The nature of institutions from a computational perspective

Claudius Gräbner[*] Wolfram Elsner[†]

Early draft, December 2015

1 Motivation and outline

Institutions are more and more considered to be one of the essential determinants of economic development. The claim that they are not only a determinant of the growth level of an economy (Acemoglu, Johnson, & Robinson, 2001; Rodrik, 2008), but also determine the distribution of income, wealth, and power (S. Bowles, 2004; Elsner, Heinrich, & Schwardt, 2014), has led to an enormous wave of empirical literature on the role of institutions (Acemoglu et al., 2001; Samuel Bowles & Gintis, 2002; Acemoglu & Robinson, 2006).[1] Yet, there is still no theoretical consensus about what an institution actually is, how it affects individual decision making, and how it evolves.[2] An important theoretical piece in the puzzle of institutions seems to be missing.

In this article I try to reconcile the most common and promising conceptions of institutions that are currently available in economic theory to detect this missing piece. All of these conceptions were developed in different contexts and either has particular advantages and disadvantages. Yet, in order to understand the role institutions play in actual economies it would be desirable to develop a general conception of institutions considering the respective approaches as special cases. The approaches I deal with in this article are summarized in table 1 are the following:

[*] Institute for Institutional and Innovation Economics, University of Bremen, Germany. Email: graebnerc@uni-bremen.de

[†] Institute for Institutional and Innovation Economics, University of Bremen, Germany. Email: welsner@uni-bremen.de

[1] Organization, previously often contrasted to institutions, are now usually considered particular institutions as well: organizations are institutions that ”involve (a) criteria to establish their boundaries and to distinguish their members from nonmembers, (b) principles of sovereignty concerning who is in charge, and (c) chains of command delineating responsibilities within the organization”. Hodgson, 2006, p. 8.

[2] A recent special issue of the Journal of Institutional Economics (Vol. 11(3) in 2015) on this topic supports this claim.

Concept	Distinctions	Examples	Methodology
Institutions as rules of the game	none	North (1990), Ostrom (1990), Bravo (2011)	Classical Game theory, ABM
Institutions as equilibria	Institutions as Nash equilibria themselves	Calvert (1995)	Classical Game Theory
	Institutions as means to accomplish Nash equilibria	Greif and Kingston (2011)	Classical Game Theory
	Institutions as correlative equilibria	Gintis (2009)	Epistemic Game Theory
Institutions as belief systems	Institutions as constitutive rules	Searle (2005, 1), Hodgson (2006)	Verbal description
Institutions as belief systems	Institutions as ideas	Dopfer, Foster, and Potts (2004), Elsner (2012)	Verbal description, Evolutionary GT

Table 1: Existing concepts of institutions

1. Institutions as the rules-or-the-game, as advocated by North (1990) and as still widely used in the literature on common pool resources (Ostrom (1990), Bravo (2011)).

2. Institutions as equilibria in a game. There are several distinct conceptions commonly summarized under this heading: The view that institutions are Nash equilibria of games (Calvert, 1995), are means to accomplish Nash equilibria (Greif & Kingston, 2011), and that institutions are correlative devices that lead to correlated equilibria rather than Nash equilibria (Aumann, 1987; Gintis, 2009). In contrast to the institutions-as-rules literature these conceptions include an explicit considerations of the motivations for following an institution.

3. Institutions as shared belief systems. This view is particular prominent in the evolutionary-institutional economics research program as represented by Searle (2005, 1), Hodgson (2006), Dopfer, Foster, and Potts (2004) or Elsner (2012).

To have so many diverse conceptions of such a central economic concept as institutions is confusing and not satisfying. Still, there is almost no attempt in the literature to reconcile these diverse conceptions.

The first main objective of the article is therefore to condense the common essence of the concepts and to integrate this essence into a new general concept.There are also aspects of the existing concepts that seem to be mutually exclusive: for example, an institution cannot be a Nash equilibrium and a mean to accomplish a Nash equilibrium simultaneously. In these cases I will provide arguments on on which of the mutually exclusive parts should be dropped. It is hoped that this step brings about theoretical clarity about what is meant by the term ”institution” and makes institutional economics even more productive in the future.

The second main question is related to the first one and concerns the adequate methodology to study institutions as defined therein. This is particularly important as some of the definitions mentioned before directly result from a particular tool: game theory, the most prominent tool applied in institutional economics. While I am sympathic to a certain degree of game theoretic reasoning when studying institutions, *defining* institutions in game theoretic terms is dangerous: not only is such a definition dependent on very strong behavioral assumptions, it also devalues any reasoning about institutions that is not based in game theory. And there is evidently a lot of valuable theoretical literature on institutions that has not yet been formalized in a game theoretic context. Some of these verbal findings are difficult to put into a game theoretic setting because the assumption to be made are too restrictive or the game theoretical approach does not allow for a sufficiently dynamics perspective on institutions. In the second part of the analysis, which is located in a subsequent paper, I therefore propose a very general and flexible formal framework for studying institutions. This framework is *computation*. This means that it draws upon some game theoretical concepts, but transfers them into a more flexible *computational* language.

There are two immediate justifications for building such a general framework on computational foundations: the first one is methodological: thanks to their flexibility, computational models allow for the consideration of most of the verbal but also empirical and experimental results already obtained. Analytical models, on the other hand, quickly become intractable if too many of such considerations are included into the model. But the computational and analytical approach is not mutually exclusive: rather, the former is a generalization of the latter: in case a more abstract analysis is attempted for, the computational approach can be simplified into a game theoretic analysis. The advantage of having the general, computational conception in the background is that the resulting game theoretic analysis is then a special case of a more general concept. It can therefore be compared much easier to other special cases of the general concept. If the general concept was missing (as it is currently the case), different applications are difficult to relate to each other.

