certain state $\mathbf{s_{t+1}} \in \mathbf{S}$, given the current state $\mathbf{s_t} \in \mathbf{S}$ and an action $\mathbf{a_t} \in \mathbf{A}$
4. a function that provides a reward ($\mathbf{r_{t+1}}$) after the transition from $\mathbf{s_t}$ to $\mathbf{s_{t+1}}$;
5. rules what an agent observes

As a result, a reinforcement learning process happens in discrete time steps.

One approach of reinforcement learning that proved particularly successful is *Q-Learning* (Watkins and Dayan, 1992). Q-Learning is a model-free approach of reinforcement learning. That means an agent does not require a full representation of its environment. It merely needs to know the state it is in and the actions available in that state. By exploring the whole problem state an agent learns the best actions in any given state. An interesting property

of Q-learning models is that they are mathematically guaranteed to achieve optimal results given enough time to explore the entire problem state (Sutton and Barto, 1998, ch. 6.5)

Action selection in Q-Learning is based on a *Q-Table*. Table 3.1 shows an example of propensities for action-state pairs in the *ant on the beach* example. Intersections describe the propensities (*Q-Values*) of an agent to perform a certain action (columns) in a given state (rows). In the case of exploitation, the agent simply takes the action with the highest value. In the case of exploration, the agent takes any of the remaining actions.

	Go Left	Go Right	Go Over or Through
Stone	50	20	5
Stream of water	15	65	0
Hole	30	20	45

Table 3.1 **Exemplary Q-Table for the *ant on the beach* example.** Columns contain actions. Rows contain states. Intersections describe propensities. In case of exploitation: the agent takes the highest rated action. In case of exploration: the agent takes any other action.

After receiving a reward, an agent needs to update its Q-Table using Formula 3.1 (Sutton and Barto, 1998). $Q(s_t;a_t)$ denotes the propensity, or Q-value, for action a_t in state s_t. α is the learning rate. The higher the value, the more important recently earned rewards ($0 < \alpha \leq 1$). r_{t+1} is the reward an agent actually receives. γ describes a discount factor. The higher the value, the more important the newly learned over old rewards ($0 \leq \gamma \leq 1$). As there are no proposed standard values in the literature for α and γ, they require either educated guesses considering the modeled scenario, or some form of parameter fitting process. $\max_a Q(s_{t+1};a)$ denotes the maximum reward an agent can expect to get after (s_{t+1}) choosing action a.

$$Q(s_t;a_t) \leftarrow Q(s_t;a_t) + \alpha[r_{t+1} + \gamma \max_a Q(s_{t+1};a) - Q(s_t;a_t)] \tag{3.1}$$

I suggest, that the VOD scenario, as tested by Diekmann and Przepiorka (2016), conforms to an MDP. Thus, Q-Learning provides an ideal framework for a model-based research. To illustrate this, consider the following example: A player observes that she had not cooperated for the last two rounds. Thus, her state s is defined by the sequence of actions in the two preceding rounds. She decides to cooperate ($C \in A$), because in the past this was the action that gave the highest rewards (highest Q-Value). After all players have chosen an action the player changes from a state where she defected twice ($s = DD$) into a state in which she defected once followed by cooperation ($s' = DC$). Finally, the VOD notifies the player in form of the reward for cooperation, before the process starts all over again.

In support of a model-based approach, (Gigerenzer, 2007) argues that many aspects in decision-making are of intuitive nature. That is why many traditional methods, like self-reporting, are not of use. Thus, formal models based on existing theory and applied to experimental data can be helpful to get a better understanding of the underlying cognitive processes. Further, reinforcement learning models proved to be successful in cognitive (neuro-)science (Dayan and Daw, 2008; Fu and Anderson, 2006; Janssen and Gray, 2012; Schultz et al., 1997; Stocco, 2017; Walsh and Anderson, 2014, e.g.,). Reinforcement learning is also used to formalize cognitive mechanisms in context of social interaction. For example, Helbing et al. (2005) showed that reinforcement learning can capture turn-taking in a two-player congestion game.

3.1.2 Model evaluation

Models based on theory allow many ways for technical realization (Lewandowsky, 1993; Marewski and Mehlhorn, 2011). As with all model-based research, there are a few limitations. For example, there is no single way to design a model based on theory (Lewandowsky, 1993; Marewski and Mehlhorn, 2011). Personal interpretations, experiences, and preferences for certain techniques play a significant role in the development of models. Further, a model designer needs to manage the balance between being precise enough to capture the issue of interest, while leaving room to allow for generalizations of the findings ("modeling problem", Kangasrääsiö et al., 2017, p. 1). Once a model design is in place, there are different approaches how to set reasonable parameters ("inference problem", Kangasrääsiö et al., 2017, p. 1). As a result, no two model-based studies on the same issue will be the same.

In order to follow a structured process of model design and minimize the risk of unintended biases, Marewski and Mehlhorn (2011) propose five methodological principles: *nested*, *constrained*, *competitive*, *predictive*, and *distributional* modeling. Nested modeling suggests that a model should follow the same structure as the original experiment. The principle of constrained modeling tells to constrict the parameter settings to the task of the original experiment. Competitive modeling says that not just a single model, but multiple alternatives ought to be compared with one another. Distributional modeling means to test whether models can also account for rare phenomena of the original data. And predictive modeling proposes follow-up experiments of model predictions that are not present in the original data. Whenever appropriate I comply to these principles.

In practice, each model has also free parameters. An interesting side question is how strongly the model depends on the model parameters for the goodness-of-fit (Roberts and Pashler, 2000). This I investigate by comparing model predictions for alternative parameter settings. In general, I compare the emerging patterns in the models and their dependence on

model fit with the actual empirical data gathered in the original empirical study. This way, I am able to present qualitative and quantitative evidence for the role of learning involved in the emergence of latent norms in the Volunteer's Dilemma.

4. Value for the field of artificial intelligence

Artificial intelligence is widely regarded as an interdisciplinary field of research in modern science, as it integrates ideas findings and methodologies from many different fields (Russell and Norvig, 2009, p. 1 ff.). The main focus of interest, however, lies on machines that mimic functions of the human mind (Russell and Norvig, 2009, p. 2). Depending on personal motivations, the main interest of research might be how computers and formal models can help to understand the inner working of the human mind. This can be done by modeling certain isolated mechanisms (e.g., Janssen et al., 2012; Payne et al., 2007), or following more holistic approaches on cognition, such as "How can the human mind occur in the physical universe?" (Anderson, 2009). In contrast, technical oriented scientists might ask questions like "How can artificial agents collaborate effectively in order to solve problems?" (e.g., Olfati-Saber et al., 2007; y López et al., 2006), "How can we interact with machines in intuitive ways?" (e.g., Kollar et al., 2010; Severinson-Eklundh et al., 2003; Wu et al., 2016) or "How can machines learn in order to adapt to novel situations?" (e.g., Sutton and Barto, 1998; Watkins and Dayan, 1992). In order to find answers to all of these questions, researchers frequently borrow from many different fields, such as psychology, sociology, computer science, game theory, philosophy and linguistics.

I consider the study at hand interdisciplinary, as well. I seek to explain a sociological phenomenon (norm emergence). This I do by investigating the cognitive mechanism of individuals (psychology). I use reinforcement learning, a formal method that is grounded in psychology and performed using computer technology. Even though my main interest is understanding how human cognition works in situations of social interaction, I believe this study will provide benefits for the field of A.I. on different levels.

For human centered research this study provides a model-based description of human behavior and the corresponding cognitive aspects in the context of social interaction. This ties in with the goal of AI to understand human behavior by modeling it. More specifically, it shows how the investigation of cognitive mechanisms at the individual level can help us

understand social phenomena like the emergence of behavioral patterns at the group level (Anderson, 2002).

Moreover, there is value for technical oriented research. The concept of norms is already utilized for distributed problem solving in multi agent systems and coordination of actions within agent societies (Wooldridge, 2009). Although many studies on the emergence of latent norms in agent-based simulations exist, there are only few human-based studies. Further, these studies mostly focus on sanctioning mechanisms (Diekmann and Przepiorka, 2016). Following a cognitive approach to modeling emergent patterns of interaction provides another perspective to already existing models (e.g., Shoham and Tennenholtz, 1992; Walker and Wooldridge, 1995). Consequently, findings of my study can be used to create coordination in artificial agent societies. This work can further help to lay ground for social behavior within artificial agents that get integrated into human society. That is, enabling agents to act according to human expectations. This would allow intuitive interaction between agent and human without the aspect of learning new ways of interaction for the human.

5. Methods

5.1 General design

Corresponding to the principle of *nested modeling* (Marewski and Mehlhorn, 2011), the layout of my research follows the same structure as the experimental setup used by Diekmann and Przepiorka (2016). I simulate three players engaged in a repeated VOD. However, instead of using human participants to investigate the group-level effects of payoff structures on norm emergence, I use cognitive models in a multi-agent simulation to investigate the role of learning within individuals on the emergence of social norms. This means, that simulated players can differ in the way they incorporate experiences from previous rounds into their decision-making process. More precisely and complying with the principle of *competitive modeling* (Marewski and Mehlhorn, 2011), I compare three different model classes:

1. *Random* – a model that serves as control and always chooses actions randomly no matter what happened before

2. *ClassicQ* – a model that learns the best action depending on game states (classical Q-Learning)

3. *CoordinateX* – a model that plans a sequence of actions ahead by considering expectations towards the actions of co-players

The key question is whether these basic models of learning can reproduce the empirical results of coordination in the Volunteer's Dilemma, as described by Diekmann and Przepiorka (2016).

My general assumption is that the main structure of the cognitive architecture is the same for all players. Therefore, within a single block of runs, all players are instances of the same model class with the same initial parameter settings. However, I am uncertain about the exact structure (modeling problem, Kangasrääsiö et al., 2017) and parameter values (inference problem, Kangasrääsiö et al., 2017) of the model. Therefore, I vary both between runs of the model. This allows to investigate each model by itself, without any interference

or side-effects resulting from cross-comparison. Just like in the original study, I use three conditions with different payoff structures: *Symmetric* ($U_{1,2,3} = 80$, $K_{1,2,3} = 50$), *Asymmetric 1* ($U_{1,2,3} = 80$, $K_1 = 30$, $K_{2,3} = 50$) and *Asymmetric 2* ($U_{1,2,3} = 80$, $K_1 = 10$, $K_{2,3} = 50$).

5.2 Simulation procedure

Fig. 5.1 shows the general procedure of each simulation. The VOD (left box) serves as the overarching framework that coordinates the repeated game playing, calculates scores, and notifies the players of these scores. The VOD behaves exactly the same in each simulation, no matter the type of players engaged in it. As a result, only player models and/or parameter settings differ between simulations.

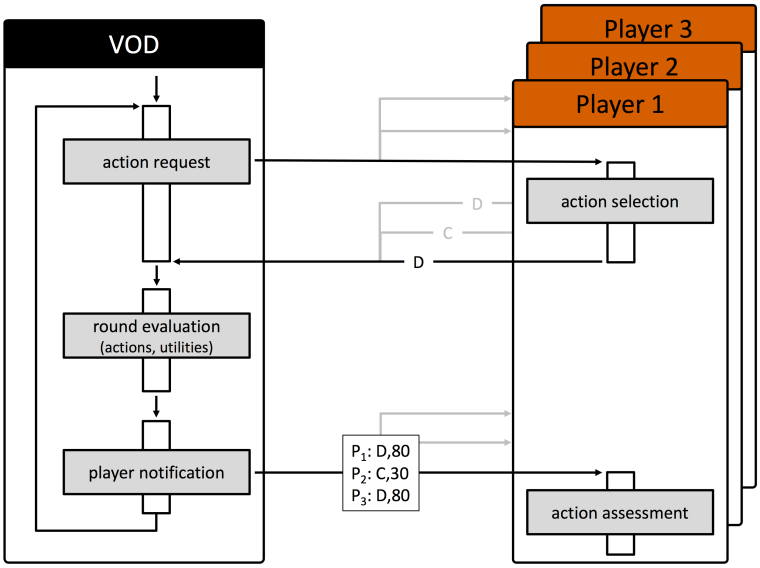

Fig. 5.1 **Simulation procedure.** A simulation consists of two entities: the VOD and a set of players. The VOD coordinates the repeated game playing. The players resemble different theories of human learning-based decision-making. Each round follows the same process: (i) the VOD requests an action for the current round from each player, (ii) players select an action according to their underlying theory of decision-making, (iii) the VOD evaluates the actions of each player, (iv) informs each player about the actions and scores for all players in the current round, (v) players can integrate the information into their decision-making process.

A single round is simulated as follows: First, the VOD requests each player to choose an action (cooperate, defect). Players then choose an action according to their underlying model as set by the modeler. The VOD collects all actions, computes the utilities for each player, and finally notifies each player of the actions and utilities of all players. Players then decide how to incorporate the information concerning that round in their decision-making process.

This procedure is generally repeated 150 times (in special cases up to 5.000, for details see section 5.4). This results in an repeated VOD with 150 rounds (or 5.000 rounds, respectively). Thus, the models simulate a higher number of rounds than in the original experiment (first part: 56 rounds; second part: 48 rounds). This increase is required to ensure a sufficient exploration of the problem state space and to allow investigations of long-term pattern stability. Furthermore, depending on parameter settings and preconditions, computational models may take many more rounds to converge than humans in similar settings (e.g., Helbing et al., 2005). Obviously, this results in a possible shift of fitting measures. For that reason I do not only compare fitting measures, but also the resulting behavioral patterns (see section 5.5).

Note that as a result of this design, only the way how players decide which action to choose (*action selection*, Fig. 5.1) and what to do with the resulting information (*action assessment*, Fig. 5.1) differs between model classes.

5.3 Model classes

In the following, I describe in detail the three different model classes (*Random*, *ClassicQ* and *CoordinateX*) used to investigate the role of learning in the emergence of social norms. For this purpose, special attention is paid to the key processes of *action selection* and *action assessment* (as noted in 5.2 Simulation procedure). For a summary overview, please refer to 5.3.3 Summary of the model classes.

5.3.1 Random

The *Random* model class does not contain any form of learning. Action selection is solely based on a probability distribution. Note that an optimal solution in each round requires exactly one out of three players to cooperate, while the other two players defect. Thus, the probability is 0.33 for cooperation and 0.67 for defection (see Fig. 5.2). Consequently, any information about the actions and utilities of all players is ignored.

The purpose of the *Random* model class is to serve as control condition with no underlying learning-based cognitive processes in play. Since any form of coordination between the players would occur due to pure chance, the results of the *Random* model class are used as a measuring stick to investigate whether learning-based models are able to improve resemblance to the empirical data.

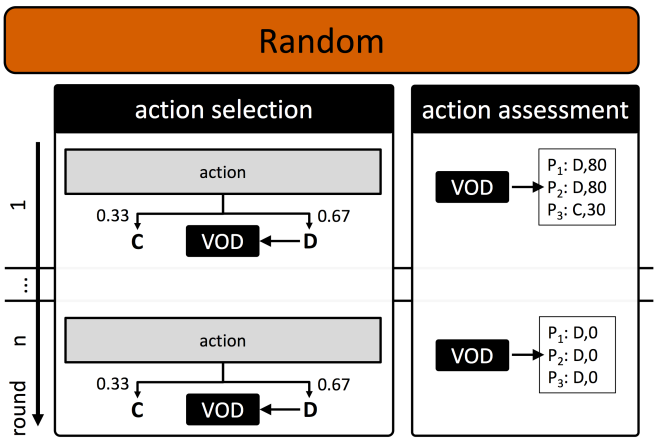

Fig. 5.2 **Decision-making in the *Random* model.** Action selection in the *Random* model does not incorporate any form of learning, but is solely based on a probability distribution of 0.33 for cooperation and 0.67 for defection. That is because the optimal solution in a VOD requires one out of three players to cooperate and two out of three players to defect.

