Chapter 5
Models of Mechanisms: The Case of the Replicator Dynamics

Till Grüne-Yanoff

Abstract The general replicator dynamics (RD) is a formal equation that is used in biology to represent biological mechanisms and in the social sciences to represent social mechanisms. For either of these purposes, I show that substantial idealisations have to be made – idealisations that differ for the respective disciplines. These create a considerable *idealisation gap* between the biologically interpreted RD and the learning interpretations of the RD. I therefore argue that these interpretations represent different mechanisms, even though they are interpretations of the same formal RD equation. Furthermore, I argue that this idealisation gap between the biological and economic models is too wide for the respective mechanisms to share a common abstract causal structure that could be represented by the general RD model.

1 Introduction

It has become fashionable in recent philosophy of science to explicate the use of scientific models by claiming that they represent mechanisms. In this chapter, I discuss the replicator dynamics (RD), an important model in biology and economics, and argue that it does not represent a mechanism. The argument proceeds in two steps. First, I show that even though the same RD model is employed in biology and economics, the different interpretations in these disciplines make it represent different mechanisms. Second, I argue that these different mechanisms do not instantiate a common, more abstract, mechanism. Rather, different kinds of idealisations are imposed on the RD model, depending on whether it is interpreted in economics or in biology. This opens an 'idealisation gap' between the different

T. Grüne-Yanoff (✉)
Avdelningen för Filosofi, Royal Institute of Technology (KTH),
Teknikringen 78 B, 100 44 Stockholm, Sweden
e-mail: gryne@kth.se

H.-K. Chao et al. (eds.), *Mechanism and Causality in Biology and Economics*,
History, Philosophy and Theory of the Life Sciences 3, DOI 10.1007/978-94-007-2454-9_5,

biological and economic models, too wide for the respective mechanisms to share a common abstract causal structure that could be represented by the general RD model.

The chapter is structured as follows. Section 2 introduces the needed distinctions between mechanism sketches, abstract models and complete models on the one hand and particular mechanisms and abstract mechanisms on the other. Section 3 surveys the formal RD model and its derivation from evolutionary game theory. Section 4 discusses its use by population biologists, who intended it as a representation of biological mechanisms. In Sect. 5, I discuss economists' use of the same RD equation to represent social mechanisms and argue that these social mechanisms are distinct from the biological ones. Section 6 contains the main argument, showing that the biological and economic models are separated by an 'idealisation gap' too wide for the respective mechanisms to share a common abstract causal structure that could be represented by the general RD model. Section 7 concludes.

2 Models and Mechanisms

The notion of mechanism has had significant impacts on the way philosophers of science account for the use of models in the sciences, in particular in biology, economics and the neurosciences. According to these accounts, models explain because they represent the mechanism that produces the phenomenon to be explained (Craver 2006, p. 367). Models help in controlling the real world, because their mechanism representations enable modellers to answer counterfactual questions (Woodward 2002, p. S371). Finally, we can make true claims with models, because they correctly represent an isolated mechanism, even when they idealise the influence of many background condition (Mäki 2009, p. 30).

In each of these functions, models *represent* mechanisms. Whatever the specific definition of mechanism is (I will remain noncommittal here, as different incompatible definitions are extant and the detail of these does not matter for my purposes here), it is clear that mechanism is considered a part of the real world, characterised, for example, as 'material structures' (Craver and Kaiser 2013, p. 130) or a 'portion of the causal structure of the world' (Craver and Kaiser 2013, p. 141).

A mechanistic model may be designed to represent more or less details of a mechanism. Here authors have distinguished between mechanisms sketches, schemata and complete mechanistic models. A *sketch* is an 'incomplete model of a mechanism' (Craver 2006, p. 360). While characterising some parts, activities and features of the mechanism's organisation, it leaves blanks. These blanks are not necessarily visible, as they may be camouflaged by 'filler terms': terms like 'activate', 'inhibit' or 'produce' that indicate activity in a mechanism without detailing how the activity is carried out. Thus, there is more to the represented mechanism than a representing model sketch says.

On the other extreme, we have an *ideally complete* model. 'Such models include all of the entities, properties, activities, and organizational features that are relevant to every aspect of the phenomenon to be explained' (Craver 2006, p. 360). Even if completeness is relativised with respect to explanatory purpose, few, if any, such complete models can be found. More relevant, thus, seems the notion of a mechanism schema, which is a somewhat complete, but less than ideally complete, mechanistic model.

For a given mechanism, a mechanism sketch thus represents less of its features than a mechanism schema does. The sketch does so either by not at all specifying some features that the schema specifies (this is easier with formal models: a set-theoretic model, say, may stay silent about the colours of the objects it represents; a computer model may successfully evade specifying the weight of the structure it represents). Alternatively, sketches often specify certain features, but users of the sketch might exclude these features from the representational function of the sketch. That is, they declare these certain features to be *idealisations*. Scale models, for example, have many features, like size and weight and materiality, that are usually considered idealisations and hence not representations of the target object's properties. By either way, a mechanistic sketch represents less features of a given mechanism than a mechanism scheme does. Consequently, mechanism sketches are more *abstract* than mechanism schemata.

Abstraction is often thought of in relation to generality.[1] A mechanism sketch, then, is more abstract than a mechanism schema, because those properties described in the sketch are a proper subset of those described in the schema. Different mechanisms, described by different schemata, may therefore be described by one and the same sketch.