The second justification for taking a computational starting point is theoretical and relates directly to the nature of institutions: In almost all existing conceptualizations of economic institutions, institutions are always about *computations*: they affect and depend upon individual decision making, and making a decision means to compute something. This means that if we wish to understand what institutions are and how they work we need to take computation seriously. Yet, the idea of *computation* is largely absent from the discussion about economic institutions. Therefore, both analytical and verbal models must necessarily remain incomplete before an adequate notion of computation is added to these models. Moreover, formalizing a verbal in a computational model is always more than just a mechanistic task: it involves theorizing itself, because writing a computer program of institutional change brings questions about the ontological status of an institution, and the precise mechanism underlying institutional change to one's attention, that may be omitted in a (necessarily more sketchy) verbal model.

The rest of the paper is structured as follows: the next section reviews the three established approaches to conceptualize institutions mentioned above: the idea to consider institutions-as-rules (IAR), the idea to conceptualize institutions-as-equilibria (IAE) and to consider institutions as systems of shared beliefs or constitutive rules. Note that the first two approaches are formalist and dependent on applying game theory to model institutions, while the last one is mostly verbal and the question of how they can be merged and enriched by computational content will be discussed at the end of the section.

Then section 3 summarizes and illustrates the mutual benefits of the conceptions and the open challenges. It thus paves the further way for an overall synthesis. Section 4 explains in a verbal way how these challenges can be dealt with in a computational model. The last section concludes.

2 Established conceptualizations of institutions

2.1 Institutions as rules

The conception of institutions as rules dates back to North (1990). He considers institutions to be ”the rules of the game in a society...[and] the purpose of the rules is to define the way the game is played” (North, 1990, p. 3).

Institutions in the sense of North can come in either of two forms: formal or informal. A formal institution comes in the form of codified law, e.g. a law of the local government that forbids people to drive faster than 30 km/h in a peasant area. Or the institution comes as an informal rule, such as the shared belief that ”it is just not proper” to drive faster than 30 km/h in a peasant area. In both cases, institutions act as constraints in the individual decision problem.[3]

2.1.1 Optimality of institutions and transaction costs

Up to this point there is no assertion of optimality of current institutions: whether a society is able to replace inefficient institutions with more efficient ones depends on the mechanism according to which institutions are formed and changed. This mechanism can be either centralized (e.g. through voting, e.g. Ostrom (1990)) or decentralized (e.g. through evolutionary selection of superior institutions, e.g. Alchian (1950)). It is ultimately an empirical question what the particular mechanism at work is.

One of the most influential branches within the IAR approach, the transaction-cost literature, takes a more pronounced view on the optimality of current institutions. The transaction-cost literature took off with Williamson (1979) and starts from the fact that all economic interactions

[3] Of course, in the end institutions can be more than mere constraints: reasonable traffic rules may constrain us from excessive speeding, but may *enable* us to move safely and without traffic jams (Hodgson, 2006).

are associated with certain costs. These costs can be reduced through institutions. Similar to the reasoning of Alchian (1950) on organizational forms, Williamson (1979) asserts that the most efficient institutions are selected by the market or by other selection mechanisms. While the existence of transaction costs is well known since Commons (1924), the claim of an optimal selection mechanism seems questionable, at least if one considers the state-or-the-art literature in evolutionary biology (Nowak, 2006), or the literature taking the path dependent character of institutions more seriously (Myrdal, (1954)2004; Acemoglu & Robinson, 2006).[4]

Also, even if institutions were identified as the optimal response to a given problem, difficulties of institutional change may stop them from being substituted by new institutions which are more adequately to the present situation. Also, limited computationally capacity of the agents can lead to suboptimal institutions if there is more than one relevant problem structure (Bednar & Page, 2007). One should therefore be very skeptical against any *a priori* conception of institutions as optimal, if no particular reference to the mechanisms that may assure this optimality is given. We will return to this in the general model below. [5]

2.1.2 Why do people follow the rules?

But there is a more fundamental problem with the IAR approach that is not specific to the transaction cost literature: it does not focus on the question of *why* people follow institutions (Greif & Kingston, 2011).

Consider the following illustration of the problem: assume a situation in which a group of people enjoys welfare $\mathfrak{w}_1$ under a set of institutions $\mathscr{I}_1$. Assume there is an alternative set of institutions $\mathscr{I}_2$ that would increase overall welfare significantly to $\mathfrak{w}_2$. However, there is an influential subgroup whose income $\mathfrak{w}_1^{\sim}$ under $\mathscr{I}_1$ is slightly higher than under $\mathscr{I}_2$. A change from $\mathscr{I}_1$ to $\mathscr{I}_2$ can only take place if this group supports the institutional change. Now assume that $\mathfrak{w}_1^{\sim} - \mathfrak{w}_2^{\sim} < \mathfrak{w}_2 - \mathfrak{w}_1$. Thus, it would be the best option for all if the small group agrees to the change from $\mathscr{I}_1$ to $\mathscr{I}_2$, and the rest of the people agrees to reimburse the group for their losses. This would still improve overall welfare. But there is a fundamental problem: once the institutions have been changed, why should the majority still reimburse the smaller group? This is the fundamental problem of *commitment* (Acemoglu & Robinson, 2002) that cannot be fully resolved in the IAR approach. One would require more precise theory about the motives and incentives that people face when making their decision to follow a particular institution. This is the starting point for the IAE approach discussed below.