5.3.2 Learning-based model classes

I created two different model classes based on Q-Learning: *ClassicQ* and *CoordinateX*. Both represent the propensities for actions (*ClassicQ*), or strategies (*CoordinateX*), in the form of a Q-Table. Thus, a first parameter is the initial setting for propensities in the Q-Table (ι). Based on the propensity update function (see Formula 3.1), both classes also possess parameters for the learning rate (α) and the discount factor for previous experiences (γ).

As described in section 3.1.1 Reinforcement Learning, both learning-based models require some form of balancing between exploitation of the currently most effective strategy and exploration of all other strategies. For each class I use four different approaches for balancing between exploration and exploitation (E), which I explain in detail below: ε-*greedy*, ε-*decreasing*, ε-*noise* and ε-*noise-decreasing*.

ε-*greedy* uses the ε parameter ($0 < \varepsilon < 1$) to define a ratio between exploration (e.g., $\varepsilon = 0.1 \implies 10\%$ of the rounds) and exploitation ($1 - \varepsilon = 0.9 \implies 90\%$ of the rounds). ε-*decreasing* follows the same general approach like ε-*greedy*. However, parameter δ ($0 < \delta < 1$) is introduced to ensure that ε decreases over time (see Formula 5.1). Thus, ε-*decreasing* represents a learning approach in which a preference is increasingly exploited after an initial phase of strong exploration.

$$\varepsilon - decreasing: \quad \varepsilon_{i+1} = \delta * \varepsilon_i \tag{5.1}$$

ε-*noise* adds a random noise value (ρ) to the actual Q-Values of the current state in the range of $-\varepsilon$ and $+\varepsilon$ (see Formula 5.2). In contrast to ε-greedy and ε-decreasing, where

less preferred strategies are still explored for a certain share of the trials, ε-noise considers the actual propensities for strategies and explores only when preferences are ambiguous. For a direct comparison between ε-greedy and ε-noise decision-making please refer to the decision-making examples in the *ClassicQ* (Fig. 5.3 and Fig. A.1) and *CoordinateX* (Fig. B.1 and 5.4) model classes.

Analogous to ε-*decreasing*, ε-*noise-decreasing* reduces the effect of ε (see Formula 5.1) and thus resembles a situation in which confidence in the existing propensities increases over time. Technically, ε-*greedy* and ε-*noise* conform to ε-*decreasing* and ε-*noise-decreasing* with $\delta = 1$.

$$\varepsilon - noise: \quad Q(s_t;a) = Q(s_t;a) + [Q(s_t;a) * \rho], \qquad \forall a : \{C,D\} \; and \; \rho \sim [-\varepsilon, \varepsilon] \quad (5.2)$$

Finally, I investigate whether social preference (*S*) affects coordination in the VOD. To do this, I compare two different simplified versions of social preference: selfishness and altruism. Selfish players maximize personal rewards. Altruistic players keep track of the rewards of all players and maximize cumulated collective rewards. Technically this is realized in the calculation of maximum expected reward (see Formula 3.1).

An altruistic player expects rewards based group payoffs. That is, a share of the collective reward when only the player with the lowest costs cooperates at a time: $\max_{D,C} Q(s_{t+1};D) = [U + U + (U - K_{\min})]/3$. This calculation is used for both model classes. The calculation for a selfish player depends on the model class (*ClassicQ* or *CoordinateX*), and is explained with the aid of specific examples in the next two sections.

I do this because in scenarios with social context, mental score representations might integrate the scores of co-players. In fact, it has been argued for a long time that we might need to consider also the outcomes of others in scenarios of social interactions (Thibaut and Kelley, 1978a). That is in contrast to some research designs of cognitive science. In situations without social context participant are frequently asked to optimize their behavior by maximizing the outcome of a score that does not depend on the actions of others (e.g., Gray et al., 2006; Janssen and Brumby, 2015; Zhang and Hornof, 2014).

5.3.2.1 ClassicQ

The *ClassicQ* model class is based on a classical Q-Learning approach by reinforcing action-state pairs (see section 3.1.1). That is, *ClassicQ* first determines the state a player is currently in. Then it infers the best action for that state, based on the propensities in the Q-Table.

The set of actions consist of cooperation (C) and defection (D). States in the *ClassicQ* models are defined by actions performed in the previous rounds and depend on two param-

eters: (i) the amount of actions per state ($A \geq 1$) and (ii) the amount of players per state ($\Phi \in \{1,3\}$). The amount of players per state consists either of the player's own actions ($\Phi = 1$) or all players' actions ($\Phi = 3$). Consequently, the amount of actions in a single state is defined by Formula 5.3, and the size of the problem state space is defined by Formula 5.4.

$$\textit{ClassicQ state length}: \quad A * \Phi \tag{5.3}$$

$$\textit{ClassicQ problem state space size}: \quad 2^{(A*\Phi)} \tag{5.4}$$

As an example, Fig. 5.3 shows an instance of *ClassicQ* with parameter settings as depicted in Table 5.1. The scenario corresponds to a symmetric VOD with the same utilities and cooperation costs for all players ($U_{1,2,3} = 80$, $K_{1,2,3} = 50$). Each round (horizontal alignment) is divided into two parts: action selection, and action assessment (see Fig. 5.1 for bigger context). Consecutive rounds of the game are aligned vertically. For a comparison of ε-greedy and ε-noise while all other parameters are the same, please see Fig. 5.3 and Appendix A.

Exploration vs. Exploitation	Social Preference	Initial Propensities	Learning Rate	Discount Rate	Explor. Rate	Explor. Decrease	Actions per State	Players per State
ε-greedy	selfish	$\iota = 50$	$\alpha = 0.4$	$\gamma = 0.6$	$\varepsilon = 0.1$	$\delta = 1$	$A = 2$	$\Phi = 1$

Table 5.1 **Exemplary *ClassicQ* parameter settings.** Each parameter describes a certain aspect of learning. The first seven parameters are shared with the *CoordinateX* class. The last two parameters define a state specific to the *ClassicQ* class.

The process for the *ClassicQ* model is as follows: In the beginning of each simulation, *ClassicQ* needs to select actions randomly to build up experience (*action selection* in Fig. 5.3). This is done with a cooperation ratio of 0.33 as in the *Random* model class. Initial built-up of experience is required, because a state is, inter alia, defined by the amount of actions in the previous rounds. Once a state can be defined (starting from round 3 in Fig. 5.3), actions can be selected according to the Q-Table. In case of ties (as in round 3), actions are chosen randomly with an equal probability of 0.5. That is to ensure that actions get the same chance to be explored when there is no clear preference. In case that there is a preference for one of the actions (as in round n), ε-greedy ensures that the less preferred action is explored with a chance of 0.1, while the more preferred action is exploited with a chance of 0.9.

In order to update the Q-Values, the maximal expected utility must be defined in each round (last step of *action selection*). In case of defection (as in round 3), a player expects another to cooperate and thus to receive a reward of $U = 80$ herself. Whenever a player

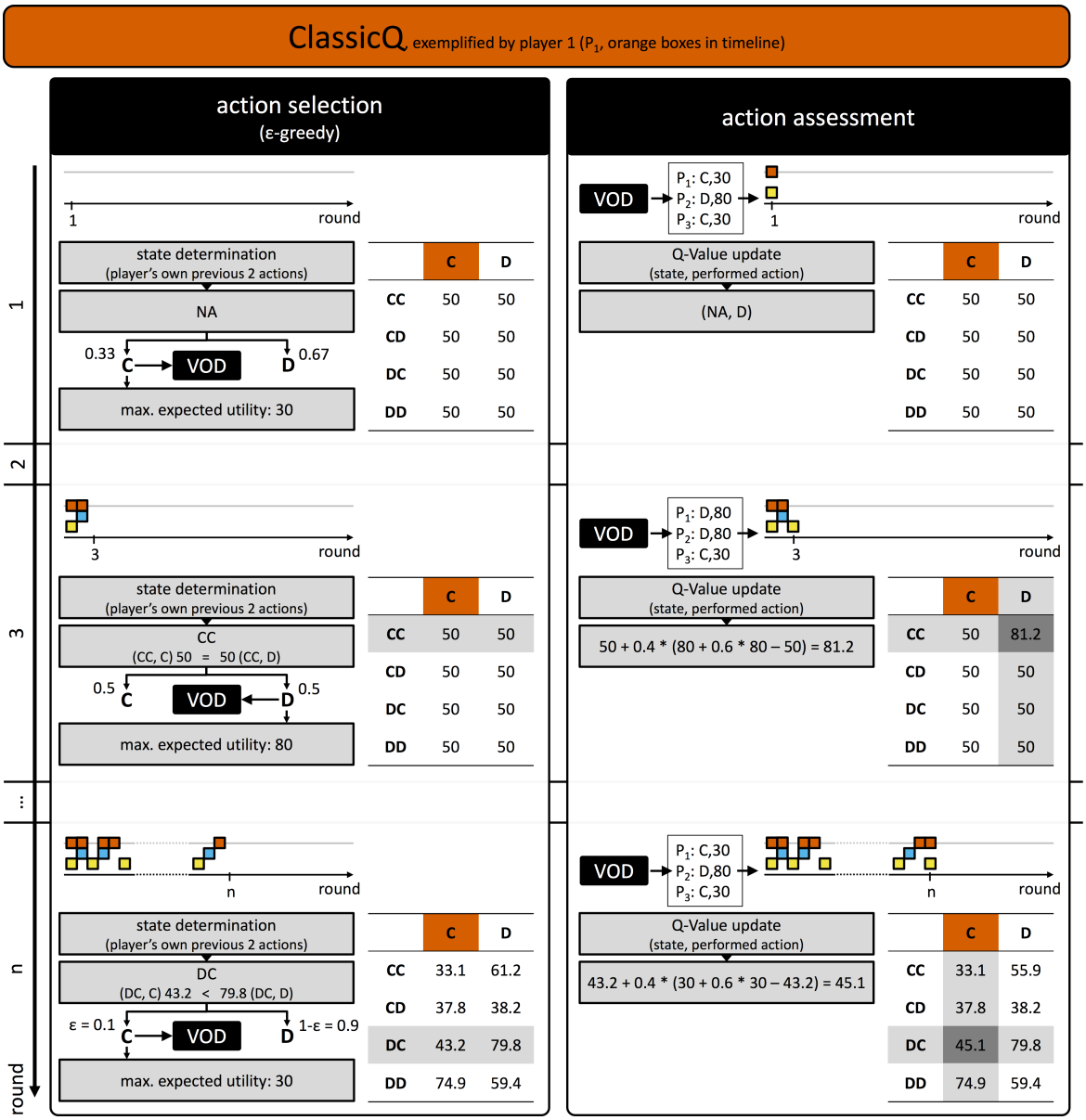

Fig. 5.3 **Learning and decision-making in the *ClassicQ* model (ε-greedy).** Learning and decision-making consists of two parts: *action selection* and *action assessment* (horizontal alignment). *Action selection* is based on propensities for action state pairs (Q-Table). Here, a state is defined by the previous two actions of the player (rows). Before a state can be determined (rounds 1,2 – vertical alignment), a player cooperates with a probability of 0.33. This is because the optimal solution for a VOD is when one out of three players cooperates in a single round. In case of ties (round 3), actions are selected randomly with equal probabilities (0.5 each), to allow a fair exploration of both actions. In all other cases, ε-greedy balances between exploration of less preferred strategies (round n). That is, with a probability of 0.1 the less preferred action is being explored, while with a probability of 0.9 the more preferred action is being exploited. As a final step, the maximum expected utility is defined – D: $U = 80$; C: $U - K = 80 - 50 = 30$. During *action assessment* the propensity for the performed action is updated using Formula 3.1.

cooperates (as in round n), the maximum expected utility is $U - K = 80 - 50 = 30$. During *action assessment* a player adds the information about the actions of all players to her private game history and updates the corresponding Q-Values according to Formula 3.1 (as in *action assessment* rounds 3 and n).

5.3.2.2 CoordinateX

In contrast to *ClassicQ*, *CoordinateX* is a modified version of Q-Learning that plans a sequence of actions ahead and assesses success by comparing the expected rewards with the actual outcome. The basic idea of *CoordinateX* is inspired by the principle of latent norms (Opp, 2004; Wrong, 1994).[1] A player incorporates expectations towards the actions of others in the decision-making process based on experiences from the past. For example: A player chooses the strategy to defect twice and cooperate once. That is because she expects any of her co-players to cooperate in the next two rounds, but to defect in the third round. The basis for the decision is the experience that this strategy worked best in previous rounds, and the expectation that it will persist.

This future-oriented behavior is represented by a set of strategies, which simply define a sequence of actions. For these strategies, X denotes the expected group size. Technically X describes the round when to cooperate the latest (see Table 5.2). A player who expects a group size of three players, has a set of four strategies: $\{D, C, DC, DDC\}$. More strategies (e.g., *DDDC*) are not feasible, because this would require a single co-player to cooperate at least twice before the player cooperates herself.

x	$X = 1$	$X = 2$	$X = 3$	...	$X = n$
0	{D,	{D,	{D,		{D,
1	C}	C,	C,		C,
2		DC}	DC,		DC,
3			DDC}		DDC,
⋮					⋮
n					[(n-1)*D]C}

Table 5.2 ***CoordinateX* strategies.** X denotes the expected group size. Technically, X specifies the round when to cooperate the latest. A player who expects to be in a group of three ($X = 3$), has four strategies: "D" (immediate defection), "C" (immediate cooperation), "DC" (defection followed by cooperation), and "DDC" (double defection followed by single cooperation).

[1] To be clear: *CoordinateX* is merely inspired by the concept of latent norms, not a full-fledged formalization.

As plain defection (*D*) is a strategy in all possible strategy sets, the the amount of available strategies is always one element bigger than X. The set of available strategies complies to the problem state space of *CoordinateX*, and is defined by Formula 5.5.

$$CoordinateX\ problem\ state\ space\ size: \quad X+1 \tag{5.5}$$

The process of learning in the *CoordinateX* model classes is realized basically in the same way as in *ClassicQ*. The only difference is that instead of reinforcing action-state pairs, *CoordinateX* reinforces the available strategies. Fig. 5.4 shows an instance of *CoordinateX* with parameter settings as depicted in Table 5.3. As before, the scenario corresponds to a symmetric VOD with the same utilities and cooperation costs for all players ($U_{1,2,3} = 80$, $K_{1,2,3} = 50$). The model instance in this example, however, uses ε-noise to balance between exploration and exploitation. For a comparison between ε-noise and ε-greedy in the *CoordinateX* model, please see Fig. 5.4 and Appendix B. Again, each round (horizontal alignment) is divided into two parts: action selection, and action assessment (see Fig. 5.1 for bigger context). Consecutive rounds of the game are aligned vertically.

Exploration vs. Exploitation	Social Preference	Initial Propensities	Learning Rate	Discount Rate	Explor. Rate	Explor. Decrease	Max. Coord. Position
ε-noise	selfish	$\iota = 50$	$\alpha = 0.4$	$\gamma = 0.6$	$\varepsilon = 0.1$	$\delta = 1$	$X = 3$

Table 5.3 **Exemplary *CoordinateX* parameter settings.** Each parameter describes a certain aspect of how learning is realized. The first seven parameters are shared with the *ClassicQ* model class. The X parameter defines the available strategies (see Table 5.2).