Such a view of mechanistic models is particularly plausible when seen from an 'exemplar' account of mechanisms. Such an account points out that mechanistic models often represent a particular, exemplary mechanism (Bechtel and Abrahamsen 2005, p. 438). Such exemplars or prototypes are particular tokens of causal structure in the world. A mechanistic model close to being ideally complete might represent just a single such exemplar. A mechanism sketch, on the other hand, might represent a large set of such exemplars. With increasing abstraction, mechanistic models get more and more general.

Scientists use exemplars and prototypes, according to Bechtel and Abrahamsen, in order to accommodate the subtle variations between related mechanisms. For example, they model a mechanism in wild-type *Drosophila* and then extrapolate from this prototype to mechanisms in other strains and species, all the while acknowledging that these are not identical mechanisms. In this view, explaining with a mechanistic model typically commences by explaining the phenomenon with

[1] Take, for example, Nancy Cartwright's Aristotelian account of abstraction: '*A* is a more abstract object than *B* if the essential properties, those in the description of *A*, are the proper subset of the essential properties of *B*' (Cartwright 1989, p. 214).

a prototype mechanism, which is then judged to be sufficiently similar to the mechanism that actually produced the phenomenon.

Yet this is not the only way how one can conceive of inferences between mechanisms. Instead of restricting one's ontology to concrete mechanisms that are only instantiated in one kind of organism, or one kind of social institution, one might also accept that there are *abstract mechanisms* that have many concrete instantiations in different kinds of organisms or institutions. This idea has been floated by some writers, who propose a sort of hierarchy of mechanisms. It is worthwhile quoting one such argument at length.

> Processes identified in the causal reconstruction of a particular case or a class of macro-phenomena can be formulated as statements of mechanisms if their basic causal structure (e.g., a specific category of positive feedback) can also be found in other (classes of) cases. The mobilization process observed in a fund-raising campaign for a specific project can, for instance, be generalized to cover other outcomes such as collective protest or a patriotic movement inducing young men massively to enlist in a war. A particular case of technological innovation like the QWERTY keyboard may similarly be recognized as a case in which an innovation that has initially gained a small competitive advantage crowds out technological alternatives in the long run. This is already a mechanism of a certain generality, but it may be generalized further to the mechanism of "increasing returns," which does not only apply to technological innovations but has also been used in the analysis of institutional stability and change … "Increasing returns," of course, is a subcategory of positive feedback, an even more general mechanism that also operates in the bankruptcy of a firm caused by the erosion of trust or in the escalation of violence in clashes between police and demonstrators. (Mayntz 2004, p. 254)

Central to this idea is that more abstract mechanisms exist in the same way as concrete mechanisms are said to exist. Abstract mechanisms are *instantiated* in more concrete mechanisms: Mayntz' positive feedback mechanism is instantiated in escalation of violence between police and demonstrators, in trust-erosion mechanisms and in increasing returns mechanisms. Mechanisms of different degrees of abstraction are also *nested*: positive feedback is instantiated in increasing returns, which in turn is instantiated in technological crowding out, which in turn is instantiated in the specific process that led to the dominance of the QWERTY keyboard.

According to this view, inferences between mechanisms do not go from prototypes to similar particular mechanisms, but they go through abstract mechanisms in the form of shared 'basic causal structure'. Explaining with a mechanistic model commences by explaining the phenomenon with an abstract mechanism and then showing that the mechanism that actually produced the phenomenon is an instantiation of the abstract mechanism.

Allowing for abstract mechanisms produces an ontological mirror image to the abstraction hierarchy of models. Unlike the exemplar account, which casts all models as more or less abstract representations of particular mechanisms, the abstract mechanism account allows models to represent both abstract and particular mechanisms. Consequently, what appears at first sight to be a mechanism sketch might either be an abstract representation of particular mechanisms or a nearly complete model of an abstract mechanism.

Prima facie, the abstract mechanisms account fits well with observed scientific practice. Scientists often speak about abstract causal structures as if they were real. They see patterns, structures and processes instantiated in various events that produce phenomena: for example, natural selection in the genesis of traits of many different organisms or positive feedback loops yielding dominance of certain set-ups in many institutions. They model these abstract patterns, structures and processes and suggest that these models represent something real.

Conversely, scientists sometimes question the legitimacy of abstract mechanistic models by arguing that an abstract mechanistic model is a mere sketch and not a representation of an abstract mechanism. For example, a paper glider might be a useful mechanism sketch of flight mechanisms in both birds and flying machines. But as parents will explain to their little paper pilots, this does not mean that bird flight and machine flight share the same basic causal structure. Rather, birds combine the function of providing both lift and thrust in their wings, while airplanes separate these functions. Such an explanation implicitly distinguishes between *genuine abstract models* that represent abstract mechanisms and *spurious abstract models* that are mere sketches of concrete mechanisms, to be filled in different and differentiating ways.

This is the problem that evolutionary game theorists face, too: they operate – amongst other formalisms – with the RD model. This model is very abstract: it is used to represent concrete mechanisms that clearly differ in some of their properties. Crucially for my question, the RD model is used to represent mechanisms both in economics and biology. The question thus arises whether the RD model represents the same abstract mechanism in both disciplines or whether it is a mere mechanism sketch that represents a set of disparate concrete mechanisms.

I argue that the RD is a spurious abstract model: a mere mechanism sketch that requires filling in to represent the relevant features of the respective biological and social mechanisms. As I will argue in Sect. 6, this ‘filling in’ of the RD follows discipline-specific paths that increase the idealisation gap between biological and social RD models. But before I can make that argument, I need to investigate the modelling projects in the two disciplines in more detail.