[4] Furthermore, if "optimal" means "minimizing transaction costs" it is not clear from which standpoint the institution is optimal. This may depend on how the costs are defined (i.e. whether all external effects are considered) and what role the institution plays for the distribution of income.

[5] This certainly has a relation to Veblen's dichotomy between ceremonial and instrumental dimensions of institutions and the institutional life cycle that will be discussed below.

2.1.3 Institutional change

Institutional change is considered to occur only in an incremental, and highly path dependent way (North, 1990). This is partly because existing institutions may change people's perceptions and preferences. Thus, after some time, people start preferring other institutions. In such a case, and assuming a well-functioning mechanism that enables ordered institutional change, a certain minimum critical mass of people usually may get united to change the institutional structure. One such mechanism could be simple majority voting.

Another form of institutional change is the transformation of formal into non-formal rules. The reliance on informal rules, however, frequently runs into trouble because all institutions involve to a certain degree an informal element (Hodgson, 2006). Furthermore, informal rules are hard to observe or to capture theoretically (Greif & Kingston, 2011, p. 24). Thus, in practice, informal rules are often left unexplained. But because informal rules play an important role in the explanations of institutional change within the IAR approach, explaining change is a particular challenge for the approach and often remains unsatisfactory. The IAR approach always works most successful if formal (or informal) rules could be reasonably inferred from the data.

2.1.4 Methodology and example

Most studies within the IAR approach rely to some extent on game theory. But there are also an increasing number of agent-based simulations, in particular in the common pool resources (CPR) literature that focus particularly on institutional change. In this context, Ostrom (1990) developed a more general framework, the IAD framework, that can be combined with either ABM or GT to assure a more comprehensive analysis (Gräbner, 2015b). As an illustration, consider the model of Bravo (2011):

In this model, agents are allocated on a grid. In every patch there is a certain amount of a common pool resource (CPR) that can be exploited by the agents. Agents have a preference about what fraction β of the initial amount of the CPR should be preserved. Additionally, agents also have beliefs about what individual level of extraction is compatible with this preference. Bravo uses a parameter κ to denote the minimum level of the resource that should be observed in the surrounding of the agent before he starts extraction: as long as the level of the resource is above κ, agents extract the resource. If it is below κ they may move to a different spot. The resource regenerates at a constant rate.

Agents update their personal κ every round by comparing their actual payoff π_t with the payoff of the previous period π_{t-1}. If $\pi_t < \pi_{t-1}$, the agent changes his κ. In case the actual total amount of the resource is lower than the preferred preservation, the agent increases κ , as he attributes the payoff loss to excessive extraction. Otherwise, he will increase it. The model also exhibits a selection mechanism according to which the worst performing agent copies the preference of the best performing one. A refinement of the model allows for institutional change: agents do not decide on the base of their individual κ, but on a global institution $\mathscr{K}$ that specifies the level beside which agents are allowed to extract. Agents can be dissatisfied with the current institution if either $\pi_t < \pi_{t-1}$ or $|\mathscr{K} - \kappa_i|$ exceeds a certain threshold value. If more than $2/3$ of the agents are unsatisfied with the institution they replace the old institution with a new one. The new institutions prescribes the average κ of the unsatisfied agents.

The model illustrates the fundamental result of the IAR approach in the context of CPR: there is no ”tragedy of the well managed CPR”, but only a ”tragey of the open-access resources”.

To look at this model is particular instructive because the simulation method chosen by the authors allows them to consider evolutionary processes, endogenous change of the resource, and the consideration of institutional change to the extent it is possible within the IAR. It therefore illustrates the fact that the IAR framework provides a suitable frame of analysis if institutional change is not at center stage, or if straightforward mechanisms for the generation of institutions exist and are known to the researcher - which is often unfortunately not the case. But the model also illustrates the fundamental weakness of the approach: its lack of a straightforward account for the question of why people follow an institution.

2.2 Institutions as equilibria

Considering institutions as equilibria addresses this major shortcoming of the IAR approach: it rests on the assertion that institutions yield situations in which agents mutually give a best response to each others behavior and that this behavior fosters the expectation of each agent to continue behaving as she is currently doing. This stability provides people an incentive to follow institutions because they help them to solve problems of coordination or cooperation: considered this way, following an institution is just a rational decision.[6]

Institutions thus represent some sort of emergent properties: they are a combination of action, or beliefs, or expectations, that manifest in a certain combination of behavior. They are emergent because they cannot be understood by studying isolated individuals *by definition*.

2.2.1 Institutions as Nash equilibria?

Within the IAE approach, there is still some disagreement about what an institution actually is: some purists take the title of the approach literally and equate institution with game theoretic Nash equilibria (Calvert, 1995). To get the basic idea, consider a normal form game $\mathscr{G}$ with players $i = 1, ..., n$ who have finite individual strategy spaces S_i such that the overall strategy space is $S = \prod_{i=1}^{n} S_i$. Preferences are captured by the individual payoff functions $\pi_i : S \to \mathbb{R}$. Different individual payoff functions give rise to different problem structures, such as coordination games, or social dilemmas. In any case, a Nash equilibrium is a combination of strategies such that no player has an incentive to change her strategy: all players give mutual best responses to each other. Formally, a Nash equilibrium is a strategy vector $s^* = \{s_1^*, s_2^*, ..., s_i^*, ..., s_n^*\}$ such

[6] Although this does not mean that the the decision is always conscious.