In contrast to *ClassicQ*, *CoordinateX* does not require to build up experience first, because the future-oriented strategies are available from the first round. Further, the depicted ε-noise approach adds noise to the propensities (rounds 1, n), so that even in situations without clear preferences (e.g., with initial propensities as in round 1) a clear decision can be made. Once a strategy has been chosen, all associated actions are performed sequentially (e.g., round 2). Only when no more actions are left, a new strategy needs to be selected (round n).

As in action assessment for *ClassicQ*, an expected utility is required to assess the success of a strategy. However, to allow a fair assessment for each strategy (no matter the amount of actions within it) the average value for expected utility in each round is considered. For example strategy *DDC* as in round 1: $[U+U+(U-K)]/3 = [80+80+(80-50)]/3 = 63.3$. If total amounts were to be taken, strategies with more actions would have a very high lever making a direct comparison of different strategies disproportionate. During *action assessment* a player first adds the information about the actions of all players to her game history and then updates the corresponding strategy according to Formula 3.1.

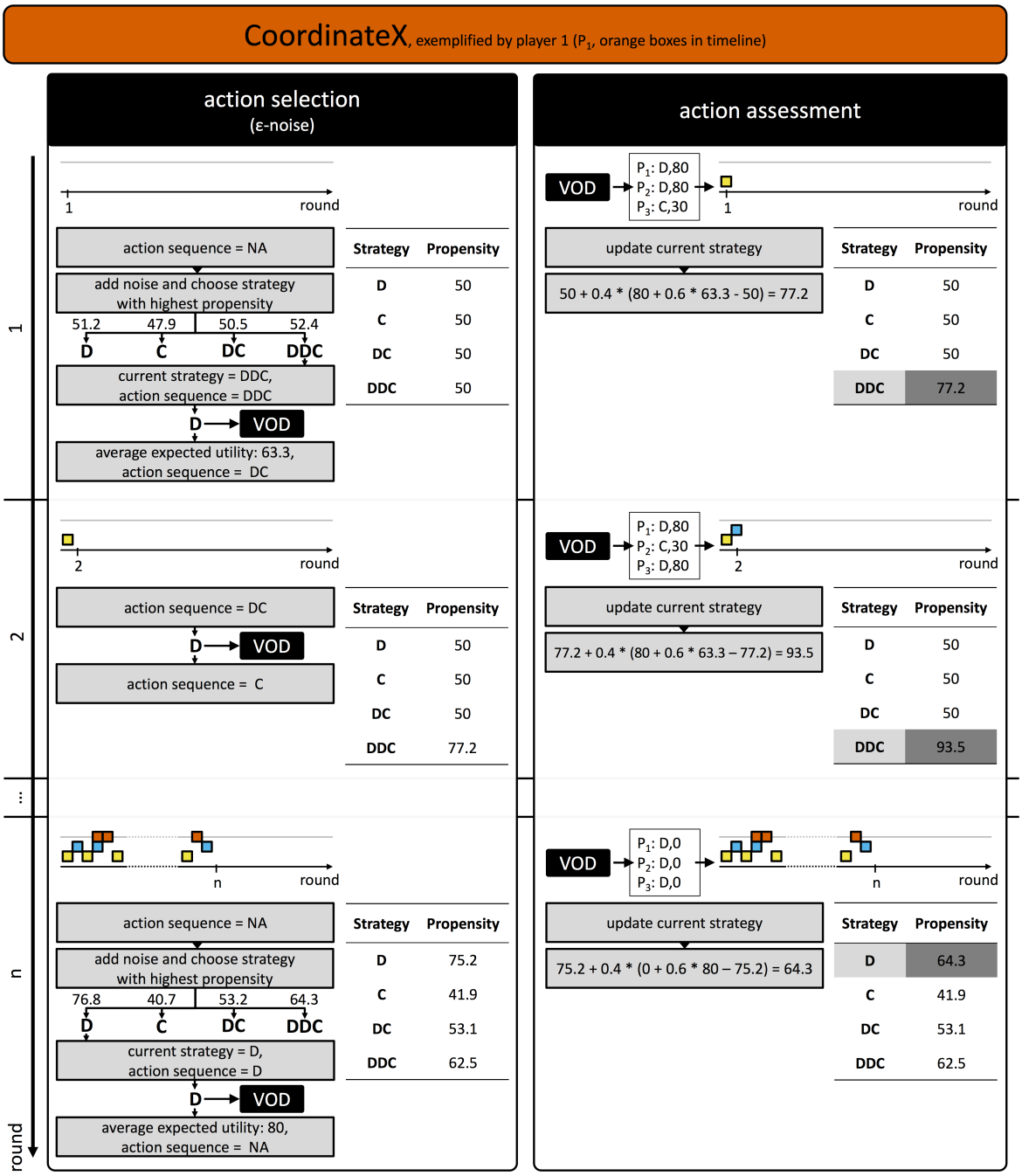

Fig. 5.4 **Learning and decision-making in the *CoordinateX* model (ε-noise).** Learning and decision-making consists of two parts: *action selection* and *action assessment* (horizontal alignment). *Action selection* is based on propensities for strategies (tables). A strategy (rows) is defined by a distinct sequence of actions, which are to be performed in the upcoming round(s). ε-noise simply adds a random noise factor to each propensity, which makes clear decisions even in situations with equal preferences possible (round 1). As long as there are still actions left in the current strategy (round 2), they will be performed sequentially. Only when no more actions are left, a new strategy is selected (round n). As a final step for strategy selection, the average expected utility is defined, in order to allow a fair assessment no matter the amount of actions within it. For example round 1: $[U+U+(U-K)]/3 = [80+80+(80-50)]/3 = 63.3$. During *action assessment* the propensity for the strategy is updated using Formula 3.1.

5.3.3 Summary of the model classes

I created a total of three different model classes: *Random*, *ClassicQ*, and *CoordinateX* (for a parameter summary, see Table 5.4). The *Random* model class serves as control condition. It selects actions randomly (cooperates with a probability of 0.33), and therefore does not use any form of learning. The *ClassicQ* and *CoordinateX* classes use versions of Q-Learning for action selection and action assessment. The main difference between *ClassicQ* and *CoordinateX* is the basis for decision-making. *ClassicQ* reinforces action-state pairs. *CoordinateX*, on the other hand, plans a sequence of actions ahead and evaluates the efficiency by comparing the expected reward with the actual reward.

5.4 Model fitting

The three model classes have multiple parameters, with little theory to guide the decision for specific parameter values. Therefore, I use a model fitting procedure to identify suitable parameters. The goal is to see whether the model is, in principle, able to fit the data and to understand under what conditions. The goal is not to achieve a perfect fit between model and empirical results, as overly fitted models are not necessarily a better evidence for the underlying theories. In fact, they might even lose fundamental characteristics like *flexibility*, *variability of data*, and *likelihood of unexpected outcomes* (Roberts and Pashler, 2000, p.359).

However, due to the amount of available parameters, the corresponding parameter ranges and a lack of agreement on general settings for some of the parameters (e.g., α, γ), which makes educated guesses hard if not even impossible, I implemented a simple parameter fitting procedure. This is done, following the principle of *constrained modeling* (Marewski and Mehlhorn, 2011), to roughly constrict the parameter settings to the task of the original experiment.

As for categorical parameters (see Table 5.4), I explored the full range of parameters: $E \in \{\varepsilon\text{-greedy}, \varepsilon\text{-decreasing}, \varepsilon\text{-noise}, \varepsilon\text{-noise-decreasing}\}$; $S \in \{\text{selfish, altruistic}\}$. Numerical parameters haven been explored using parameter sets in an iterative approach. Two parameter sets have been selected based on educated guesses. First, $\varepsilon \in \{0.05, 0.1, 0.2, 0.4\}$ resembles an agent with different exploration behavior ranging from very conservative to highly exploratory. Second, $\delta \in \{0.98, 0.995\}$ resembles agents that are either more or less explorative in the late stage of a game.

ClassicQ players regarded either two or three previous rounds ($A \in \{2,3\}$), and either their own actions or the actions of all players ($\Phi \in \{1,3\}$). *CoordinateX* players expected group sizes of two, three, and four players ($X \in \{2,3,4\}$). Initial propensities for actions

	Random	ClassicQ	CoordinateX
I. Principles of decision-making			
Formal Framework	Randomness	Q-Learning	Modified form of Q-Learning
Action Selection	random	based on propensities for actions	based on propensities for strategies
Action Assessment	–	propensity update	propensity update
Decision Basis	–	action-state pairs	future-oriented strategies
II. Categorical Parameters			
Exploration vs. Exploitation	–	$E \in \{\varepsilon$-greedy, ε-decreasing, ε-noise, ε-noise-decreasing$\}$	$E \in \{\varepsilon$-greedy, ε-decreasing, ε-noise, ε-noise-decreasing$\}$
Social Preference	–	$S \in \{$selfish, altruistic$\}$	$S \in \{$selfish, altruistic$\}$
III. Numerical Parameters			
Coop. Ratio	0.33	0.33	–
Initial Propensities	–	$\iota \geq 0$	$\iota \geq 0$
Learning Rate	–	$0 < \alpha \leq 1$	$0 < \alpha \leq 1$
Discount Rate	–	$0 \leq \gamma \leq 1$	$0 \leq \gamma \leq 1$
Explor. Rate	–	$0 < \varepsilon < 1$	$0 < \varepsilon < 1$
Explor. Decrease	–	$0 < \delta < 1$	$0 < \delta < 1$
Actions per State	–	$A \geq 1$	–
Players per State	–	$\Phi \in \{1, 3\}$	–
Exp. group size	–	–	$X \geq 1$

Table 5.4 **Summary of the model classes.** The *Random* model class has only one parameter, namely cooperation ratio. This is a fixed value of 0.33 in order to resemble that an optimal solution in each round requires exactly one out of three players to cooperate, while the other two players defect (for details see section 5.3.1). Note, that the cooperation ratio for *ClassicQ* models only applies in early stages of the simulation. That is when not enough rounds were simulated to define the current state. All other parameters of *ClassicQ* and *CoordinateX* define the characteristics of learning and decision-making (for details see section 5.3.2).

and strategies were defined to cover a broad range of values, because the actual effects were uncertain ($\iota \in \{0, 60, 100, 120, 300\}$). α and γ were selected randomly, as there is no consensus standard settings for these values. All parameter sets have been combined with another, resulting in a total amount of 4.864 unique parameter combinations (ClassicQ: 1.948, CoordinateX: 2.916). A single simulation consisted of a unique parameter combination for the three conditions (Symmetric, Asymmetric 1 and Asymmetric 2) and ten repeated VOD scenarios per condition.

The vast majority of simulations for *ClassicQ* (1.944) used only a small problem state, with only the player's own actions defining a state ($\Phi = 1$). That is due to the fact that the problem state size increases exponentially (see Formula 5.4), resulting in 512 different states for all three players ($\Phi = 3$) and three actions per state ($A = 3$). The standard amount of 150 rounds for a repeated VOD simulation does not allow to explore the full problem state space, not to mention exploitation of most preferred strategies. For that reason and due to computational constraints, another four simulations using the best parameter fit for *ClassicQ* and the actions of all three players defining a state ($\Phi = 3$) were performed. These simulations were used to compare the different four balancing strategies for exploration and exploitation (ε-greedy, ε-decreasing, ε-noise, ε-noise-decreasing) and consisted of repeated VODs with 5.000 rounds.

5.5 Analysis of the simulation and model validation

Based on exemplary model predictions, I compare model and empirical results in three steps. In the first step I investigate whether learning is a crucial factor for the emergence of tacit norms in general. Or in other words: Do the models predict the same behavioral patterns as they occur in the empirical data? To answer this question, I first calculate the *Latent Norm Index* (*LNI*), as described by (Diekmann and Przepiorka, 2016, p. 1318 f.). This measure provides the ratio of behavioral patterns compared to the overall amount of rounds. Just as in the experiment, I analyze three different types of patterns: solitary volunteering, turn-taking between two players, and turn-taking between three players.

One important formal requirement is that, in a three player game, a pattern needs to be stable for at least three consecutive rounds. Consider, for example, the following sequence of actions. Numbers denote a cooperative player in a single round: 1112312123. In order to comply with 'solitary volunteering', one player needs to cooperate three times in a row while no other player cooperates at the same time. 1112312123 has a sequence of player 1 cooperating three times in a row. A game of 10 rounds therefore results in a ratio of 30% solitary volunteering.

In order to comply with 'turn-taking between two players', two players need to cooperate successively for at least three consecutive rounds. Here, 11123121213 maps to 40% turn-taking between two players. 1112312123, 1112312123, 1112312123, and 1112312123 alternate only in two consecutive rounds. Thus, these sections do not comply with 'turn-taking between two players'.

In order to comply with 'turn-taking between three players', all three players need to cooperate successively for at least three consecutive rounds. Consequently, 1112312123 maps to 70% turn-taking between three players. For a detailed description of the LNI algorithm, please refer to Diekmann and Przepiorka (2016, p. 1318 f.).

In contrast to the empirical data, which is described using the mean *LNIs* over all experimental trials, I use the median *LNIs* over all 10 simulations for each unique parameter combination. This is due to heavily skewed distributions of the mean score differences in some of the model predictions.

Following the LNI computations, I use root-mean-square errors (*RMSE*) and normalized root-mean-square errors (*NRMSE*) to describe the error size between model predictions and empirical data. NRMSE is computed by comparing the value of RMSE to the standard defection of the empirical data, with an RMSE that has the value of 1 SD equating to 100%. To investigate how much of the variation in the empirical data can be explained using the models, I use R-Squared (R^2) measures.

For the learning-based models, I define a *relative model fit*. That is when RMSE and NRMSE show lower values and R2 shows a higher value compared to the predictions of the Random model. This is because any form of coordination in the Random model would be due to pure chance. And therefore, an increase of model fit over Random predictions indicates an improvement of coordinative behavior.

In the second step I investigate whether emerging behavioral patterns have similar qualitative aspects as the empirical data. That means, I investigate (i.) how many rounds a pattern needs to emerge, (ii.) whether patterns persist once they have emerged, and (iii.) whether certain parameters have a direct influence on speed and stability of tacit norm emergence.

Consider a ratio of 50% turn-taking between three actors. This can occur for various reasons. One example of such a pattern is a block of trials in which the first half does not have any turn-taking, but the later half does. Another example for 50% turn-taking is when turn-taking occurs from the beginning, but is periodically disrupted (e.g., three rounds of turn-taking followed by three rounds of no cooperation at all).

Furthermore, I mostly compare 150 simulated rounds with about 50 experimental rounds (see section 5.2). Suppose that humans and simulations require 10 rounds to coordinate

(humans: 20% of all rounds; simulation: 6.67% or all rounds). Consequently, this results in different fitting measures, but shows the same speed and stability of pattern emergence.

As an alternative, I could compare the same amount of rounds. For example, I could analyze the last 50 simulated rounds. However, this would not provide information on how patterns emerged. Another alternative could be to simulate only 50 rounds in total. But as described earlier (see section 5.2), for big problem state spaces it might not suffice to converge. Further, this would not allow to examine long-term pattern stability.

In the wake of this, I also investigate whether rare patterns of the empirical data are predicted by the models. For example, turn-taking between two players sometimes occurs in symmetric VODs. This follows the principle of *distributional modeling* (Marewski and Mehlhorn, 2011) and serves as another indicator of model validity.