3 The Replicator Dynamics

Evolutionary game theory (EGT) investigates the compositional stability of a population as the result of interaction amongst its members. One of its most prominent modelling approaches derives a differential equation for the population composition from the game matrices that detail payoffs from interaction for each individual in the population. Thus, in contrast to classical game theory, EGT focuses not on decisions of individual players, but on properties of the whole population and on the effect of properties of previous populations on future population. This effect is represented through various population dynamics, first and foremost the replicator dynamics (RD).

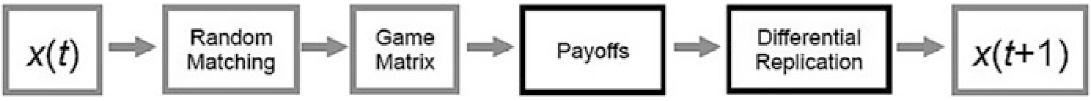

Fig. 5.1 The general RD model

Let me describe the RD model in more detail. A *population* is a set of individuals. Individuals are programmed to play one strategy. A *strategy* is a complete plan of action for whatever situation might arise; this fully determines the player's behaviour. A *population state* is defined as the vector $x(t) = (x_1(t),\ldots, x_k(t))$, where each component $x_i(t)$ is the frequency of strategy i in the population at time t.[2] The replicator dynamics is a function that maps a population state at time t onto a population state at $t + 1$. It exists both as a discrete version, in which $x(t + 1) = f(x(t))$, and as a continuous version, in which for each i, $\mathrm{d}x_i/\mathrm{d}t = f(x(t))$.

The RD function relates to the interaction of individuals in the population through the following five steps. First, a population of individuals is presented and the variation of strategies in the population described in the population state. Second, in each period, every individual is paired at random with another individual from the population. These individuals play the strategies that they are programmed to play against each other. Third, a *game* is specified that members of the population play between each other. Commonly, this game is a two-player simultaneous-move game that for each player includes all strategies present in the population state. For each *strategy profile* (i, j) – a combination of strategy i of one player and strategy j of another player – the game specifies a *payoff* $u_k(i, j)$ for each player $k = \{1, 2\}$. Fourth, the payoff individual received from the interaction is interpreted as affecting the replication of this individual: how many individuals will play strategy i in the next period is proportional to how well individuals playing i in this period did vis-à-vis other individuals. Fifth, proportionality of replication and payoffs leads to differential representation of strategies in the population in the next period. Over many periods, this differential representation may lead to the convergence of stable state, in which differential representation of traits becomes stable over time, unless disturbed exogenously. Alternatively, differential representation might change in a regular fashion, for example, in regular oscillations or circles. Tracking the outcome of the dynamics over time reveals such stability or regularity results. Figure 5.1 depicts these five steps graphically.[3]

Mathematically, these steps are represented as follows. Given a population state $x(t)$, the expected payoff to any pure strategy i in a random match is $u(i, x)$: an

[2] The population state is formally identical to a mixed strategy. Its support is the set of strategies played by individuals in the population.

[3] These and the following graphs are schematic representations of models – of the formal RD equation *and* its respective interpretations. I use these graphs in order to make comparison between the different models more palpable.

individual that plays i against a randomly drawn opponent faces every strategy present in the population with the associated frequency with which that strategy occurs in the population. This is formally identical to this individual playing against an opponent who plays a mixed strategy $x(t)$. The associated population average payoff is $u(x, x) = \Sigma_i x_i * u(i, x)$.

The frequency of strategy i changes to the degree that the expected payoff $u(i, x)$ differs from the population average payoff $u(x, x)$. If $u(i, x)$ is greater than $u(x, x)$, the number of individuals playing i in the next period will grow more than the population average. If $u(i, x)$ is smaller than $u(x, x)$, the number of individuals playing i in the next period will shrink more than the population average. This relative growth is assumed to be linearly proportional to the difference between strategy payoff and the population average payoff.[4] Consequently, the continuous RD is specified as follows:

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = [u(i,x) - u(x,x)] * x_i \qquad \text{(Weibull 1995, p. 72)} \tag{5.1}$$

That is, the change in x_i's population share is determined by x_i's current population share and the difference between its expected payoff and the population average payoff.

Through analysis of a phase diagram of these dynamics, convergent trajectories, stable states and regular changes can be identified. Under the biological interpretation, regular changes identify the temporal predominance of certain traits in the population, while stable states identify results of adaptation of organisms to their environment.

4 The Biological RD

The RD was first derived in the late 1970s and quickly became the most prominent model of evolutionary dynamics in EGT.[5] The RD is derived from EGT by implicitly presupposing EGT to describe an underlying biological mechanism. The core idea of EGT in biology is that organisms often find themselves in strategic situations, in which the fitness-relevant outcome of their behaviour at a certain time depends on the behaviour of the other organisms in the population at that time. The fitness of an organism thus is influenced by the frequency of behaviour in that population. Consequently, there is a systematic relationship between the kind of

[4] The relation between proliferation and payoffs characterises different classes of selection dynamics. While a linear relation characterises the RD, wider classes are characterised by payoff positivity and payoff monotonicity, respectively (Weibull 1995, pp. 139–152). Yet the RD, which takes payoffs to represent fitness differences, is the most prominent selection dynamic in EGT and therefore will be discussed here.

[5] For a historical survey, see Grüne-Yanoff (2011a).