that $\pi_i(s_i^*, s_{-1}^*) \geq \pi_i(s_i^{\sim}, s_{-1}^*) \forall i$ with $s_i^* \neq s_i^{\sim}$ and $s_{-i} = \{s_1, .., s_{i-1}, s_{i+1}, ..., s_n\}$. Such a situation captures the notion of stability that is often associated with institutions. It also does not claim any form of optimality or fairness: Nash equilibria can be highly inefficient (like in the prisoners' dilemma), or may give players very asymmetric payoffs (such as in the battle of the sexes game). Nevertheless, by asserting that "there is, strictly speaking, no separate animal that we can identify as an institution. There is only rational behavior, conditioned on expectations about the behavior and reactions of others..."Institution" is just a name we give to certain parts of certain kinds of equilibria" Calvert takes a very strong position that is hard to defend for at least four reasons:

Firstly, from a practical perspective, such a conception of institutions significantly limits the ability to study institutional change: Nash equilibria represent a situation of stability and consistency of mutual expectation without any incentive to change one's behavior. The conception is also of limited help if one wishes to understand how institutions emerge, and how people adapt their behavior to them. The equilibrium is only the ultimate result of such a process.

Secondly, from a theoretical perspective, it does not provide any reasonable information about how an institution actually affects individual behavior: the explanatory burden is given entirely to individual rationality, while it is not explained where this comes from, and how it may change over time.

Thirdly, from a methodological perspective, it suggests a focus exclusively on game theory. Given all the informative studies on institutions that do not make use of game theory, such a focus seems to be too restrictive.

Lastly, there is a somehow disturbing ontological implication of this claim: institutions as such would not be an interesting object of investigation as they are mere derivative of individual rationality and the whole task of the researcher would be to explain the particular preference ordering of the agents involved.

2.2.2 Institutions as correlative devices

The Nash equilibrium is not the only solution technique for games. And in fact, in many instances, *correlative equilibria* form a far more intuitive path to understanding institutions (Aumann, 1987; Gintis, 2009). If we solve a game using a correlative equilibrium, we consider institutions to be *means* to accomplish equilibria, rather than equilibria themselves.

Formally, consider a normal form game $\mathscr{G}^N$ as defined above. An *epistemic game* $\mathscr{G}^E$ is the corresponding normal form game $\mathscr{G}^N$ that has a current state $\omega \in \Omega$ and a knowledge partition $\mathscr{P}_i$ for every player. Furthermore, every player has a certain belief about the current state of the game. If the real state of the game is ω, then $p_i(\omega'; \omega)$ denotes the probability that agent i expects the actual state to be ω'. The knowledge partition of the player specifies which states she is able to identify correctly, and which states she cannot distinguish. The beliefs are therefore a consequence of the knowledge partition.[7]

The fundamental idea that matters here is that before the players play the game as it has been conceptualized in $\mathscr{G}^N$, a correlating device will issue directives to each player in the game. If to follow the directives of this device is a Nash equilibrium, then the resulting equilibrium is called a *correlated equilibrium*.

The correlating device could be a physical object that has received particular meaning through a *constitutive rule*, as defined below. Signs of traffic lights are examples. But it could also be a convention: consider the property rights game from before: As illustrated by Hindrinks and Guala (2015), one could think of a simple rule like ”the land belongs to the person who first harvested on the land”. In such a case, there would be a clear signal to the agent. Yet the concept is still much richer than the pure IAR approach, since it requires the situation in which the players follow the directive to be an equilibrium.

And besides the intuition, there is another advantage of the correlated-equilibrium concept over the Nash equilibrium: it requires less demanding assumptions. Besides the standard assumptions about individual preference relations, Nash equilibria require the strong assumption of common knowledge of rationality (CKR), i.e. the fact that all players are rational and know that all other players are rational as well and so on. The self-referential process is not only empirically incorrect (Gintis, 2009), it is also implausible. [8]

The correlated equilibrium as a solution concept does not require CKR but only rational players that share the same priors about state space of the game. This assumption is still demanding, but firstly it can be justified on evolutionary grounds (Gintis, 2009), and secondly, the assumption of CKR is far more problematic.

The true usefulness for the analysis of institutions, is, however, the fact that the correlative device in an epistemic game captures the most important aspect of an institution identified above: it coordinates the behavior and the expectations of the players and helps them to settle on equilibria they could otherwise no have achieved.In this interpretation, institutions yield behavior that is part of an equilibrium in the sense that the people involved give mutually best responses and no one has an incentive to deviate from current practice. [9]

Yet, there are also some open questions: firstly, there is still no mature theory of institutional change within in the framework of correlated equilibria. Secondly, but related to point one, institutions are not always optimal: this refers in particular to the transition of instrumental to ceremonial institutions. When and why do people accept correlating devices that stabilize certain power relations rather than solving a situation effectively? While the concept of quasi-parameters

[7] I will not go into the particular details of the concept here, see e.g. Gintis (2009) for further information.

[8] This distinguishes it from the consistency assumptions on individual preferences that, at least to a certain degree, can be adapted to the concrete situation under study, and can be justified on evolutionary grounds (Gintis, 2009).

[9] Concerning the concrete functioning, an institutions works through a mediated downward effect on the players beliefs leading to a certain action (Hedoin, 2012, 03). This conceptual trick avoids the typical problems arising from a mind-body problem like difficulty that arises if one assumes causal effects between different ontological layers. In any case, Hedoin (2012, 03) seems to be not entirely correct, as he sometimes defines institutions as correlated equilibria (page 338), and sometime, more correctly, as correlating devices.