In the third and final step I describe whether certain characteristics of learning are necessary for the emergence of tacit norms. Or in other words: Do certain parameter settings suppress the emergence of tacit norms? To analyze this, I investigate the effects of different isolated parameter settings on the fitting measures. That means, I compare the average model fit for certain parameter settings (e.g., social preference = selfish vs. altruistic) with the fitting measures of the *Random* model. An increase in model fit for only one of the values is a strong indicator for a necessary parameter setting.

6. Results

In the following sections I investigate three things. First, I investigate whether certain model instances predict the same patterns as in the original experiment by Diekmann and Przepiorka (2016). Second, I investigate whether certain characteristics of learning have major influence on speed and stability of norm emergence. Third, I investigate whether certain characteristics of learning are necessary for norms to emerge.

Eight representative model instances provide the basis for my analysis: *ClassicQ*: CQ.362, CQ.1280, CQ.1442, CQ.1947; *CoordinateX*: CX.1009, CX.1981, CX.1983, CX.2807. The names describe the model type (CQ, CX), and the number of the simulation per model type. They do not provide further information about specific parameter settings or instance characteristics. For a detailed overview of the instances' parameter settings and fitting measures, please refer to Table 6.1.

I selected these instances, because they had the best fitting measures and most representative behavioral patterns. Thus, they show the general capabilities of the model classes. Further, all instances within each model class have only slight variations in parameter settings. This allows to compare the effects of parameter settings on model predictions.

6.1 General performance

Fig. 6.1 compares the observed empicial pattern (black bars) with model predictions (colored bars) for the three model types (rows). That is, for *Random* (gray) and the best performing model instances for *ClassicQ* (orange) and *CoordinateX* (blue).

The *Random* model's reference fit was: $RMSE = 34.2$ ($NRMSE = 165.8\%$, $R^2 = 0.12$). The *ClassicQ* and *CoordinateX* models had better fits overall. *CoordinateX* had the best fit with $RMSE = 6.92$ ($NRMSE = 33.5\%$, $R^2 = 0.94$). *ClassicQ* by comparison had a moderate fit with $RMSE = 20.83$ ($NRMSE = 101\%$, $R^2 = 0.48$).

	Random	**ClassicQ**				**CoordinateX**			
		CQ.362 (best fit)	CQ.1280	CQ.1442	CQ.1947	CX.1009	CX.1981	**CX.1983 (best fit)**	CX.2807
I. Parameter Settings									
Exploration vs. Exploitation	–	**ε-greedy**	ε-noise	ε-noise	ε-noise	ε-decreasing	ε-noise-decreasing	**ε-noise-decreasing**	ε-noise
Social Preference	–	**selfish**	selfish	selfish	selfish	selfish	selfish	**selfish**	selfish
Coop. Ratio	0.33	**0.33**	0.33	0.33	0.33	–	–	**–**	–
Initial Propensities	–	**$\iota = 120$**	$\iota = 60$	$\iota = 120$	$\iota = 60$	$\iota = 100$	$\iota = 100$	**$\iota = 100$**	$\iota = 300$
Learning Rate	–	**$\alpha = 0.25$**	$\alpha = 0.25$	$\alpha = 0.25$	$\alpha = 0.25$	$\alpha = 0.4$	$\alpha = 0.4$	**$\alpha = 0.4$**	$\alpha = 0.6$
Discount Rate	–	**$\gamma = 0.6$**	$\gamma = 0.6$	$\gamma = 0.6$	$\gamma = 0.6$	$\gamma = 0.5$	$\gamma = 0.5$	**$\gamma = 0.5$**	$\gamma = 0.9$
Exploration Rate	–	**$\varepsilon = 0.05$**	$\varepsilon = 0.2$	$\varepsilon = 0.2$	$\varepsilon = 0.2$	$\varepsilon = 0.05$	$\varepsilon = 0.05$	**$\varepsilon = 0.05$**	$\varepsilon = 0.1$
Exploration Decr.	–	**$\delta = 1$**	$\delta = 1$	$\delta = 1$	$\delta = 1$	$\delta = 0.98$	$\delta = 0.98$	**$\delta = 0.98$**	$\delta = 1$
Actions per State	–	**$A = 3$**	$A = 3$	$A = 3$	$A = 3$	–	–	**–**	–
Players per State	–	**$\Phi = 1$**	$\Phi = 1$	$\Phi = 1$	$\Phi = 3$	–	–	**–**	–
Exp. group size	–	**–**	–	–	–	$X = 2$	$X = 2$	**$X = 4$**	$X = 3$
II. Goodness of Fit									
II.a. Combined									
RMSE	34.2	**20.83**	25.19	21.65	28.3	29.44	32.08	**6.92**	19.14
NRMSE	165.8	**101**	122.1	105	137.6	142.7	155.5	**33.5**	92.8
R^2	0.12	**0.48**	0.45	0.48	0.43	0.39	0.38	**0.94**	0.55
II.b. Symmetric									
RMSE	32.3	**34.41**	35.63	35.79	35.06	35.04	37.85	**5.79**	11.76
NRMSE	141.6	**150.8**	156.1	156.9	153.7	153.6	165.9	**25.4**	51.5
R^2	0.17	**0.11**	0.12	0.12	0.18	0.14	0.16	**0.96**	0.78
II.c. Asymmetric 1									
RMSE	27.96	**9.76**	20.99	10.49	29.52	32.3	34.75	**9.44**	27.38
NRMSE	165.4	**57.7**	124.2	62.1	174.7	191.1	205.6	**55.8**	162
R^2	0.5	**0.97**	0.49	0.95	0.29	0.25	0.21	**0.93**	0.48
II.d. Asymmetric 2									
RMSE	41.05	**4.7**	13.9	3.85	17.72	18.14	21.13	**4.57**	14.54
NRMSE	149.5	**17.1**	50.6	14	64.6	66.1	77	**16.7**	53
R^2	0.01	**0.97**	0.91	0.99	0.87	0.85	0.83	**0.99**	0.66

Table 6.1 **Parameter settings and fitting measures of representative model instances.** Columns contain nine model instances: Random, four ClassicQ and four CoordinateX. Rows show parameter settings (I.) and fitting measures (II.). Fitting measures are divided into combined fit for all types of VODs (II.a) and each type of VOD individually (II.b.–d.). All model instances have representative fitting measures and representative behavioral patterns.

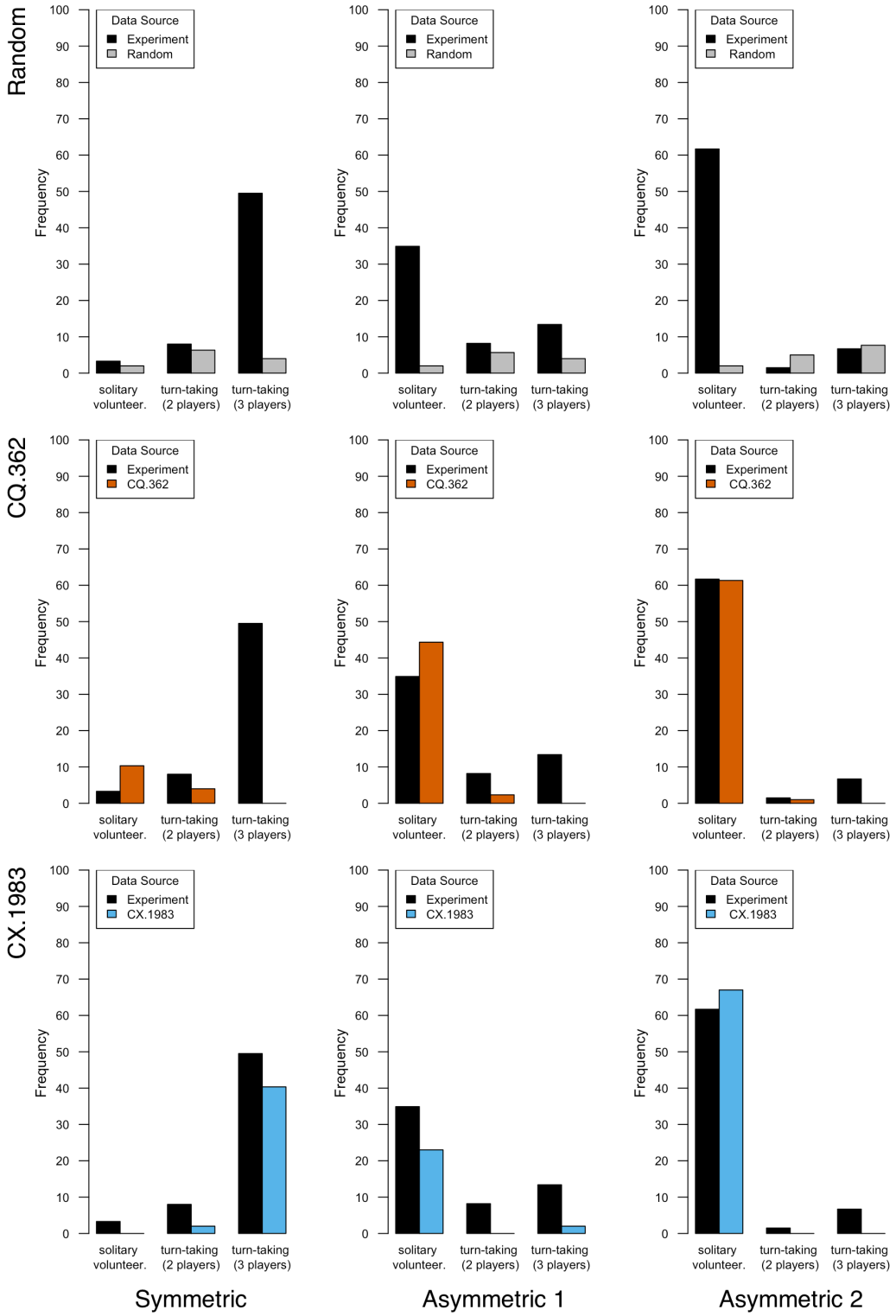

Fig. 6.1 **Ratios of behavioral patterns for the best performing model instances.** The figure is divided into nine plots. Rows contain plots for the different model instances (Random, CQ.362, CX.1983). Columns contain plots for the different conditions (Symmetric, Asymmetric 1, Asymmetric 2). Each plot consists of six bars, depicting pattern ratios: solitary volunteering, turn-taking between two players, and turn-taking between three players. Three bars show the ratios for the empirical data (black). The remaining bars depict the ratios of the corresponding model instances (gray: *Random*, orange: CQ.362, blue: CX.1983).

For asymmetric VODs, both model classes had similarly good fits. Especially the reference fit for Asymmetric 2 was clearly improved. The fit of the *Random* model was $RMSE = 41.05$ ($NRMSE = 149.5\%$, $R^2 = 0.01$). *CoordinateX* had a very good fit with $RMSE = 4.57$ ($NRMSE = 16.7\%$, $R^2 = 0.99$). *ClassicQ* also showed a very good fit with $RMSE = 4.7$ ($NRMSE = 17.1\%$, $R^2 = 0.97$). Further, Fig. 6.1 shows that both learning-based models predicted similar ratios of solitary volunteering for the asymmetric VODs as found in the experiment. Note that the lower ratio for solitary volunteering in Asymmetric 1 compared to Asymmetric 2 is also predicted accurately.

For the symmetric VOD however, only *CoordinateX* showed a good fit. The reference fit of the *Random* model was: $RMSE = 32.3$ ($NRMSE = 141.6\%$, $R^2 = 0.17$). *CoordinateX* had a very good fit with $RMSE = 5.79$ ($NRMSE = 25.4\%$, $R^2 = 0.96$). *ClassicQ* however, had a poor fit with $RMSE = 34.41$ ($NRMSE = 150.8\%$, $R^2 = 0.11$). Fig. 6.1 shows that CQ.362 predicted only a very small ratio of turn-taking between two actors. Turn-taking between three actors, the most common behavioral pattern in the experiment, was not predicted at all. *CoordinateX* on the other hand, predicted almost the same ratio of turn-taking between three actors as in the experiment.

In summary, these results show that both learning-based model classes successfully predict similar ratios of solitary volunteering for asymmetric VODs as found in the empirical data. Further, only *CoordinateX* predicts turn-taking between three actors for the symmetric VOD. The ratios produced by *CoordinateX* for all types of VOD are very close to the ratios of the empirical data.

6.2 Speed and stability of pattern emergence

An open question is whether the models that have the best quantitative fit also show the best learning speed and stability of the human pattern. In order to allow all models to explore the full range of parameter space and investigate long-term pattern stability, I simulated 150 rounds instead of about 50 rounds as in the experiment (see section 5.2). This has a direct effect on fitting measures (see section 5.5). For that reason, I investigate in the following section whether the patterns of model predictions show similar speed and stability of pattern emergence, as in the experiment.

6.2.1 ClassicQ

As the pattern ratios in Fig. 6.1 suggest, there is no form of stable and reliable coordination for *ClassicQ* in the symmetric condition. Fig. 6.2 shows two exemplary interaction and

convergence patterns for the symmetric VOD. The upper plot for each instance (CQ.362 and CQ.1947) shows the interactions of all three actors. An 'x' denotes 'cooperation'. The lower plot shows whether a stable pattern emerges. Remember that for the formal recognition of a pattern, at least three consecutive rounds of a sequence is required (see section 5.5).

Consider, for example, player 1 in simulation CQ.362. In rounds 29 till 34, player 1 cooperates five times in a row without any other player cooperating at the same time. This translates into a pattern of solitary volunteering. However, in rounds 56 and 57, player 1 cooperates twice in a row. This sequence does not meet the requirement of three consecutive rounds and thus is not recognized as a pattern. A different pattern can be found in rounds 2, 3 and 4. Here, player 1 and 2 alternate which translates into a pattern of turn-taking between two players.

Although there are some short stretches of patterns for CQ.362 in the symmetric VOD, none of them is stable over a longer period of rounds. In fact, behavior seems random throughout the whole course of the simulation. Further, sub-optimal behavior predominates most of the rounds. That is, either none or more than one actor cooperates at the same time.

Various parameter combinations are possible within the *ClassicQ* model. In the following, I discuss the impact of specific combinations on model fit, and explain why the parameter values lead to a specific result.

Just as CQ.362, CQ.1947 does not show any form of coordination in the symmetric condition either. In contrast to CQ.362, CQ.1947 has an increased problem state space. That is because decision-making of CQ.1947 is not only based on the player's own actions in the previous three rounds (CQ.362: $A = 3$, $\Phi = 1$), but based on the last three actions of all players (CQ.1947: $A = 3$, $\Phi = 3$). This means CQ.1947 has a much more complex representation of memory, considering all the information available in the previous rounds.

For both instances, CQ.362 and CQ.1947, the amount of rounds should suffice to converge at some form of optimal behavior. That is because the problem state space in both cases is much smaller than the amount of rounds (CQ.362: $2^{A*\Phi} = 2^{1*3} = 8$; CQ.1947: $2^{A*\Phi} = 2^{3*3} = 512$; for details see Formula 5.4). This means that all states should have had the chance to be visited at least once during the course of a simulation. Furthermore, a total of 19.480 simulations were computed (1.948 unique parameter combinations, with 10 simulations each). None of them showed any form of stable coordination in the symmetric VOD.