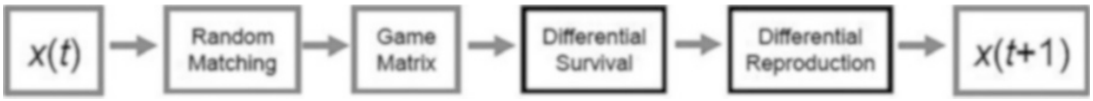

Fig. 5.2 The biological interpretation of the RD

composition of a population at a certain time and the differential reproduction of the respective strategies in that population at the next time step.

In particular, the biological interpretation gives causal substance to the formal five steps of the RD model above. First, individuals are interpreted as organisms and their strategies as certain inheritable behavioural traits. Second, organisms interact, for example, by fighting, mating, exchanging or collaborating. In this interaction, each organism exhibits the behavioural trait it is endowed with. Third, each organism receives an outcome from that interaction – for example, territory, food and mating partner – depending on its own behavioural trait and that of the organism it interacted with. Fourth, this outcome determines the number of offspring the organism has in the next period. Fifth, weighing growth of each trait by overall population growth yields the differential reproduction of each kind of organism.

The RD model is used to represent this mechanism. But it is not the formal RD model alone that performs this representational function, but rather a biologically interpreted RD model. In particular, the causally relevant properties are not found in the mathematical expressions of RD, but in its biological interpretation. This interpretation has turned the RD formalism into representations of specific causal forces and specific arrangement of these forces. Figuratively speaking, it fills in the black boxes of Fig. 5.1 to yield a causal process from population at t to population at $t + 1$, through interaction and differential reproduction, as shown in Fig. 5.2.

The biological interpretation of the mathematical expressions specifies the causal properties that bring about the result and that tell us *how* the population state changes from $x(t)$ to $x(t + 1)$. Hence, the mathematical RD model *in conjunction with* the biological interpretation represents the causal process from initial conditions to specific outcome, not the formal model alone.[6] The RD model is thus a mere sketch of the biological mechanisms it represents, as various gaps are filled in by the biological interpretation. Let me therefore distinguish the – more sketchy – formal RD model from the – less sketchy – biologically interpreted BRD model.

Even if it is less sketchy than the RD, the BRD does not represent a particular mechanism. Instead, it is a schema that represents mechanisms differing in many

[6] I have elsewhere argued that models consist of a formal structure *and* a story (Grüne-Yanoff and Schweinzer 2008). In the case I am discussing, the RD equation (5.1) constitutes the formal structure. The biological interpretation of its terms, and the account of interaction yielding a fitness-relevant outcome, leading to differential reproduction, constitutes the story.

details. BRD *abstracts* from these details. For example, it deals with strategies, abstracting from any concrete content of behavioural plans. Furthermore, it deals with generic organisms, not specific species or individuals. Finally, it abstracts from any differences between organisms, as when it assumes that organisms have the same fitness base rate.

Besides omitting and hence abstracting from many features, BRD also makes many specific assumptions about the processes it purportedly represents, even though these assumptions are likely to be false of many of these processes. Typical *idealisations* of this sort include the assumption that organisms match and interact with others randomly, hence idealising possible local interactions and network structure. It also idealises inheritance, assuming away epigenetic effects and sexual reproduction. BRD thus is an abstract and idealised representation of a process that supposedly can be found in many different concrete instantiations.

That it is seen as a representation of one abstract mechanism lies in the success of its application: many phenomena – in particular those involving frequency-dependent selection, like sex ratios, fighting behaviour or cooperation – have been successfully explained by reference to this abstract mechanism.

5 The Social RD

From the 1980s onwards, social scientists have increasingly adopted EGT for their own explanatory purposes. In particular, EGT has been used in order to explain the evolution of social institutions, in particular of conventions, norms and fairness preferences.

Sometimes, social scientists not only employed the general RD model but also resorted to its biological interpretation. For various reasons, this is today not considered adequate for most social science purposes.[7] Instead, specifically social interpretations of the RD have been proposed. These social interpretations represent mechanisms that account for the social interaction between individuals and for the social replication of these individuals' traits. A particularly important class of such mechanisms has been described as *learning*. Learning is an extremely open concept, and in the following I will only concentrate on those kinds of learning that are

[7] These difficulties spring from many sources; I just want to sketch three reasons here. First, while animals largely exist on the subsistence level, humans mainly do not. It is consequently much less clear what the causal effect of, say, adherence to norms is for survival and reproduction in humans, than what the causal effect of daily competition for food, shelter and mating opportunities is for survival and reproduction in nonhuman animals. Secondly, while it may be plausible that some basic animal behaviour is encoded in ways that can be inherited through reproduction, it is much less clear that complex human behavioural characteristics, like compliance with norms, can. Thirdly, the speed of cultural evolution is often much higher than human reproduction. Conventions in small groups, for example, can emerge or change within days, thus making reference to player reproduction inadequate. For these as well as other reasons, strategy replication often has to be thought of in ways independent of player reproduction.

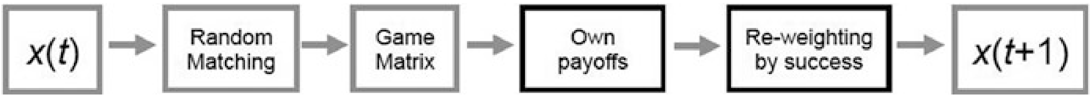

Fig. 5.3 The reinforcement interpretation of the RD

purportedly described by the RD. Within that class, three kinds of learning can be distinguished: reinforcement, imitation and belief learning.