(Greif & Laitin, 2004, 4) may be used to explain this fact, such an application has not been accomplished yet.

Thirdly, the concept of a correlated equilibrium in its pure form does not help to explain how institutions can solve social dilemmata. Even if there was a correlative device issuing the directive to the players to cooperate, this would not be best response, thus an unstable situation, not compatible with the notion of an institution.

Conjecture 2.1. *An epistemic dilemma game $\mathscr{G}^E$ with $s_i = \{c,d\} \forall i$ and the common payoff structure of the PD has no correlated equilibrium at mutual cooperation.*

Proof. The most accessible case would be if the correlative device issues the directive to the mutual cooperation to all players. A correlated equilibrium at (C,C) would require that $\mathbb{E}(\pi_i(c_i,c_{-i})) \geq \mathbb{E}(\pi_i(d_i,c_{-i}))$. But by the definition of the dilemma game $\pi_i(d,c) > \pi_i(c,c)$. □

A solution could be to consider repeated epistemic games. In such a case, if the discount parameter is high enough, and players play a TFT strategy in contrast to simple cooperation, an equilibrium of mutual cooperation could emerge. But there is not yet any work that has accomplished this. Also, one may think of sanctioning mechanisms that discourage defection. But in case sanctioning is costly, the dilemma structure still remains at the level of how the sanctioning mechanism gets stabilized.

2.2.3 Summary of the IAE approach

Wrapping up, the IAE approach seems to be more diverse, but also richer than the IAR approach. It captures the property of institutions that they assure a consistency among agent's expectations, and their behavior. We have also seen, however, that a conception of institutions *as* equilibria in the literal sense is too restrictive. Rather, the most promising conception of institutions considers them as *means to achieve* certain equilibria in games. The corresponding concept is that of a correlating equilibrium and a correlative device. This approach is, however, not yet in its maturity: there are no analytical in-depth studies of dilemma-like situations. And it is particular this kind of situations that requires social institutions. But this gap in the literature will - at least partly - be addressed below. Secondly, the approach is methodologically quire narrow: game theory seems to be the exclusive tool to address institutions. This comes with a qualification, however: A considerable part of the work is to identify the concrete nature of the correlative device. This requires one to look besides games theoretic concepts. And we will see later in this paper, how the approach can benefit from the application of computational approaches.

The third problem of the approach is that there is no viable theory of institutional change in the context of correlated equilibria. Within the context of the classical Nash equilibrium, however, Greif and Laitin (2004, 4) illustrates well how institutional change could be modeled, and their work will be an important source of inspiration to develop a similar concept for the context of correlated equilibria.

		Person B	
		Not use	Use
Person A	Not use	0 0	1 2
	Use	2 1	1 1

Figure 1: The private property game.

2.3 Institutions as nonphysical elements

If we consider broader definitions originating from the IAE approach we encounter a definition offered by Avner Greif, who defines institutions ”as a system of human-made, nonphysical elements - norms, beliefs, organizations, and rules - exogenous to each individual whose behavior it influences that generates behavioral regularities” (Greif & Laitin, 2004, 4, p. 635).

Here, institutions are understood as a system of elements generating behavioral regularities. The ontological statement underlying this definition contrasts sharply with Calvert’s account: while Greif perceives institutions as elements exogenous to the individuals, Calvert stresses that institutions are just a combination of behavior that is the result of individual rationality.

To illustrate the different degrees of explanatory power, consider the well known *property rights game* illustrated in figure 1 (Hindrinks & Guala, 2015).

This game considers a situation in which two persons have to decide whether they are using a particular piece of land or not. If both want to use the same piece of land, they will end up fighting and not use the land purposefully. If both abstain from using the land, the opportunity was missed by both. The best solutions are the situations in which one players uses the land and the other one abstains from using the land. These situations are Nash equilibria. But the interesting question would be which of the Nash equilibria emerges. This is where pure game theory cannot give an answer.

The task for the researcher would be to derive the concrete incentives (Greif’s ”system of human-made, nonphysical elements”) that make the agents play the one equilibrium or the other. It is hereby not sufficient to identify certain rules, like ”never cheat in a prisoner’s dilemma-like situation”, but one has to identify the mechanisms that would sustain such a rule. Put differently: why do agents follow this rule?

To frame the question more pronounced in the context of the PD: what kind of mechanisms could make a situation of mutual cooperation a stable equilibrium of the game? This is partly an empirical questions and it has been found out that reputation mechanisms were found to be effective in ensuring mutual cooperation in iterated PD (Greif, 1994).

2.4 Institutions as systems of shared beliefs and constitutive rules

Greif's defintion of institutions as "systems of human-made, nonphysical elements" shares the fundamental ingredients of Hodgsons definition of "institutions as systems of established and prevalent social rules that structure social interactions." (Hodgson, 2006). This suggests potential way of formalizing the rich ideas of original institutionalists such as Hodgson, with the formal tools of game theory. Before identifying the missing parts of such a formal theory, I will build upon the argument of Hindrinks and Guala (2015) who argue for a convergence of the formal conception of institutions as equilibria and the philosophical (and institutionalist) concept of institutions as constitutive rules.[10]

According to Searle (2005, 1), an institution is a system of constitutive rules. Constitutive rules are regularized rules of the form *X counts as Y in context C*. The use of such rules is very particular to human beings. An obvious example of how constitutive rules are used is money (Searle, 2005, 1): *A piece of paper issued by the European Central Bank counts as a medium of exchange in the Eurozone*. Hindrinks and Guala (2015) develop the concept further by distinguishing between a *status rule* and a *base rule* where the former specifies what is means to have a certain status (e.g. to be counted as money), while the latter specifies what it *requires* to have a certain status (e.g. when something is counted as money). Using these two concepts, it can be shown that the correlative devices in an epistemic game can be translated into constitutive rules (Hindrinks & Guala, 2015).[11] Combined with the concept of *shared intentionality* according to which a constitutive rules can get accepted and practiced by many members of a social group, one ends up with a philosophically well grounded theory of institutions.