However, stable coordination did emerge in the asymmetric VOD. The lower six plots of Fig. 6.2 depict exemplary results for the asymmetric conditions: five interaction patterns (Asymmtric 2: CQ.362, CQ.1442, CQ.1280, CQ.1947; Asymmetric 1: CQ.1280) and one convergence pattern (Asymmetric 2: CQ.1947). All five simulations show solitary

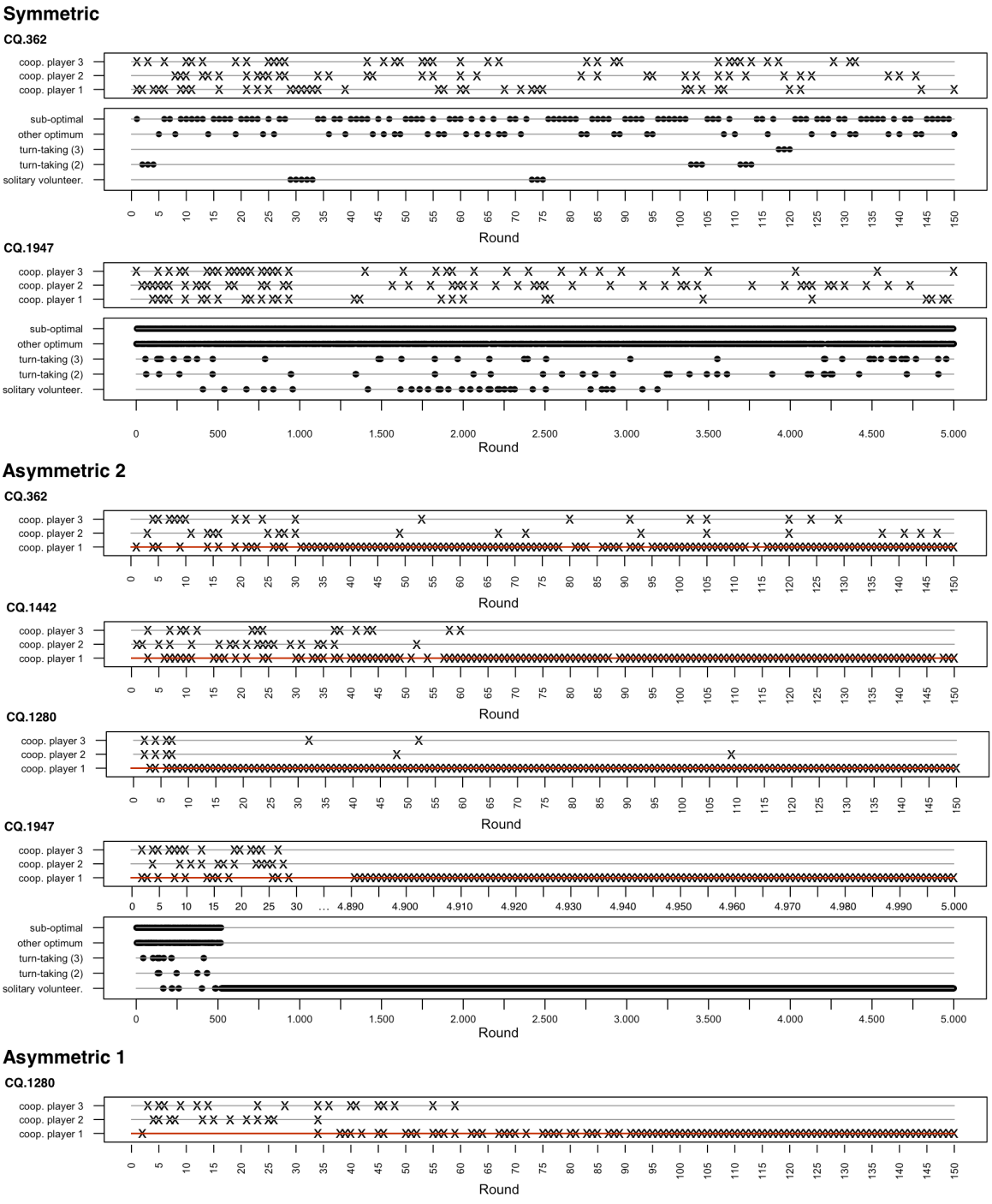

Fig. 6.2 **Exemplary patterns (*ClassicQ*).** Two types of patterns are presented: behavioral patterns and convergence patterns. Behavioral patterns depict cooperation of players in corresponding rounds with an 'x'. Convergence patterns show whether a sequence of actions complies with one of the considered patterns (marked with a dot). Note that a sequence of at least three consecutive actions is required to be recognized as a pattern. For example, 'solitary volunteering' requires a single player to volunteer three times in a row, while no other player cooperates at the same time (e.g., Symmetric, CQ.362, player 1, rounds 29–34). 'Turn-taking (3)' requires a sequence of cooperation where each player successively cooperates for at least three rounds (e.g., Symmetric, CQ.362, rounds 118–120). For details see section 5.5. Red lines in the asymmetric conditions mark the player with the lowest costs to cooperate.

volunteering of the actor with the lowest cooperation costs (player 1, highlighted with a red line). However, they differ in speed and stability of pattern emergence.

CQ.362 showed the best overall fitting measures. This means a similar ratio of solitary volunteering is predicted by the model as found in the empirical data. However, the models simulate 150 rounds and are compared to about 50 experimental rounds. As a result, a stable pattern requires quite long to emerge ($\sim$ 30 rounds). Further, it is occasionally disrupted by either coordination of other players or defection of player 1.

Instead of ε-greedy, as used by CQ.362, CQ.1442 uses ε-noise to balance between exploration and exploitation. This still results in about equally good fitting measures (see Table 6.1). However, the patterns are much more stable, but require around 60 rounds to emerge.

CQ.1280 also uses ε-noise. Additionally, initial propensities for the available strategies are lower: 60 instead of 120. This means that a player is less optimistic about the success of strategies in the beginning of the game. In fact, she is more realistic because the payoff structure allows only a maximum reward of 80 points (actor defects while another actor cooperates). In case of defection, the actual reward is even lower due to cooperation costs. These parameter settings result in the quickest and most stable pattern emergence. Stable patterns emerge after about 10 to 15 rounds.

CQ.1280 further shows how the degree of asymmetry affects pattern emergence. While solitary volunteering emerges after 10 rounds in the Asymmetric 2 condition, CQ.1280 requires more than 90 rounds to converge to solitary volunteering in the Asymmetric 1 condition. That is because player 1 shows a long stretch of fluctuation between cooperation and defection (rounds 34 – 91), while the other two players stopped comparatively early in the game: after 34 rounds for player 2, and after 59 rounds for player 3.

Memory representation is the same for all three models. That is, they consider their own actions of the last three rounds, resulting in 8 different action combinations ($2^{A*\Phi} = 2^{1*3} = 8$; for details see Formula 5.4). Thus, optimistic settings for initial propensities (here: 120), require more rounds to lower expectations to realistic values, and thus to settle on a preferred action than initial realistic propensities (here: 60). CQ.1947 even shows that realistic initial propensities require each problem state to be visited only once to settle at preferable actions. That is, the problem state space consists of 512 action combinations for all three players in the previous three rounds. Further, the simulation requires just over 500 rounds to converge at solitary volunteering of the player with the lowest cost to cooperate.

Taken together, the results show that *ClassicQ* model cannot predict turn-taking in the symmetric condition. Solitary volunteering, however, is reliably predicted in the asymmetric VODs. Speed and stability of pattern emergence depend on parameter settings. ε-noise produces the most stable patterns. Realistic initial propensities (here: $\iota = 60$) produce quicker convergence than optimistic initial propensities (here: $\iota = 120$).

6.2.2 CoordinateX

CX.1983 shows good performance on different levels. It has a good overall model fit and pattern ratios for all types of VODs (see Fig. 6.1). In contrast to *ClassicQ* model instances, this also holds for symmetric VODs. Fig. 6.3 shows that *CoordinateX* instances predict very stable turn-taking beginning after a minimum of 25 rounds. These patterns are consistently optimal in the sense that only one actor cooperates at a time.

However, sometimes even more complex patterns than in the original experiment emerge. For example, the first interaction pattern of CX.1983 shows a stable sequence of cooperation for players: 12131213... This complies formally with turn-taking between two and turn-taking between three players at the same time. This pattern is only possible, because players expect a maximum amount of four players in the group ($X = 4$). As a result, the most promising strategy for players 2 and 3 is to defect three times before coordinating once (for details on strategy creation, see section 5.3.2.2).

Just like *ClassicQ*, the *CoordinateX* model allows different parameter combinations. In the following I discuss effects of specific combinations on model fit, and explain why certain parameter setting lead to specific results.

The effect of lowering the expected amount of players to the actual group size ($X = 3$) is depicted by the patterns for CX.2807. Both simulations end up with strict turn-taking between three players. More complex patterns do not emerge anymore.

The first interaction pattern of CX.2807 shows another remarkable feature. After 20 rounds, player 1 stopped to cooperate. This lead, after 42 rounds, to successful coordination and thus turn-taking between players 2 and 3. However, this did not prove to be successful on the long run and player 2 dropped out after round 87. Then, after only nine rounds and starting from round 96, the actors managed to coordinate and converge towards turn-taking between all three. This ultimately resulted in evenly distributed rewards and cooperation costs for all players.

The lower three interaction patterns of Fig. 6.3 show exemplary simulations for Asymmetric 2. Asymmetric 1 is omitted, due to strong similarities of the results. Patterns for CX.2807 are also omitted, as they don't add further insights. Solitary volunteering by the strong player was also dominant for CX.2807. Red lines mark the player with the lowest cooperation costs.

As discussed before, CX.1983 showed the best fit and ratio for behavioral patterns of all model instances. However, the emerging patterns are not optimal. Player 1 cooperates consistently beginning from round 16. The other two players however, do not recognize that cooperation on their behalf is not necessary anymore. On the one hand, this leads to a stable

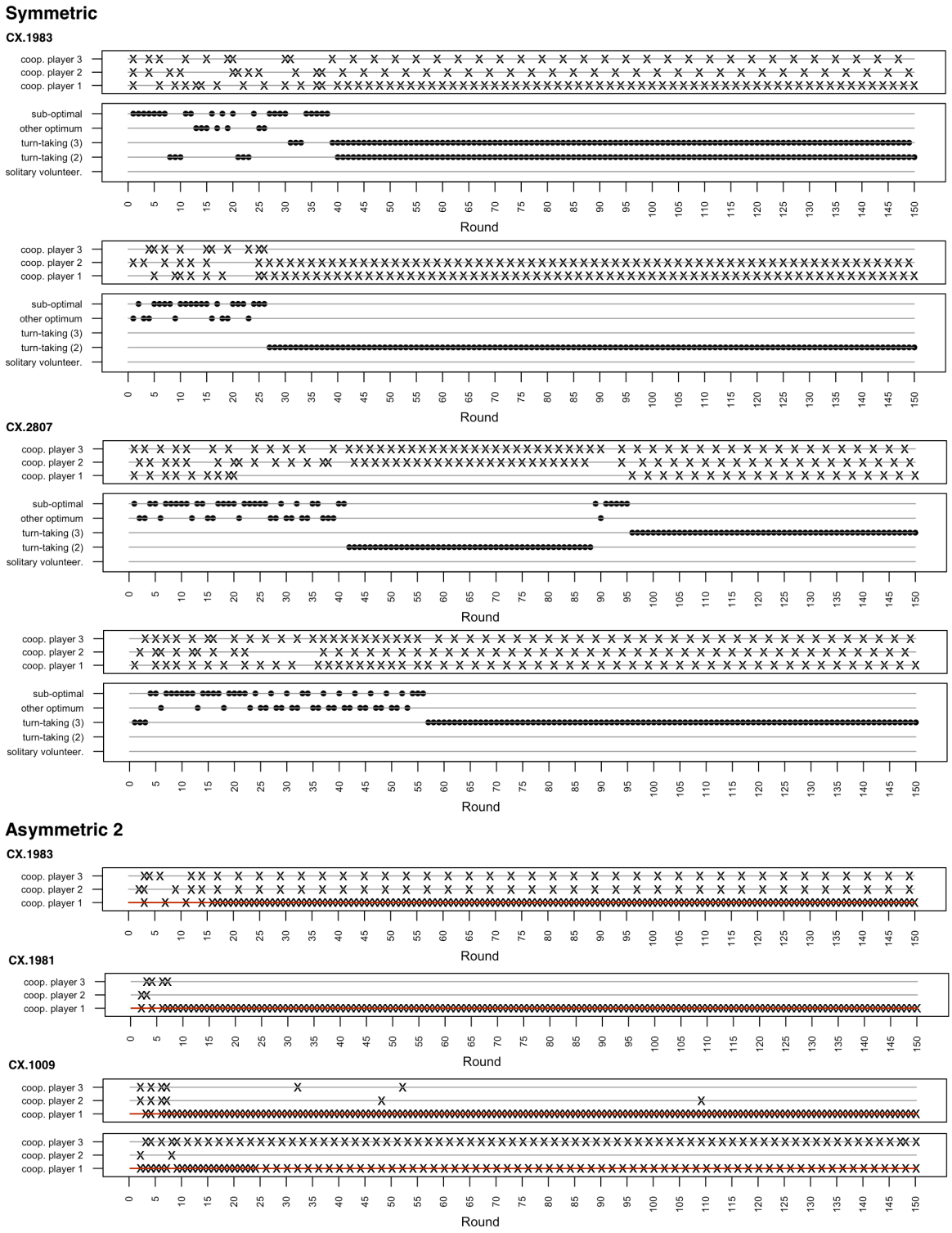

Fig. 6.3 **Exemplary patterns (*CoordinateX*).** Two types of patterns are presented: behavioral patterns and convergence patterns. Behavioral patterns depict cooperation of players in corresponding rounds with an 'x'. Convergence patterns show whether a sequence of actions complies with one of the considered patterns (marked with a dot). Note that a sequence of at least three consecutive actions is required to be recognized as a pattern. For example, 'turn-taking (3)' requires a sequence of cooperation where each player successively cooperates for at least three rounds (e.g., Symmetric, CX.2807, rounds 96–150). For details see section 5.5. Red lines for Asymmetric 2 mark the player with the lowest costs to cooperate.

pattern that complies with the same ratio as in the original experiment. On the other hand, coordination is disrupted every four rounds.

In contrast to CX.1983, CX.1981 has a lower expected group size, namely from $X = 4$ to $X = 2$. Consequently, CX.1981 has a smaller strategy space with only three strategies in total (defect, cooperate, defect-cooperate). This results not only in much quicker coordination (starting from round 8), but also in a pattern of consistent turn-taking between three players.

The effect of ε-greedy on pattern stability is pictured by the comparison of CX.1981 and CX.1009. All other parameters being equal, CX.1009 shows occasional disruptions of coordination, while coordination of CX.1981 is stable after 8 rounds. This confirms the effects of the two balancing strategies (ε-noise and ε-greedy) on pattern stability as already found for *ClassicQ*.

A final notable property of *CoordinateX* is reflected by the second simulation for symmetric CX.1983 and the second simulation for asymmetric CX.1009. In both cases, two players engage in stable turn-taking. Turn-taking between two players sometimes also occurs in symmetric VODs of the original experiment. Further, general turn-taking in asymmetric VODs can also be found in a few experimental games. Therefore, *CoordinateX* can also account for infrequent events of the original experiment and thus meets the principle of *distributional modeling* (Marewski and Mehlhorn, 2011).