In *reinforcement learning*, a player's received payoffs from past interactions are her only feedback information. That is, the probability of a strategy to be played in the future is proportional to the success it gave the player in the past. Börgers and Sarin (1997) present a well-known model of such learning, which conforms to the replicator dynamic. In their model, a player at stage n plays a mixed strategy $P(n) = (P_1(n),\ldots, P_J(n))$ that includes all possible pure strategies $S_1,\ldots, S_J$ in the population. The player i observes the (pure) strategy S_k and its payoff $u_i(S_k, S_{-k})$, normalized to lie between 0 and 1, that is realised when she implements her mixed strategy against other players playing S_{-k}. She then 'learns' by adjusting the weight P_k of S_k in her mixed strategy in proportion to the payoff that S_k gave her by the following rule:

$$P_k(n+1) = u_i(S_k, S_{-k}) + (1 - u_i(S_k, S_{-k})) * P_k(n) \tag{5.2}$$

$$P_{k'}(n+1) = (1 - u_i(S_k, S_{-k})) * P_{k'}(n) \quad \text{for all } k' \neq k$$

For the specific case of only two actions, the expected movement of action probabilities based on this model equals the RD, rescaled by a constant (Börgers and Sarin 1997; Börgers et al. 2004). More generally, if the decision-maker uses Cross' learning rule, (and satisfies the model's other requirements), then the learning dynamics satisfies *monotonicity* and *absolute expediency* (Börgers et al. 2004). Both of these properties are also satisfied by the RD. Thus, there is an analogy between Cross learning and the RD. Börgers et al. (2004, p. 358) conclude from this that their results 'strengthen the case of the use of RD dynamics in contexts where learning is important'. They also speculate that it may be possible to adopt their results 'to an evolutionary setting' (Börgers et al. 2004, p. 400) but refrain from making any specific claims about this.

The reinforcement interpretation of the RD model can be graphically presented as shown in Fig. 5.3.

This interpretation differs in a number of features from BRD. It commences with agents playing mixed strategies (where all organisms share the same support) rather than pure strategies. These strategies are not inherited, but adopted and adjusted by the agents. It does not interpret payoffs as fitness, but as subjectively evaluated outcomes. It is these subjective evaluations that cause the agent's adjustment of her own strategies. And it is this adjustment, and not differential reproduction, that constitutes differential representation in the population.

In *imitation* learning, players occasionally sample other players in the population and learn about their strategy and the payoff they realised in the last round. They

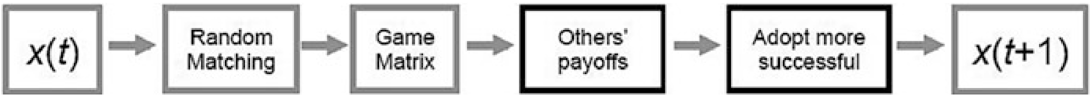

Fig. 5.4 The imitation interpretation of the RD

then switch their strategies according to the following rule: if in a population with state $x(t)$ the agent i's payoff is $u_i(x)$, and the agent samples an agent j with payoff $u_j(x)$, the agent switches with probability[8]

$$q_i = \max\{0, b(u_j(x) - u_i(x))\} \tag{5.3}$$

(Schlag 1998, p. 150, cf. also Weibull 1995, pp. 152–161). That is, she retains her strategy if her realised payoffs are greater than that of the sampled player. Otherwise, she adopts the strategy of the sampled player with a probability proportional to the difference between her and the sampled payoff. For this reason, such models are sometimes seen as closely related to the meme concept (Börgers 1996). The resulting population dynamic – in a large but finite population – is approximated by a deterministic dynamic that is analogous to the discrete RD (Schlag 1998, p. 152). Schlag furthermore points out that his model arrives at this result solely based on individual information and induced performance, while reinforcement learning models discussed above 'contain axioms concerning the functional form of a desirable learning curve' (Schlag 1998, p. 153).

The imitation interpretation of the RD model can be graphically presented as shown in Fig. 5.4.

This interpretation differs in a number of features from BRD. Although agents here also play pure strategies, it drops the heritability of strategies. Like the interpretation of Fig. 5.3, it does not interpret payoffs as fitness, but as subjectively evaluated outcomes. But unlike the reinforcement schema, the imitation schema models agents as evaluating not only their own but also others' outcomes. It is these subjective evaluations that may cause the agent to adopt another agent's strategy if she finds it more successful than her own. And it is this conditional adoption, and not differential reproduction, that constitutes differential representation in the population.

The previous two kinds of models cast learning as an influence of past payoffs (either of the player herself or of other players) on future behaviour. *Belief learning*, in contrast, models learning as experience influencing beliefs, and only through this influence, there is an indirect effect on behaviour. Hopkins (2002, p. 2144) has termed the particular kinds of belief learning modelled with EGT 'hypothetical reinforcement'. This is because players are modelled as calculating what they

[8] The function b ensures that the difference are normalised – that is, for any payoffs u_i, u_j in the population, $0 \leq b(u_j(x) - u_i(x)) \leq 1$.

would have received had they chosen some other action, on the basis of knowledge of their own payoff matrices and observations of their opponents actions.