The above discussed game theoretic concepts seem to be able to formalize these theoretical ideas. But while the conceptualization of institutions as particular correlation devices captures a large part of what is considered to be an essential part of institutions, there are some missing aspects that I will address in the next section.

3 Taking stock

We have discussed three fundamental approaches to model institutions: the IAR represented the starting point for the formalization of insitutions. While it is instructive in many instances, it put insufficient weight on the question of *why* people follow institutions. This has been taken up by the IAE approach. This approach has been found to be much more diverse than IAR: Considering institutions as literal equilibria has been found to be theoretically and empirically unsound. A more promising way was to consider as institutions as *means* that yield equilibria in games. The concept of a correlated equilibrium was found to be of particular theoretical

[10] The only disagreement I have with Hindrinks and Guala (2015) is that I think institutions must be considered as correlating devices *leading to* equilibria, rather than institutions as equilibria themselves.

[11] Sometimes such a translation is not required as the correlative device already has the form of a constitutive rule.

appeal.[12] In the proceeding section we have seen that the the concept of correlated equilibria is a natural formalization of the philosophical-institutionalist concept of institutions as shared systems of beliefs or, more precisely, constitutive rules.

But there are a number of open questions that should be addressed in a general conception of institutions:

1. The phenomenon of institutional change has not yet been adequately dealt with in the context of the correlated equilibrium approach.

2. The institutional life cycle and the dichotomy between instrumental and ceremonial (i.e. suboptimal) institutions is a fundamental aspect of institutionalist theory. Any formalization must be able to capture these concepts. This has not yet been shown for the correlated equilibrium.

3. The correlated equilibrium is, in its basic form, a static concept. It would be desirable to have a dynamic concept.

4. While the assumptions for correlated equilibria are less demanding than for Nash equilibria, there are still quite demanding requirements concerning individual rationality. It would therefore be desirable to have a formal theory that is at least able to relax these assumptions and allows for boundedly rational, or even inductively and adaptive agents.

5. Simirlarly, it would also be desirable to have the possibility to consider the social embededness of individuals, rather than considering isolated individuals. This also concerns the problem structure themselves: agents do not face problem structures in an isolated way, but are usually required to deal with different problem structures at a time. Therefore, institutions are often not addressed on one particular problem structure, but may be addressing several problem structures at once.

While the first issue requires more attention and requires certainly a more elaborated theory than will be offered in this paper, the other points can be straightforwardly be addressed via computational methods. The appealing property of such a computational approach is that the different aspects may be considered as extensions to one basic concept, if this is required: it may not always be necessary to keep track of the interaction structure of agents, but if this is indeed is the case, than the framework used to conceptualize institutions should be able to cope with this challenge without changing its fundamental character. In the following section, I will show that this is the case for the computational approach advocated here.

[12] Note that this concept also allows to take up several insights of the IAR approach.

4 Reconceptualizing institutions

I will now theoretically describe how the points listed above can be accounted in a computational setting.[13]

The institutional life cycle and sub-optimal institutions Instrumental institutions are problem-solving and help agents to deal with a particular problem structure effectively. Ceremonial institutions are not effective at problem solving, but help to sustain the privileges of particular interest groups. Ceremonial institutions usually do not start as ceremonial institutions, but develop out of instrumental institutions during time. This has been one of the fundamental contributions of Veblen (1898). The formal implication of this theoretical construction is, of course, a change in the underlying payoff structure of an epistemic game played by the agents. But then a theory about why this does not immediately affects choices is required. Such a theory needs to be computational because it is about the concrete mechanisms underlying people's decision making.

A timetable for the events as they could occur in a simulation could be:

1. People agree to a choreographer at t_1 because it is an effective way to solve a, say, battle-of-the-sexes game (*BoS*). Because the mean of the probability distribution of which the directive is drawn is neutral, the outcome is egalitarian and efficient for both players.

2. Then either the mean of the distribution or the asymmetry of the payoffs may change incrementally over time so that the correlated equilibrium is not efficient any more. But because agents update their perceptions of the real world only once in a while, or, if the payoff differences are significant, most people stick to the correlative device which now becomes a ceremonial institution. Or it may be the case than sticking to the (asymmetric and unjust) equilibrium is still better to play any other strategy (this may be the case if the resulting game is of the *BoS* type, but with a more pronounced asymmetric payoffs.)

3. A new correlative device restoring the egalitarian payoffs is not straightforward to emerge, in particular, a certain minimum mass of followers is usually required and there is no obvious incentive for the first agents to switch to the new choreographer. The originally instrumental choreographer has become ceremonial and the development is not easily reversed because of the path dependent nature of the process.