Taken together, the results show that *CoordinateX* is the only model class that creates patterns for all conditions that are close to the original data. More precisely, *CoordinateX* predicts turn-taking in the symmetric VOD, and solitary volunteering in the asymmetric VODs. Simple patterns of turn-taking between three players occur consistently when expected group size meets actual group size ($X = 3$). As for *ClassicQ*, fully stable patterns are only predicted by ε-noise. Finally, *CoordinateX* also predicts rare patterns from the original data, like turn-taking between two players in the symmetric and asymmetric VODs.

6.3 Necessary characteristics of learning

Fig. 6.4 provides a direct comparison of model fits for best fitting model instances (plot a.), average fits (plot b.), and the effects of selected parameter settings on model fit (plots c., d.: social preference; plots e., f.: initial propensities). Each plot shows average fit (y-axis) for *Random* (gray), *ClassicQ* (orange) and *CoordinateX* (blue) per condition (x-axis).

As I described earlier, both learning-based models, *ClassicQ* and *CoordinateX*, had a better fit to the human data compared to the *Random* model. This is further illustrated by Fig. 6.4, plot a. However, the average fits for all simulations are comparatively poor (see Fig. 6.4, plot b. and Appendix **??**, Tables **??** and **??**). That is, *ClassicQ* has a bad combined fit with

$RMSE = 37.89$ ($NRMSE = 183.7\%$, $R^2 = 0.14$). *CoordinateX* has only slightly improved combined fit with $RMSE = 30.33$ ($NRMSE = 147\%$, $R^2 = 0.18$). As a reminder, *Random* had a combined fit of $RMSE = 34.2$ ($NRMSE = 165.8\%$, $R^2 = 0.12$). The question therefore arises whether there is any structure in the model fits; or do some particular parameters produce better results than others?

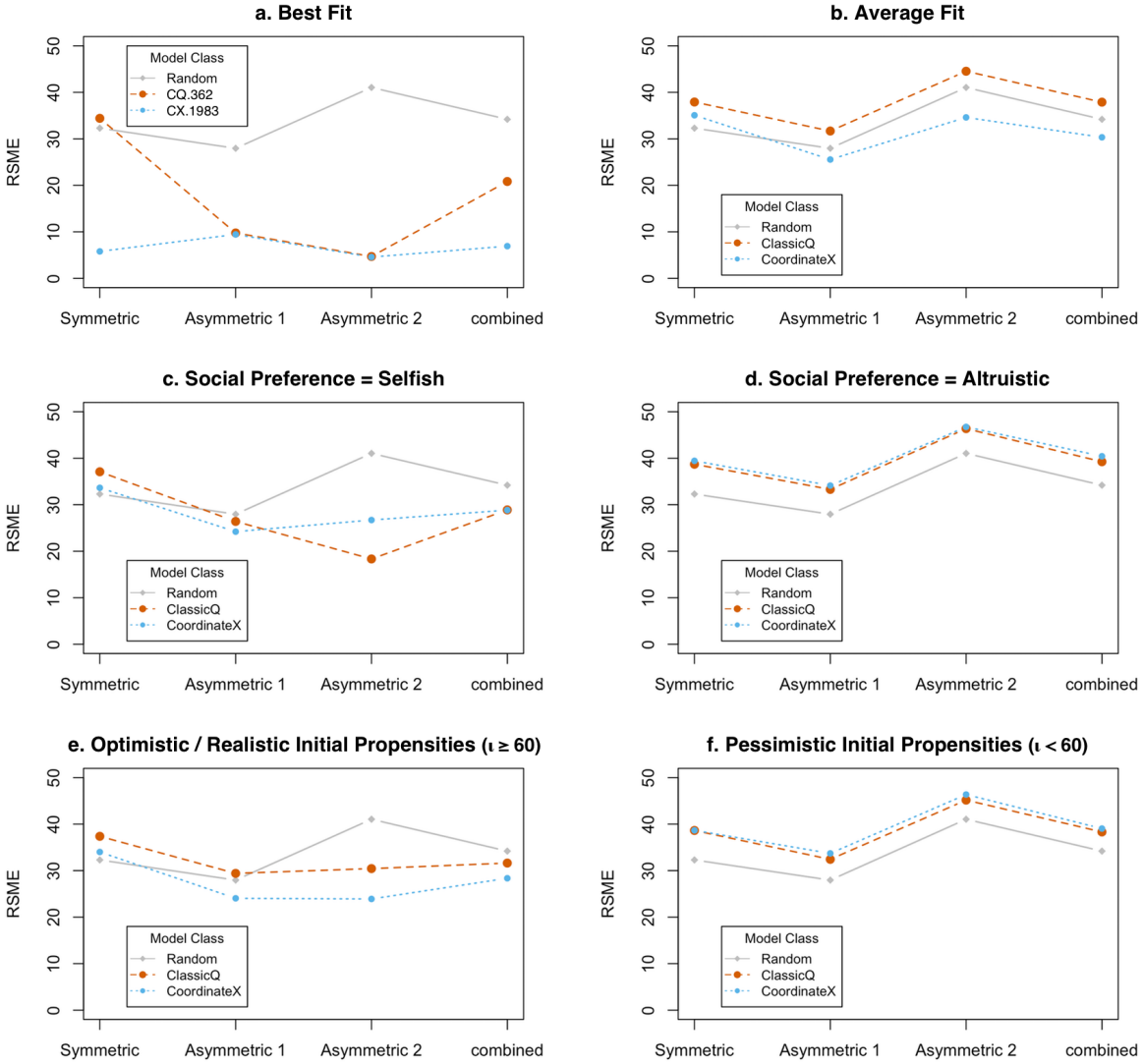

Fig. 6.4 **Effects of parameter settings on model fit.** Each plot displays model fit (y-axis: *RMSE*) for all three model types (gray: *Random*, orange: *ClassicQ*, blue: *CoordinateX*) and each type of VOD (x-axis: Symmetric, Asymmetric 1, Asymmetric 2, combined). Plot a. compares fitting measures for the best performing model instances. Plot b. shows the average fits over all instances per model type. Plot c. and d. compare model fit between the two different settings for social preference: selfish and altruistic. Plot e. and f. compare model fit between optimistic and realistic initial propensities ($\iota \geq 60$) and pessimistic initial propensities ($\iota < 60$).

Of all 9 parameters for *ClassicQ* and 8 parameters for *CoordinateX*, two parameters had an immediate effect on model fit. First, social preference. Altruism lead to worse model fits for all types of VOD compared to the *Random* model (see Fig. 6.4, plot d.). Selfishness on the other hand, showed an increase of model fit, especially for Asymmetric 2 (see Fig. 6.4, plot c.).

The second parameter that had an immediate effect on model fit was initial propensity. Pessimistic settings for initial propensities ($\iota < 60$) also resulted in worse model fits for all VOD types compared to *Random* (see Fig. 6.4, plot f.). Optimistic or realistic settings for initial propensities however, lead to an increase of model fit (see Fig. 6.4, plot e.). Again the biggest effect was for Asymmetric 2.

These results, and further examination of behavioral patterns, showed that altruism and pessimistic initial propensities suppressed any form of coordination. That is, there was no form of turn-taking in the symmetric VOD. Further, there was no form of stable solitary volunteering in the asymmetric VODs.

6.4 Summary of the results

In the previous sections I showed three things. First, I showed that models based on reinforcement learning predict the emergence of behavioral patterns. I showed that both models of learning successfully predict solitary volunteering in the asymmetric VODs. That is, reinforcement of action-state pairs by the *ClassicQ* model on the one hand. And the use of future-oriented strategies (i.e. planning a sequence of actions ahead with subsequent evaluation of success) by the *CoordinateX* model on the other hand.

However, turn-taking in the symmetric VOD can be only predicted by planning ahead (*CoordinateX*). This form of learning produces a very good fit, and ratios of behavioral patterns that are close to the empirical data. It also predicts patterns that occur on a very limited basis in the empirical data. These are turn-taking between two actors in the symmetric VOD and general turn-taking in the asymmetric VOD. *CoordinateX* is therefore the most likely model to capture the cognitive mechanisms of latent norm emergence in the VOD.

Second, I showed that certain characteristics of learning have a significant influence on speed and stability of pattern emergence. I showed that ε-noise produces very stable patterns as found in the empirical data, while ε-greedy shows occasional disruptions. Further, realistic initial propensities for certain actions or strategies, meaning propensities close to actually received rewards, lead to quick coordination. This effect was confirmed by both learning-based model. Furthermore, *CoordinateX* showed very stable turn-taking between three players, when the expected group size matches the actual group size (here: three actors, thus $X = 3$).

Third and finally, I showed that certain characteristics of learning are necessary for patterns to emerge. I showed that two parameters are necessary for coordination. These are selfishness and non-pessimistic initial propensities for actions or strategies. In fact, altruism and pessimistic initial propensities suppressed any form of coordination for both learning-based models and all types of VOD.

An overview of these results is also presented in Table 6.2.

	Random	ClassicQ	CoordinateX
I. Complexity			
Amount of parameters	1	10	8
II. Goodness of fit of the best fitting models			
RMSE	34.2	20.83	6.92
NRMSE	165.8	101	33.5
R^2	0.12	0.48	0.94
III. Coordination in symmetric VOD			
Existent	✗	✗	✓
Emerging patterns	–	–	turn-taking (3 and 2 players)
Minimum of rounds required	–	–	27
IV. Coordination in asymmetric VODs			
Existent	✗	✓	✓
Emerging patterns	–	solitary volunteering	solitary volunteering, turn-taking
Minimum of rounds required	–	13	8
V. Effects of parameter settings			
V.a. Balancing between exploration and exploitation			
ε-greedy, ε-decreasing	–	disruptions in patterns	disruptions in patterns
ε-noise, ε-noise-decreasing	–	stable patterns	stable patterns
V.b. Initial propensities			
Pessimistic ($\iota \ll 60$)	–	no coordination	no coordination
Realistic ($\iota \approx 60$)	–	quick coordination	quick coordination
Optimistic ($\iota \gg 60$)	–	slow coordination	slow coordination
V.c. Social preference			
Selfish	–	coordination possible	coordination possible
Altruistic	–	no coordination	no coordination

Table 6.2 **Summary of results.** Columns contain the different model types. Rows show the most important findings, as discussed in the previous section. Fitting measures are displayed for the best performing model instances (*ClassicQ*: CQ.362, *CoordinateX*: CX.1983).

7. General discussion

7.1 Summary of the study

In the present study, I investigated whether reinforcement learning (Sutton and Barto, 1998) can explain the emergence of social norms in the Volunteer's Dilemma (*VOD*). A sociological study on the three player repeated VOD served as the basis for my work: Diekmann and Przepiorka (2016) showed that the reward structures of the game have an immediate effect on manifestation of behavioral patterns, precursors of social norms (Opp, 2004; Wrong, 1994). That is turn-taking for symmetric VODs (when rewards are distributed equally among the group), and solitary volunteering for asymmetric VODs (when one player has lower costs for cooperation than the others). It is, however, unknown whether and how cognitive mechanisms contribute to the emergence of social norms. Using two classes of computational cognitive models based on reinforcement learning, I investigated whether coordination in the Volunteer's Dilemma can be explained with cognitive mechanisms of the individuals.

Results showed that the future-oriented model (*CoordinateX*), which compared a set of pre-defined action sequences, predicted the results of the empirical data very closely (see sections 6.1, and 6.2.2). The memory-based model (*ClassicQ*), which makes decisions based on reinforcement of action-state pairs, predicted only solitary volunteering in asymmetric VODs (see sections 6.1, and 6.2.1). This makes *CoordinateX* the most likely model to capture the cognitive mechanisms of latent norm emergence in the VOD.

Further, the data showed that different parameter settings had an immediate effect on the the emergence of behavioral patterns (see section 6.3). Altruistic behavior, that is maximization of collective rather than personal rewards, suppressed the emergence of patterns completely. Realistic initial propensities, that is a Q-Table[1] with initial values close to actual rewards, lead to quick pattern emergence.

[1] For an exemplary Q-Table, see Table 3.1.

7.2 Learning as a key cognitive mechanism for norm emergence

The *CoordinateX* model predicted quick, stable and reliable patterns across all conditions, consistent with the results from the empirical data (see Fig. 6.3). That is, it predicted turn-taking between three players in the symmetric VOD and solitary volunteering in the asymmetric VODs. Both learning-based model classes, *CoordinateX* and *ClassicQ*, further showed that the degree of asymmetry in reward structures has an immediate effect on the likelihood to solitarily volunteer for the actor with the lowest costs. The study therefore shows that complex outward behavior can emerge from simple underlying mechanisms (Anderson, 2002; Pfeifer and Scheier, 2001; Simon, 1969). On this account, I propose learning as the key cognitive mechanism in the emergence of behavioral patterns, and thus a precursor of social norms (Wrong, 1994).

One of the most striking results was that both learning-based models predicted solitary volunteering for the asymmetric VODs, while only *CoordinateX* successfully predicted turn-taking in the symmetric VOD. In the following, I first describe that coordination in the asymmetric conditions is a necessary consequence of asymmetry itself. In the course of this, I also describe how the degree of asymmetry in the VOD affects the ratio of solitary volunteering. In the second step, I explain how future-oriented strategies facilitate turn-taking in the symmetric VOD.

7.2.1 Solitary volunteering as necessary consequence of asymmetric VODs

At the system level, that is considering the collective rewards of all agents, solitary volunteering by the strong player is clearly optimal in the asymmetric conditions. Any other outcome creates more costs for the group (i.e. is Pareto inferior). An asymmetric condition therefore provides a focal point, as there is only one single system optimum. Logical inference, or top-down analysis, allows to find the focal point. The reinforcement learning models, however, do not use top-down analysis. They do not use any form logical inference. Further, the focal point in form of the system optimum does not necessarily mean that it is the agents' optimum. Nevertheless, reinforcement learning leads to reliable and stable solitary volunteering in the asymmetric conditions. But why is that? In the following, I illustrate certain aspects with the aid of *ClassicQ* examples. The same reasoning, however, also applies to *CoordinateX*.

First, let's recall how learning and decision-making works in general: At the beginning of a round each actor chooses an action that gives the most expected reward. Once all actors

have performed the selected action, they receive feedback in form of actually received reward. This information is then used to reinforce actions that meet or exceed expected reward and demote actions that fall short of expected reward. Consequently, the actual received reward and the expected reward are crucial factors in the decision-making process, as they provide the yardstick by which success is being measured. Technically this is reflected in the propensity update function (see Formula 3.1), with (r_{t+1}) being received reward, and $(\max_a Q(s_{t+1};a))$ being expected reward. This function holds the key to why solitary volunteering is a dominant solution in the asymmetric conditions.

Consider a selfish player with optimistic initial propensities (e.g., CQ.1442 with $\iota = 120$). First, a selfish player maximizes personal rewards (for details, see sections 5.3.2.1, and 5.3.2.2). Thus, she expects to receive $80-K$ points in case of cooperation, with K being the personal costs. Second, asymmetry results in lower cooperation costs for exactly one player, the *strong player*. As a result, the strong player expects and receives a higher reward for cooperation than the other two players. All other parameters being equal, the propensity update function computes higher results when actual and expected rewards increase. As the strong player has the highest values for rewards, the propensity for cooperation levels off at the highest value for the strong player in a repeated game. Thus, rewards have an immediate effect on propensities: the higher the reward for an action, the higher the potential propensity for that action.