This approach reinterprets EGT in general, and strategy selection in particular, as a theory of individual mental processes. Under this interpretation, all references to payoffs of others in a given environment are understood counterfactually as the payoffs that one would get in that environment if one adopted the other's strategy. For example, if a player knows the payoffs of each strategy profile, and knows the frequency with which strategies are played in the population, she can compare the expected payoffs of these strategies based solely on her own preferences. Having compared the strategies according to her own preferences, she can then choose that strategy that is either better than the current strategies or a best reply to her belief about the frequencies in the population. Variants of such models have been proposed by Sugden (1986), in Kandori et al.'s (1993) 'stochastic fictitious play' and in Young's (1993) 'adaptive play'.

Take, for example, Young's (1993) model. He defines play at time t as the strategy-tuple $s(t) = (s_i(t), \ldots, s_n(t))$, consisting of each player's strategy choice at time t. At period t, each player samples the past play h of a certain number of past periods. From this sample, the player constructs strategy-tuple s_h by weighing the past play in some way. Strategy-tuple s_h constitutes her estimate how other players will play in the next period. Thus, for the next period, agent i chooses s_i as the best reply to s_h. By choosing s_i, the player replaces the history of past play h with a new history h', in which the earliest period is removed and the most recent play added. This yields a process

$$P^0_{hh'} = \Pi p_i(s_i|h) \tag{5.4}$$

Where $P^0_{hh'}$ is the probability of moving from h to h′, determined as the product of the player's probabilities of choosing s_i given sample h. Young calls this process *adaptive play*.

Young's model is an example of what I call a *mental play* interpretation of EGT. What is relevant for a certain strategy to be selected no longer is the effect of actual interaction in a real population, but rather the consequence of an individual player evaluating various options, based on her subjective value criteria and her beliefs what her opponents will play. She forms these beliefs from her perception of and through reasoning about others' past play. She chooses her strategy by mentally representing her various options in the anticipated environment, figuring out the consequences of these counterfactual scenarios and choosing the one with the outcomes she values better or best.

Consequently, because the causal relation is between interaction and individuals' mental attitudes, no interpersonal payoff comparison is necessary. Players only observe their own payoffs from past play, and this affects only their own attitudes towards future play. Effects on aggregate properties are not directly modelled.

If noise is introduced into models of fictitious play, the expected motion of fictitious play becomes a form of noisy replicator dynamic (Hopkins 2002, p. 2149).

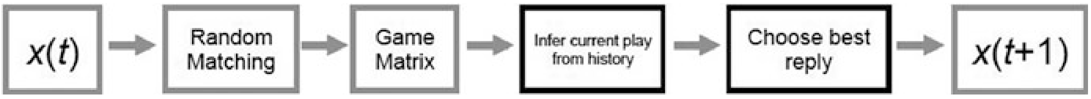

Fig. 5.5 The belief-learning interpretation of the RD

The only way that learning behaviour generated by stochastic fictitious play differs from the population dynamics of the two previous models is that they may differ in speed of passage along similar paths.

The belief-learning model can be graphically presented as shown in Fig. 5.5.

This interpretation differs in a number of features from BRD. It commences with agents playing mixed strategies (where all organisms share the same support) rather than pure strategies. These strategies are not inherited, but adopted and adjusted by the agents. Furthermore, it makes the crucial extra assumption that the whole population and all its strategies and payoffs are mentally represented by each organism. Based on this representation, the agent estimates how other players will play in the next period. Furthermore, the schema does not interpret payoffs as fitness, but as subjectively evaluated outcomes. Based on the estimation of others' future play, and her own subjective evaluations, the agent then chooses her action as a best reply. It is this deliberation, and not differential reproduction, that causes differential representation in the population.

6 Relating the Mechanisms

It should be clear from the comparison of the previous section that the three learning mechanisms are not identical with what the BRD represents. In particular, what kind of strategies individuals play, how payoffs are realised and what information and what mental capacities individuals employ in replicating strategies differ considerably between BRD and the learning interpretations of the RD (as well, to a lesser extent, between these interpretations themselves). Thus, the BRD and the respective learning interpretations of the RD represent *different* mechanisms, even though all these mechanisms are represented by the same RD model.

The RD model thus appears in the first instance as a highly abstract mechanism sketch. It is used to represent different kinds of mechanisms, but for each of these representation tasks, it needs to be filled in with a more domain-specific interpretation or story.

Nevertheless, one might still want to defend the claim that the general RD represents one mechanism – namely, by arguing that the BRD and the learning mechanisms all instantiate a more abstract mechanism and that this abstract mechanism is represented by the general RD model.

This idea seems prima facie plausible, particularly when one recalls that the BRD and the learning models themselves are abstract representations of mechanisms. As I discussed in Sect. 4, the BRD abstracts from any concrete content of behavioural plans, from specific species or individuals and from any differences between

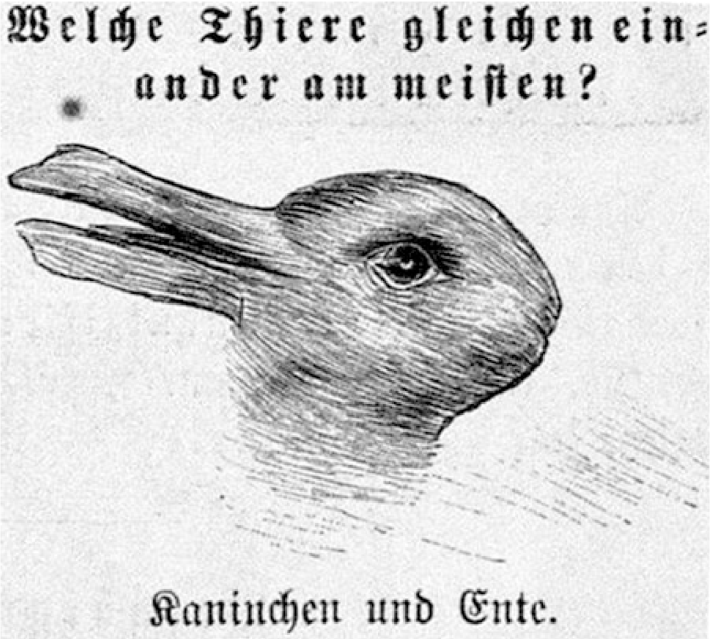

Fig. 5.6 *Fliegende Blätter* (Oct. 23, 1892, p. 147, Nr. 2465)