A dynamic version of the correlated equilibrium In its analytic form the correlated equilibrium is a static concept. But institutions are not born out of nowhere. They emerge from the interactions among the agents in the population in a dynamic process. ABM allow to an-

[13] For a more general discussion of how agent-based simulations can be used to formalize institutionalist theory see Gräbner (2015a).

alyze the generative process and to ensure that the equilibria can indeed be reached:[14] agents evolve certain strategies via an evolutionary process where the worst agents copy the strategies of the best-off agents via some stochastic process, and, if one wishes to account for this fact, entirely new strategies may emerge due to random alteration of existing strategies. The sophistication of the strategies can be controlled for through the number of bits the agents have available to store the strategy: simple strategies for the PD are simply of the form ”DEFECT!”, ”FOLLOW-CHOREO”, or ”COOPERATE”, a little bit more sophisticated strategies are TFT or GRIM-TRIGGER, and even more sophisticated strategies are like TF2T.

Boundedly rational agents Standard game theoretic agents are making their decisions in a deductive way. This decision-making process is not always easily implemented in computational models. The reason is that optimization problems including the expectation about others behavior are usually ill-defined as the agents are referring to each other in an infinite regress. One solution is the use of evolutionary methods that yield optimal behavior through the mechanisms of selection and imitation.[15] As indicated in the paragraph before, the computational capacity of agents can be controlled for by modeling agents as finite state machines. Finite state machines are automaton models that are specified by the finite number of their potential states, and via transition rules that specify when the agent changes from one state into another. Strategies can be conveniently be encoded as finite state machines. The strategy TFT for example could be encoded as $\{0,(C,0,1),(D,0,1)\}$. The notation means: start in state 0 (encoded as $(C,0,1)$). State zero means that the agent cooperates, then remains in state zero if the other agent has cooperated as well and changes to state 1 if the other agent has defected. If the agent is in state 1 (encoded as $(D,0,1)$) then the agent defects, and changes into state 0 if the other agent has cooperated, and stays in state 1 if the other agent has defected. The number of possible states an agent can be in is a measure for her computational capacity: more states mean that the agent can evolve more sophisticated strategies, that are usually more effective in achieving high payoffs. This also means that agents start making their decisions inductively. Such an account of agents allows to study the out-of-equilibrium dynamics of the model explicitly and to make sense of institutions as means to *reduce the complexity* of decision problems agents face. This is an acknowledged aspect of institutions (Elsner, 2014, p. 13), that can only be studied of the computational capacities if the agents are actually limited.

Embedded agents The assumption that agents interact with each other at equal probability (and thus assuming a complete interaction network) is often convenient. Sometimes, however, it

[14] Proofs for Nash equilibria are usually existence proofs. This means that they show that the equilibrium actually exists, but they by no means proof that the equilibrium is actually reachable by the agents of the relevant population. See Epstein and Hammond (2002), where the authors study a very simple game of spatial dynamics where almost all analytic equilibria cannot be reached by rational agents.

[15] This strategy also has theoretical support as rationality assumptions are often justified on the basis of exactly this evolutionary mechanisms (Friedman, 1966).

is necessary to consider the particular interaction structure among agents explicitly in the model because it matters that not all individuals interact with each other with equal probability.

In such a case it is adequate to study the behavior of the model with either idealized networks such as scale-free of core-periphery networks that represent ideal cases of empirical social networks, or, if one has access to the relevant data, with empirical interaction networks. In such cases, networks can be easily integrated into an ABM. To ensure a maximum of comparability among the different versions, the baseline models should be programmed with a explicit, complete graph structure. This is what in most analytic models is assumed implicitly. But implementing it explicitly makes the comparison among the different models easier and the isolation of the single mechanisms more straightforward.

When can the interaction structure become important? Consider for example of cases if we want to consider how different strategies spread over the population and certain agents are more likely to copy the behavior of some subset of the whole population. Or agents have different probabilities to interact with each other, how does this affect they way they choose their strategies? These questions are of course far beyond the simple games we normally consider. But there are situations in which such a level of complexity might be necessary.

Several problem structures Usually, game theory is used to model one particular interaction situation. This is extremely useful because it allows to focus on the mechanisms that are associated with this particular structure. But in other situations it can also be a poor choice: what if people face many different problem structures, and their cognitive limitations do not allow them to derive an optimal strategy for all (or any) of them? Such cases can be modeled with multiple games: agent play different games simultaneously. Such an approach is due to Long (1958, 3), and was recently picked up by Bednar and Page (2007). For some cases, in which the number of games and potential strategies is limited, analytical results can be obtained, provided no other additional mechanisms are considered in the model. But especially for the cases in which agents are heterogeneous or boundedly rational, computational methods are usually required.

5 Preliminary conclusions and open questions

The framework outlined above will prove useful in studying institutions and their evolution from a computational perspective and how such an approach can build upon existing conceptions of institutions. The usefulness of the approach stems from the fact that it can be finetuned concerning the different mechanisms that are actually considered in the model: in the simplest case it is just a dynamic and generative version of the correlative equilibrium approach, embedded into evolutionary-institutional theory. Such a conceptualization is often adequate and allows a focus on a few relevant mechanisms and parameters. But thanks to the modular structure of (object-oriented) agent-based models, the framework can be expanded to situations in which the agents are increasingly heterogeneous, interact with each other in a structured way, face different problems structures at once or adapt their strategies inductively to their environment. By adding complexity stepwise to the model using different modules allows a clear isolations of how different mechanisms contribute to the model outcome.

For further work, it would be desirable to first establish analytical proofs for further reduced version of the model and then to move on to the more complicated computational models. In order to keep the latter as tractable as possible, the first aim of the computational model should always be the replication of the mathematical results obtained before. Such a strategy is a reasonable disciplining strategy to avoid overparametrization of the simulation model and ensures the identification of particular mechanisms and their mutual influences.