To illustrate this, consider the early stages of a game. No form of coordination has emerged yet. Sub-optimal outcomes occur frequently. As a result, initially high propensities lower constantly for all actions (cooperation and defection) within all players. At some point a weak player cooperates. She still has relatively high propensities for cooperation. But due to her high cooperation costs, she receives only a small reward. Thus, propensity for cooperation decreases, as the received value is below current propensity. The next round the strong player cooperates. She still has a relatively high propensity for cooperation, as well. Due to her low cooperation costs, however, her current propensity is met by the actual reward and cooperation gets reinforced. Consequently, cooperation for the strong player gets reinforced in a stage of the game when cooperation for the weak players gets demoted. Over time, this imbalance results in repeated cooperation by the strong player. Defection can now be reinforced repeatedly for the weak players, and solitary volunteering has emerged.

This example also provides insights on how the degree of asymmetry affects the ratio of solitary volunteering, an effect also found by Diekmann and Przepiorka (2016). CQ.1280, for example, predicts stable solitary volunteering in both asymmetric conditions (see Fig. 6.2). In Asymmetric 2 solitary volunteering is stable after 7 rounds. However, more than 90 rounds are required in Asymmetric 1. Note that the only difference between the two

simulations is the cooperation costs of the strong player. In Asymmetric 2 the strong player gets 70 points for cooperation; in Asymmetric 1 only 50. In both cases defection results in 80 points, given another player cooperates. All other conditions are the same (i.e. model type and parameter settings). Thus, the different results depend solely on the degree of asymmetry in the two asymmetric conditions.

From the strong player's perspective asymmetry has the following effect: When rewards for cooperation and defection are close to another (i.e. Asymmetric 2: 70 vs. 80 points), convergence towards cooperation happens fast. When rewards for cooperation and defection are further apart (i.e. Asymmetric 1: 50 vs. 80 points), convergence requires more rounds. This is because cooperation comes with a guaranteed reward. Defection, however, requires another player to cooperate. Therefore, defection is a very unreliable source of utility, which results in a lower average reward for defection than the potential 80 points. It follows that the closer the reward for cooperation is to the reward for defection, the quicker it can outperform unreliable defection.

7.2.2 Turn-taking in symmetric VODs

As shown in the previous section, asymmetry allows reinforcement learning models to converge at different propensity levels for cooperation, depending on cooperation costs. In the symmetric VOD, however, all actors have the same cooperation costs. Any behavioral pattern of a single cooperating player creates a system optimum. That means there is an almost innumerable amount of system optima in repeated symmetric VODs. This erases different propensity levels and solitary volunteering disappears. As a result, *ClassicQ* is not able to achieve any form of system optimum in the given time. That is, no form of coordination between the players is achieved, and sub-optimal action combinations occur frequently (see Fig. 6.2). In contrast, *CoordinateX* models coordinate in the given time. Players achieve optimal solutions where only a single player cooperates at a time. In some simulations players take turns in twos, in most simulations players take turns in threes. Obviously, *ClassicQ* lacks a crucial property in order to coordinate – a property that exists in *CoordinateX*. But what might that property be?

Previous research with memory-based models (e.g., Juvina et al., 2015; Martin et al., 2014; Stevens et al., 2016) showed that the amount of information players have on the options and rewards of their co-players has a direct effect on the likelihood of coordination. In other words, successful coordination requires a player to select actions in accordance with the behavior of others. *ClassicQ* however, a model based on action-state pairs, never achieves any form of coordination in the symmetric VOD no matter how much information is integrated into the decision-making process. Consider CQ.1947, an instance of *ClassicQ* that integrates

the actions of all players in the previous rounds into the decision-making process ($\Phi = 3$, $A = 3$; see Table 6.1). As a result, CQ.1947 has a big problem state space with 512 states (see Formula 5.4). However, CQ.1947 never achieves any form of coordination in the symmetric VOD in the given time (see Fig. 6.2). *CoordinateX* models, on the other hand, have small problem state spaces. Consider CX.1983 which expects to be in a group of four ($X = 4$; see Table 6.1), and therefore has a problem state space of 5 states (see Formula 5.5). CX.1983 however achieves coordination in form of turn-taking. Thus, pure amount of information is not a crucial factor for coordination in the VOD.

The difference lies in the different mechanisms for learning and decision-making. Let's recall how the two models learn and make decisions. First, *ClassicQ* (see also section 5.3.2.1): *ClassicQ* reinforces action-state pairs. A state is defined by a sequence of previously performed actions. A player selects the action with the highest propensity according to her current state. Propensities are affected by rewards (see Formula 3.1). Thus, a player chooses the action that produced the highest rewards in the same previously experienced situation. To illustrate this, consider a player who determines that she has defected in the previous two rounds. She then decides to cooperate, because from experience she knows that cooperation was usually the best action to perform after she defected twice.

Second, *CoordinateX* (see also section 5.3.2.2): *CoordinateX* reinforces future-oriented strategies, inspired by the concept of latent norms. *CoordinateX* players choose the strategy that produced the highest rewards in the past. A strategy consists of a sequence of actions a player performs in the upcoming rounds, considering what the player expect her co-players do. To illustrate this, consider a player who plans to defect twice (because she expects any of her co-players to cooperate) and then to cooperate once (because she expects nobody else to volunteer in the third round), because this usually produced the best outcome.

The main difference between the two models is that *CoordinateX* selects from predefined action sequences, while *ClassicQ* selects from single actions. The advantage of action sequences over single actions is that sequences add structure. That is, they incorporate the potential for effortless and consistent repetitions. For example, a *CoordinateX* player plans to defect twice and cooperate once (*DDC*). In case this sequence meets the propensities, the strategy is replayed and creates a complex six move pattern with only two decisions. *ClassicQ* players however, can only plan for the upcoming round. A six move pattern requires six decisions for *ClassicQ* models. Thus, the *CoordinateX* model reduces complexity, as there are only a few action sequences available, that have the potential to create complex behavior. Thus, *CoordinateX* has a structural advantage over *ClassicQ*.

But why is turn-taking the dominant pattern in the symmetric VOD? The answer is that the structural advantage in the *CoordinateX* model comes along with a quantifiable

advantage for propensities. Consider, a player who believes to be in a group of three ($X = 3$, for details see section 5.3.2.2). The set of available strategies is: $\{D, C, DC, DDC\}$. For *DDC* the player's expected (and potential) reward per round is: $80 + 80 + (80 - 50)/3 = 63.3$ points. In contrast, *DC* gives only a potential reward of $80 + (80 - 50)/2 = 55$ points. As discussed in the previous section, rewards have an immediate effect on propensities. Thus, the propensity for *DDC* levels off at a higher value than for *DC* and *C*. This causes that *DDC* is considered more often than *DC*, and *C*. Note that the reward for *DDC* is only a potential reward, as it requires coordination with the other players. *D* has a higher potential reward (80 points vs. 63.3 points), but it is a single action strategy as in *ClassicQ*. The previous paragraph shows that coordination for single actions has not been established. Thus, actual reward for *D* drops below *DDC* and turn-taking has emerged.

However, turn-taking could also occur in many different ways. For example, players could alternate by volunteering for three rounds in a row each. Complex patterns like this, however, do not show in the empirical data (see Diekmann and Przepiorka, 2016). This is in line with observations in previous studies, which also showed a predominance of simple patterns (e.g., Goldstone et al., 2015; Helbing et al., 2005; Juvina et al., 2015). In fact, complex patterns would require a higher cognitive load within each individual: In order to cooperate three times in a row and to start doing so in seven rounds from now, a player needs to keep track of more information than a player who simply cooperates once every three rounds. Humans choose feasible strategies based on constraints given by the task and their own cognitive architecture (Howes et al., 2009). Simple turn-taking provides a profitable strategy (as in *CoordinateX*), while minimizing cognitive load (small problem state space). Simple turn-taking therefore presents a feasible strategy. It provides a simple solution for the task with few demands on the cognitive architecture. More complex patterns are less feasible, as they merely increase memory, but do not necessarily change the overall outcome of the task. This is also in line with the literature on strategic game playing, claiming that humans only use a minimal amount of information to form cognitive strategies (Camerer, 2003; Colman, 2003; Juvina et al., 2015).

Additionally, predictions of models with lower demands on memory are closer to the empirical data. This becomes evident from three different aspects of the simulations. First, *ClassicQ* instances, that consider more information in the decision-making process require more rounds to coordinate in the same way like players that consider a minimal amount of information. CQ.1947, for example, keeps track of the previous three actions of all players ($\Phi = 3$, see Table 6.1). CQ.1280 however, considers only her own previous three actions to make a decision ($\Phi = 1$, see Table 6.1). This results in a problem space of $2^{3*3} = 512$ states for CQ.1947, and $2^{1*3} = 8$ states for CQ.1280 (see Formula 5.4). As a result, solitary

volunteering requires more than 500 rounds for CQ.1947 to emerge, while CQ.1280 shows the same pattern after only 7 rounds.

Second, *CoordinateX* players, who create their strategies according to the actual group size ($X = 3$, see section 5.3.2.2), do not show more complex patterns than simple turn-taking. In fact, more complex patterns emerge only when expected group size exceeds actual group size. CX.1983 ($X = 4$), for example predicts a pattern where player 1 cooperates every second round, while players 2 and 3 cooperate every fourth round (see Fig. 6.3).

Third and finally, altruistic players keep track of the utilities of all players, in order to maximize group rewards. Selfish players, however, keep track only of their own utilities in order to maximize personal rewards. As a result, only selfish players coordinate successfully with their co-players. Interestingly, pre-dominance of myopic behavior can also be found in other contexts of social interaction, such as network formations (e.g., Van Dolder and Buskens, 2014).

7.3 Implications

In the current study I developed computational cognitive models to investigate whether reinforcement learning models can explain the emergence of social norms in the Volunteer's Dilemma. These have practical and theoretical value for a wide range of disciplines. With regard to cognitive psychology, I showed that simple mechanisms can help to understand complex behavior (Pfeifer and Scheier, 2001; Simon, 1969). Specifically, I showed that research at the cognitive band (milliseconds, seconds) can inform about phenomena on the social band (days, weeks, months) (Anderson, 2002). More precisely, I showed that reinforcement learning (cognitive band) explains the emergence of behavioral patterns, precursors of social norms (social band). Further, I provided a formal description of a cognitive mechanism in the context of social interaction. Finally, I showed that strategies adjusted to the problem at hand (*CoordinateX*), provide a structural and quantifiable advantage for decision-making processes, that allow to reduce cognitive demands and increase the ability to coordinate in a social scenario.

With regard to sociology, my study provides a new perspective on norm emergence. That is, rather than looking at group-level aspects (e.g., Diekmann and Przepiorka, 2016), my research focused on the individuals involved in norm emergence. I showed that expectations towards the actions of others (i.e. future-oriented strategies in the *CoordinateX* model) are necessary for successful tacit coordination. Additionally, I provided predictions for human behavior. Specifically, the model predicts that the stronger the disparity in the asymmetric conditions, the more rounds are required to coordinate. Finally, the models also predicted that

altruistic players fail to coordinate entirely. These results can help to establish a bottom-up theory on norm emergence, as started by Diekmann and Przepiorka (2016).

With regard to (behavioral) game theory, my study provides a formal description of a human solution concept in strategic game-playing. I showed that humans learn from experience and integrate knowledge as presented during the course of the game. That is a different approach to top-down analyses, such as backwards induction, as typically proposed by classical game theory (Camerer, 2003; Colman, 2003). Further, the way humans make decisions in context of the VOD applies to utility maximization within the bounds of the given scenario. Finally, the scenario itself, the Volunteer's Dilemma, is a three-player game. That is particularly interesting for game theoretic research, as games with more than two players are clearly under-represented in scientific studies (e.g., Helbing et al., 2005).

With regard to applied artificial intelligence, my study provides a mechanistic and self-learning model for human behavior in the domain of social interaction. Especially in the design of multi-agent systems and multi-agent learning, game theory forms the basis for autonomous rational behavior within individual agents (e.g., Peters, 2015; Shoham et al., 2009). Norms are further used to achieve coordination in artificial agent societies (Wooldridge, 2009, p. 173 ff.). A formal description of the emergence of social norms based on a game-theoretic setting can therefore be instrumental to create coordinative behavior within societies of artificial agents. This can support an easy deployment of artificial agents into human societies. That is because such agents would allow for intuitive interaction with human counterparts without the need to get acquainted with unfamiliar behavior.

7.4 Limitations and future work

As with all model-based research, there are a few limitations. As described in section 3.1.2, model-based research is under the influence of personal interpretations, experiences and preferences. Thus, no two model-based studies on the same issue are the same. This also holds for the current study. Juvina et al. (2015) for example, introduced the concept of trust to account for the actions of other players. In my study however, I use future-oriented strategies to integrate expectations towards the actions of co-players.

Aside from the general limitations of model-based research, there are also some specific limitations on my study. First, my simulations currently assume all players to possess equal characteristics. That is, a comparison of different initial parameter settings and thus different a priori characteristics were left out. Other research does account for a direct comparison of personal differences, such as myopic vs. altruistic behavior (e.g., Bogaert et al., 2008). This approach I deliberately omitted to avoid interferences or side-effects due to cross-comparison.

My models however, can easily simulate such a priori differences, by simply initializing players in the same VOD with different parameter settings.

A second shortcoming of the study is that the representation of altruism might be too simple. In my study altruistic players merely maximize group rewards, while selfish players maximize personal rewards. This black and white representation of social preference is very unlikely in real life and a better differentiation might lead to different results. As an example, Van Dolder and Buskens (2014) represent social preference with a formula that incorporates weights for own rewards, rewards of others and the preference for equality. This could be easily integrated into my current models and would allow a more differentiated view on social preference.

Third, the *ClassicQ* model might require more time to coordinate. However, 150 rounds for a problem state space size of 9 (CQ.1 – CQ.1944), and 5.000 rounds for a problem state space size of 512 (CQ.1945 – CQ.1948) provide plenty of opportunity to explore all available states multiple times.

Fourth, to further generalize my results additional steps are required. That is, predictions I made (disparity of asymmetry affects time to coordinate, altruistic players fail to coordinate) need to be tested in sufficient follow-up experiments. This step is also described by Marewski and Mehlhorn (2011) as the principle of *predictive modeling*.

One opportunity for future work I would like to point out relates to the mental representation of rewards. As described by Camerer (2003) and Colman (2003) this is also one of the central issues in behavioral game theory. Currently, my models use rewards as presented by the game. However, mental representation of rewards might be different than actual rewards given by the VOD. For example, in case of no cooperation between players, 0 points are rewarded. This translates into *no gain* for that round. For a human player, however, a situation of dysfunctional coordination might feel like a loss, synonymous with a negative reward. This could be easily tested with minor adjustments in the current model implementation. The idea of different mental representations of rewards goes in line with Janssen and Gray (2012). They argue that rewards can take many different forms within the same task, based, for example, on either accuracy or speed of task performance.

7.5 Conclusion

My results show that learning is the key cognitive mechanism in the emergence of latent norms. In particular, future-oriented strategies based on expectations towards the actions of others provide a structural and quantifiable advantage in decision-making processes, that facilitate quick coordination with minimum cognitive load.

References

Allison, S. T. and Kerr, N. L. (1994). Group correspondence biases and the provision of public goods. *Journal of Personality and Social Psychology*, 66(4):688–698.

Anderson, J. R. (2002). Spanning seven orders of magnitude: A challenge for cognitive modeling. *Cognitive Science*, 26(1):85–112.

Anderson, J. R. (2009). *How can the human mind occur in the physical universe?* Oxford University Press.

Bogaert, S., Boone, C., and Declerck, C. (2008). Social value orientation and cooperation in social dilemmas: A review and conceptual model. *British Journal of Social Psychology*, 47(3):453–480.