organisms. For example, it omits representations of how organisms reproduce and instead describes the stage as a general process of reproduction in all its possible forms. So if the BRD is an abstract representation of a class of mechanisms, why should the general RD not be an even more abstract representation?

Furthermore, the fact that both the BRD and the learning models use the RD for their representational tasks seems to provide evidence that indeed there is an abstract mechanism instantiated both in the more concrete biological and social mechanisms and that this more abstract mechanism is represented by the RD model. Because the RD contains those features shared by the BRD and the learning models, it might seem plausible to conclude that the RD represents that abstract causal structure shared by the biological and the social mechanisms.

Against this appearance, I will now argue that the general RD model is *not* a representation of an abstract mechanism, instantiated by both the biological and the social mechanisms. Rather, the way the two disciplines 'fill in' the RD model in order to represent their respective mechanisms differs considerably. Users of the RD model, when filling it in, make systematically different kinds of idealisations, depending on whether it is interpreted in economics or in biology. This leaves little to be shared between the respective represented mechanisms – little that could be represented by a single RD, however interpreted. Instead, the RD model faces an *idealisation gap*: it can be interpreted *either* biologically *or* in one of the learning senses, but it cannot be interpreted to capture the essence of all, because there is little essence to capture. To clarify my argument, let me illustrate it with a joke.

The joke's not mine – it was published 120 years ago in the *Fliegende Blätter*, a German satirical weekly. Most philosophers know its subject, the duck-rabbit, from Wittgenstein's discussion of aspectual perception or from Kuhn's discussion of a paradigm shift. What those discussions ignore is the way the joke was posed, as shown in Fig. 5.6. The German headline reads 'which animals are most similar?', and the answer is 'rabbit and duck'.

The author of this little vignette thus did not solely intend to entertain with the *Gestalt* shift, but rather used this shift in order to infer an obviously absurd and hence satirical conclusion: because the same image represents both a rabbit and duck, it is suggested, we must conclude that rabbit and duck are indeed most similar.

Obviously, this inference is absurd for a number of reasons. I want to focus here on a rather subtle one, namely, that the same image relates to the two objects it supposedly represents in different ways. When we use the above image as a representation of a rabbit, we make certain kinds of idealisation. For example, we idealise the size of the rabbit's mouth and nose, as well as the shape of its ears. When we use the image as a representation of a duck, however, we make *different* idealisations: the back of a duck's head looks different, and it has different markings on its feathers. Thus, when using the image to represent either the one or the other, we make different allowances for which part of the image may not be representationally accurate. The ingenuity of the draughtsman lay in creating one image that allowed us to make the respective idealisations in such a way that it can function either as a representation of a rabbit *or* a duck. By making these different idealisations, we adapt the image for its respective uses. Although a duck shares some features with the image, and a rabit also share some features with the image, *these are not the same features*. Thus understood, there is little reason to believe in the similarity of rabbits and ducks because they are representable with the same image.

The same holds for the RD model. To use a model as a representation, we always have to make some idealising assumptions. But when interpreting the RD model biologically, we make idealisations that systematically differ from those we make when interpreting the RD model socially. Let me list some of these differences.

First, all three learning models require that players in some way identify actions and strategies – either of their own or possibly of others. If agents could not identify strategies in this way, they would not be able to link a diagnosis of 'success' with the choice of a successful strategy. This stands in contrast to the biological model, where the strategy notion only fulfils a theoretical role: differential reproduction does not require that the organism identify the strategies.

This additional requirement pushes these learning models beyond a simple notion of copying. Rather, it involves the ability to attribute goals and intentions. 'Something other than copying is taking place' (Sperber 2000, p. 171), and this other factor may have the power to lead the process in directions that mere copying would not. Yet such factors are idealised away in all of the three learning models.

Second, unlike the biological model, the learning models make specific assumptions about the learning rules players employ, at the exclusion of other, possible rules. In the biological model, if the payoffs are interpreted as fitness, there is a natural justification for a linear relationship between payoffs and differential reproduction. Yet in the learning models, specific imitation and reinforcement rules have to be chosen to arrive at a linear relationship.

> Other imitation rules – as plausible as Schlag's – yield processes different from any biological ones. (Börgers 1996, p. 1383)

This is even more obvious with respect to belief learning. For example, choice of different reasoning principles or heuristics may lead to different beliefs about strategies, strategy outcomes, etc., even when based on the same actual interactions. This sensitivity of the population dynamic to the specifics of the learning rules increases the 'idealisation gap' between the biological and the learning models.

Third, and related to the previous point, all learning models have to make strong assumptions about players not making mistakes – they never switch from a better to a worse strategy. This is a real possibility in all learning models – as agents have to actively identify strategies, associate payoffs with them and choose their own actions on that basis – while it has no significance in the biological model. The way this is dealt with usually involves taking expected values. Averaging this way over the possible behaviours of an agent idealises the influence of players' mistakes away: even if there is a positive probability that a player will switch from better to worse, on average the player will not (cf. Gintis 2000, p. 192).