Camerer, C. F. (2003). Behavioural studies of strategic thinking in games. *Trends in cognitive sciences*, 7(5):225–231.

Colman, A. M. (2003). Cooperation, psychological game theory, and limitations of rationality in social interaction. *Behavioral and brain sciences*, 26(02):139–153.

Dayan, P. and Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, 8(4):429–453.

Diekmann, A. (1985). Volunteer's dilemma. *Journal of Conflict Resolution*, 29(4):605–610.

Diekmann, A. (1993). Cooperation in an asymmetric volunteer's dilemma game theory and experimental evidence. *International Journal of Game Theory*, 22(1):75–85.

Diekmann, A. and Przepiorka, W. (2016). "take one for the team!": Individual heterogeneity and the emergence of latent norms in a volunteer's dilemma. *Social Forces*, 94(3):1309–1333.

Fu, W.-T. and Anderson, J. R. (2006). From recurrent choice to skill learning: A reinforcement-learning model. *Journal of experimental psychology: General*, 135(2):184.

Gigerenzer, G. (2007). *Gut feelings: The intelligence of the unconscious*. Penguin.

Goldstone, R. L., de Leeuw, J. R., and Landy, D. H. (2015). Fitting perception in and to cognition. *Cognition*, 135:24–29.

Gray, W. D., Sims, C. R., Fu, W.-T., and Schoelles, M. J. (2006). The soft constraints hypothesis: a rational analysis approach to resource allocation for interactive behavior. *Psychological review*, 113(3):461.

Helbing, D., Schönhof, M., Stark, H.-U., and Hołyst, J. A. (2005). How individuals learn to take turns: Emergence of alternating cooperation in a congestion game and the prisoner's dilemma. *Advances in Complex Systems*, 8(01):87–116.

Howes, A., Lewis, R. L., and Vera, A. (2009). Rational adaptation under task and processing constraints: implications for testing theories of cognition and action. *Psychological review*, 116(4):717.

Janssen, C. P. and Brumby, D. P. (2015). Strategic adaptation to task characteristics, incentives, and individual differences in dual-tasking. *PloS one*, 10(7):e0130009.

Janssen, C. P., Brumby, D. P., and Garnett, R. (2012). Natural break points: the influence of priorities and cognitive and motor cues on dual-task interleaving. *Journal of Cognitive Engineering and Decision Making*, 6(1):5–29.

Janssen, C. P. and Gray, W. D. (2012). When, what, and how much to reward in reinforcement learning-based models of cognition. *Cognitive Science*, 36(2):333–358.

Juvina, I., Lebiere, C., and Gonzalez, C. (2015). Modeling trust dynamics in strategic interaction. *Journal of Applied Research in Memory and Cognition*, 4(3):197–211.

Kangasrääsiö, A., Athukorala, K., Howes, A., Corander, J., Kaski, S., and Oulasvirta, A. (2017). Inferring cognitive models from data using approximate bayesian computation. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 1295–1306. ACM.

Kollar, T., Tellex, S., Roy, D., and Roy, N. (2010). Toward understanding natural language directions. In *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on*, pages 259–266. IEEE.

Lewandowsky, S. (1993). The rewards and hazards of computer simulations. *Psychological science*, 4(4):236–243.

Lewis, D. (1969). *Convention: A Philosophical Study*. Harvard University Press.

Marewski, J. N. and Mehlhorn, K. (2011). Using the act-r architecture to specify 39 quantitative process models of decision making. *Judgment and Decision Making*, 6(6):439.

Martin, J. M., Gonzalez, C., Juvina, I., and Lebiere, C. (2014). A description–experience gap in social interactions: Information about interdependence and its effects on cooperation. *Journal of Behavioral Decision Making*, 27(4):349–362.

Newell, A. (1990). *Unified theories of Cognition*. Cambridge University Press.

Olfati-Saber, R., Fax, J. A., and Murray, R. M. (2007). Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, 95(1):215–233.

Oliver, P., Marwell, G., and Teixeira, R. (1985). A theory of the critical mass. i. interdependence, group heterogeneity, and the production of collective action. *American journal of Sociology*, 91(3):522–556.

Opp, K.-D. (2004). " what is is always becoming what ought to be." how political action generates a participation norm1. *European Sociological Review*, 20(1):13–29.

Osborne, M. J. and Rubinstein, A. (1994). *A course in game theory*. MIT press.

Payne, S. J., Duggan, G. B., and Neth, H. (2007). Discretionary task interleaving: Heuristics for time allocation in cognitive foraging. *JOURNAL OF EXPERIMENTAL PSYCHOLOGY GENERAL*, 136(3):370.

Peters, H. (2015). *Game theory: A Multi-leveled approach*. Springer.

Pfeifer, R. and Scheier, C. (2001). *Understanding intelligence*. MIT press.

Rapoport, A. (1974). Prisoner's dilemma: recollections and observations. *Game theory as a theory of conflict resolution. Dordrecht: Reidel*, pages 17–34.

Roberts, S. and Pashler, H. (2000). How persuasive is a good fit? a comment on theory testing. *Psychological review*, 107(2):358.

Russell, S. J. and Norvig, P. (2009). Artificial intelligence: a modern approach (3rd edition).

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599.

Severinson-Eklundh, K., Green, A., and Hüttenrauch, H. (2003). Social and collaborative aspects of interaction with a service robot. *Robotics and Autonomous systems*, 42(3):223–234.

Shoham, Y., Leyton-Brown, K., et al. (2009). Multiagent systems. *Cambridge Books*.

Shoham, Y. and Tennenholtz, M. (1992). On the synthesis of useful social laws for artificial agent societies (preliminary report). *AAAI*, pages 276–281.

Simon, H. A. (1969). The sciences of the artificial. *Cambridge, MA*.

Stevens, C. A., Taatgen, N. A., and Cnossen, F. (2016). Instance-based models of metacognition in the prisoner's dilemma. *Topics in cognitive science*, 8(1):322–334.

Stocco, A. (2017). A biologically plausible action selection system for cognitive architectures: Implications of basal ganglia anatomy for learning and decision-making models. *Cognitive Science*.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

Thibaut, J. W. and Kelley, H. H. (1978a). Interpersonal relations: A theory of interdependence. *Aufl. New York ua*.

Thibaut, J. W. and Kelley, H. H. (1978b). *The Social Psychology of Groups*. Oxford: John Wiley.

Van Dolder, D. and Buskens, V. (2014). Individual choices in dynamic networks: An experiment on social preferences. *PloS one*, 9(4):e92276.

Voss, T. (2001). *Game-theoretical perspectives on the emergence of social norms*. na.

Walker, A. and Wooldridge, M. (1995). Understanding the emergence of conventions in multi-agent systems. *ICMAS*, 95:384–389.

Walsh, M. M. and Anderson, J. R. (2014). Navigating complex decision spaces: Problems and paradigms in sequential choice. *Psychological bulletin*, 140(2):466.

Watkins, C. J. and Dayan, P. (1992). Q-learning. *Machine learning*, 8(3-4):279–292.

Wooldridge, M. (2009). *An introduction to multiagent systems*. John Wiley & Sons.

Wrong, D. (1994). *Problem of Order*. Simon and Schuster.

Wu, E., Gopalan, N., MacGlashan, J., Tellex, S., and Wong, L. L. (2016). Social feedback for robotic collaboration.

y López, F. L., Luck, M., and d'Inverno, M. (2006). A normative framework for agent-based systems. *Computational & Mathematical Organization Theory*, 12(2-3):227–250.

Zhang, Y. and Hornof, A. J. (2014). Understanding multitasking through parallelized strategy exploration and individualized cognitive modeling. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 3885–3894. ACM.

A. Learning and decision making (*ClassicQ*, ε-noise)

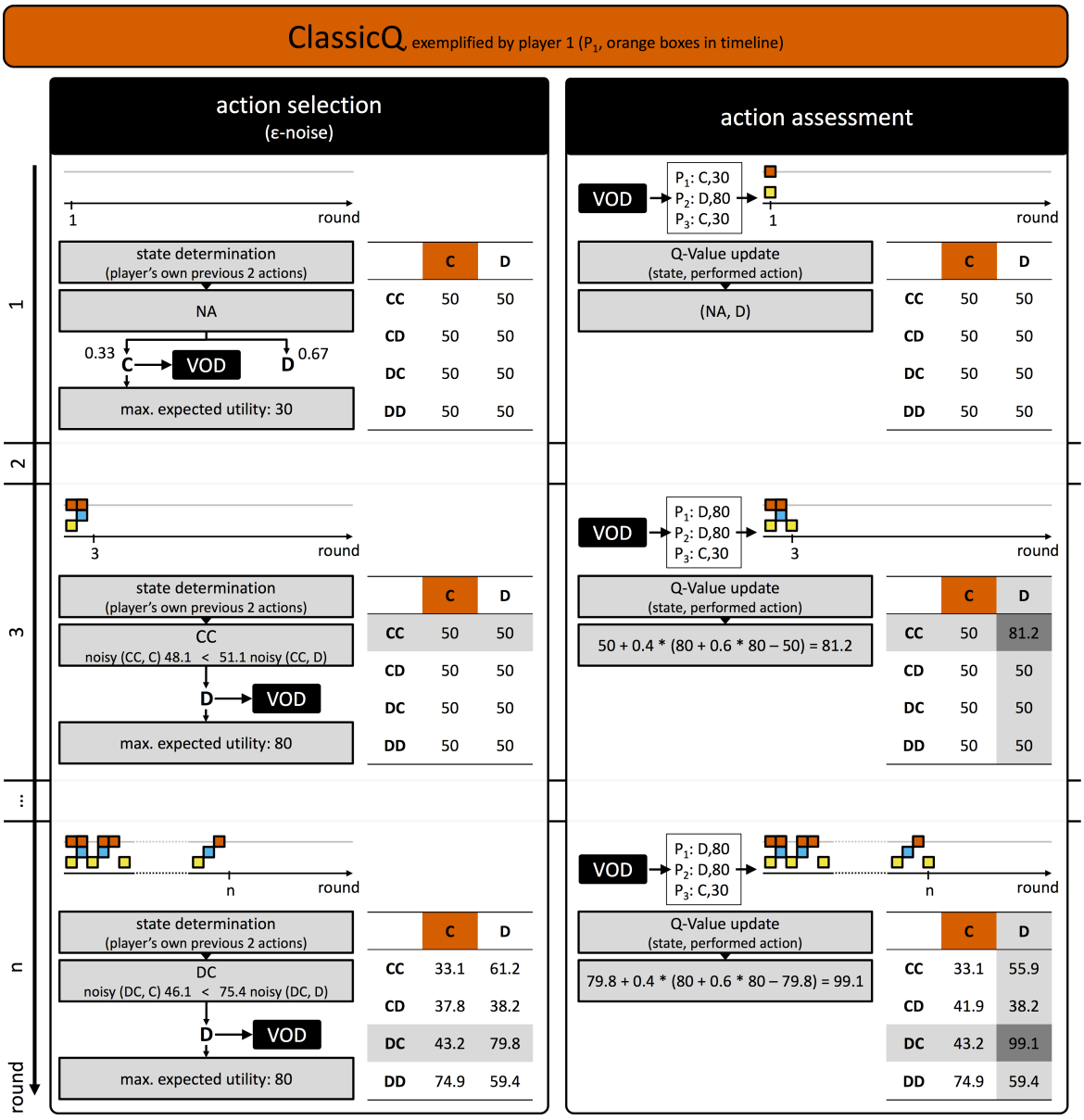

Fig. A.1 **Learning and decision-making in the *ClassicQ* model (ε-noise).**

B. Learning and decision making (*CoordinateX*, ε-noise)

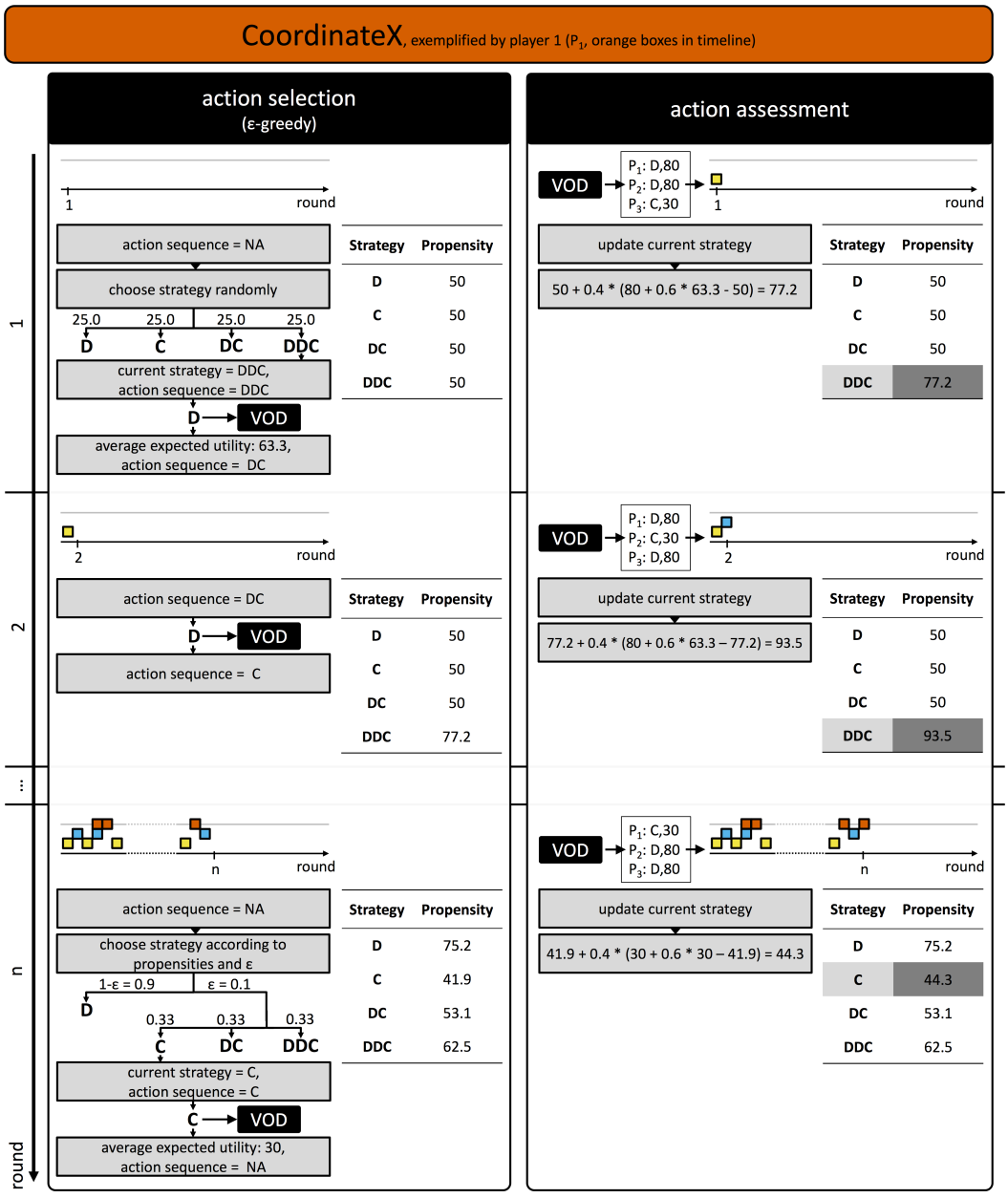

Fig. B.1 **Learning and decision-making in the *CoordinateX* model (ε-greedy).**