Fourth, stochastic fictitious play models face the particular problem of excessive time horizons. As Sobel starkly puts it,

> the long-run predictions [of stochastic fictitious play] only are relevant for cockroaches, as all other life forms will have long been extinct before the system reaches its limits. (Sobel 2000, p. 253)

To turn the stochastic belief-learning models into representations of social mechanisms, the time horizons thus must be idealised.

Fifth, the imitation learning model faces the particular problem of requiring interpersonal comparisons of utility (Grüne-Yanoff 2011b). The biological RD model does that, too – yet while this requirement is innocuous under the fitness interpretation, it is highly problematic when payoffs are interpreted as numerical representations of preferences. Thus, this extra requirement constitutes an important difference between the belief-learning models and the other models discussed here.

Certain substantial idealisations need to be taken also when the RD model is interpreted biologically. A different set of substantial idealisations needs to be taken when the RD model is interpreted socially. By making these different idealisations, we adapt the model for its respective representative uses. This is standard scientific practice: most, and possibly all, model uses involve idealisations.

Yet when the same formal structure is employed to construct different, more specific mechanistic models, and each of these models involves different idealisations, one has to be careful when inferring purported similarities between these different mechanisms based on the common formal structure. Like the duck-rabbit, the RD equation is adapted for its respective representative tasks. In the course of each adaptation, certain features of the RD are drawn on – others are accepted as useful or at least harmless idealisations. Which features are drawn on and which are accepted as idealisations differ with each adaptation. The mechanism that each adaptation of the RD represents is substantially different from each other and does not share any or little causal structure between each other. Thus, there is no abstract mechanism that is instantiated by the biological and learning mechanisms, and consequently the RD cannot represent such a mechanism.

7 Conclusions

The general RD is a model that is used in biology to represent biological mechanisms and in the social sciences to represent social mechanisms. Substantial idealisations have to be made for these purposes – idealisations that differ for the respective disciplines. These create a considerable idealisation gap between the BRD and the learning interpretations of the RD. This gap is sufficiently large to conclude that the general RD does not represent an abstract mechanism that subsumes both the biological and the social cases. Just like the duck-rabbit image does not represent the essence of both duck and rabbit, but rather either a duck or a rabbit (depending on what idealisations one accepts), so the general RD represents either biological or social mechanisms, but not the shared causal structure of both.

References

Bechtel, W., and A. Abrahamsen. 2005. Explanation: A mechanist alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences* 36: 421–441.

Börgers, T. 1996. On the relevance of evolution and learning to economic theory. *The Economic Journal* 106: 1274–1385.

Börgers, T., and R. Sarin. 1997. Learning through reinforcement and replicator dynamics. *Journal of Economic Theory* 77: 1–14.

Börgers, T., A. Morales, and R. Sarin. 2004. Expedient and monotone learning rules. *Econometrica* 72: 383–405.

Cartwright, N. 1989. *Nature's capacities and their measurement*. Oxford: Clarendon.

Craver, C.F. 2006. What mechanistic models explain. *Synthese* 153: 355–376.

Craver, C.F., and M.I. Kaiser. 2013. Mechanisms and laws: Clarifying the debate. In *Mechanism and causality in biology and economics*, ed. Hsiang-Ke Chao, Szu-Ting Chen, and Roberta L. Millstein, 125–145, Dordrecht: Springer.

Gintis, H. 2000. *Game theory evolving*. Princeton: Princeton University Press.

Grüne-Yanoff, T. 2011a. Models as products of interdisciplinary exchange: Evidence from evolutionary game theory. *Studies in History and Philosophy of Science* 42: 386–397.

Grüne-Yanoff, T. 2011b. Evolutionary game theory between interpersonal comparisons and natural selection: A dilemma. *Biology and Philosophy* 26: 637–654.

Grüne-Yanoff, T., and P. Schweinzer. 2008. The roles of stories in applying game theory. *Journal of Economic Methodology* 15(2): 131–146.

Hopkins, E. 2002. Two competing models of how people learn in games. *Econometrica* 70: 2141–2166.

Kandori, M., G. Mailath, and R. Rob. 1993. Learning, mutation, and long run equilibria in games. *Econometrica* 61: 29–56.

Mäki, U. 2009. MISSing the world: Models as isolations and crediblesurrogate systems. *Erkenntnis* 70(1): 29–43.

Mayntz, R. 2004. Mechanisms in the analysis of social macro-phenomena. *Philosophy of the Social Sciences* 34: 237–259.

Schlag, K. 1998. Why imitate, and if so, how? A boundedly rational approach to multi-armed bandits. *Journal of Economic Theory* 78(1): 130–156.

Sobel, J. 2000. Economists' models of learning. *Journal of Economic Theory* 94: 241–261.

Sperber, D. 2000. An objection to the memetic approach to culture. In *Darwinizing culture: The status of memetics as a science*, ed. Robert Aunger, 163–174. Oxford: Oxford University Press.

Sugden, R. 1986. *The evolution of rights, cooperation, and welfare*. Oxford: Basil Blackwell.

Weibull, J.W. 1995. *Evolutionary game theory*. Cambridge, MA: M.I.T. Press.

Woodward, J. 2002. What is a mechanism? A counterfactual account. *Philosophy of Science* 69: S366–S377.

Young, H.P. 1993. The evolution of conventions. *Econometrica* 61(1): 57–84.