Finally, this will result in a general computational approach to study institutions where the latter are understood as particular networks of decision making algorithms. These algorithms represent the very fundamental computations carried out by any human being when making a decision. Building the concept of the institution on such a fundamental aspect of economic activity is both scientifically appealing and highly promising.

References

Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). The colonial origins of comparative development: an empirical investigation. *American Economic Review*, *91*(5), 1369–1401.

Acemoglu, D. & Robinson, J. A. (2002). The political economy of the kuznets curve. *Review of Development Economics*, *6*(2), 183–203.

Acemoglu, D. & Robinson, J. A. (2006). De facto political power and institutional persistence. *American Economic Review*, *96*(2), 325–330.

Alchian, A. (1950). Uncertainty, evolution and economic theory. *Journal of Political Economy*, *58*, 211–221.

Aumann, R. J. (1987). Correlated equilibria as an expression of bayesian rationality. *Econometrica*, *55*(3), 1–18.

Bednar, J. & Page, S. (2007). Can game(s) theory explain culture?: the emergence of cultural behavior within multiple games. *Rationality and Society*, *19*(1), 65–97.

Bowles, S. [S.]. (2004). *Microeconomics. behavior, institutions and evolution*. New York: Princeton University Press.

Bowles, S. [Samuel] & Gintis, H. (2002). The inheritance of inequality. *Journal of Economic Perspectives*, *16*(3), 3–30.

Bravo, G. (2011). Agents beliefs and the evolution of institutions for common-pool resource management. *Rationality and Society*, *23*(1), 117–152.

Calvert, R. L. (1995). Rational actors, equilibrium, and social institutions. In J. Knight & I. Sened (Eds.), *Explaining social institutions* (pp. 57–93). Ann Arbor: University of Michigan Press.

Commons, J. R. (1924). *Legal foundations of capitalism*. Clifton: Augustus M. Kelley Publishers.

Dopfer, K., Foster, J., & Potts, J. (2004). Micro-meso-macro. *Journal of Evolutionary Economics*, *14*(3), 263–279. doi:10.1007/s00191-004-0193-0

Elsner, W. (2012). The theory of institutional change revisited: the institutional dichotomy, its dynamic, and its policy implications in a more formal analysis. *Journal of Economic Issues*, *46*(1), 1–43.

Elsner, W. (2014). Social economics and evolutionary institutionalism today. theoretical components and heterodox convergence in a socio-economic perspective. *Forum for Social Economics*.

Elsner, W., Heinrich, T., & Schwardt, H. (2014). *The microeconomics of complex economies*. Amsterdam et al.: Elsevier.

Epstein, J. M. & Hammond, R. A. (2002). Non-explanatory equilibria: an extremely simple game with (mostly) unattainable fixed points. *Complexity*, *7*(4), 18–22.

Friedman, M. (1966). The methodology of positive economics. In M. Friedman (Ed.), *Essays in positive economics* (pp. 3–16, 30–43). University of Chicago Press.

Gintis, H. (2009). *The bounds of reason. game theory and the unification of the behavioral sciences*. Princeton, NJ: Princeton University Press.

Gräbner, C. (2015a). Agent-based computational models - a formal heuristic for institutionalist pattern modelling? *Journal of Institutional Economics*. doi:10.1017/S1744137415000193

Gräbner, C. (2015b). Formal Approaches to Socio Economic Policy Analysis - Past and Perspectives. *Forum for Social Economics*. doi:10.1080/07360932.2015.1042491

Greif, A. (1994). Cultural beliefs and the organization of society: a historical and theoretical reflection on collectivist and individualist societies. *Journal of Political Economy*, *102*(5), 912–50.

Greif, A. & Kingston, C. (2011). Institutions: rules or equilibria? In N. Schofield & G. Caballero (Eds.), *Political economy of institutions* (pp. 13–44). Berlin: Springer.

Greif, A. & Laitin, D. D. (2004). A theory of endogenous institutional change. *American Political Science Review*, *98*, 633–652.

Hedoin, C. (2012). Linking institutions to economic performance: the role of macro-structures in micro-explanations. *Journal of Institutional Economics*, *8*, 327–349.

Hindrinks, F. & Guala, F. (2015). Institutions, rules, and equilibria: a unified theory. *Journal of Institutional Economics*, *11*(3), 459–480.

Hodgson, G. M. (2006). What are institutions? *Journal of Economic Issues*, *40*(1), 1–25.

Long, N. E. (1958). The local community as an ecology of games. *American Journal of Sociology*, *64*, 251–261.

Myrdal, G. ((1954)2004). *The political element in the development of economic theory*. New Brunswick, New Jersey: Transaction Publishers.

North, D. C. (1990). *Institutions, institutional change and economic performance*. Cambridge University Press.

Nowak, M. A. (2006). *Evolutionary dynamics. exploring the equations of life*. Cambridge, Massachusetts; London, England: The Belknap Press of Harvard University Press.

Ostrom, E. (1990). *Governing the commons: the evolution of institutions for collective action*. New York, NY: Cambridge University Press.

Rodrik, D. (2008). Second-best institutions. *American Economic Review*, *98*(2), 100–104. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=eoh&AN=0966397&site=ehost-live

Searle, J. R. (2005, June). What is an institution? *Journal of Institutional Economics*, *1*, 1–22.

Veblen, T. (1898, July). Why is economics not an evolutionary science? *The Quarterly Journal of Economics*, *12*(4), 373–397.

Williamson, O. E. (1979). Transaction-cost economics: the governance of contractual relations. *Journal of Law and Economics*, *22*(2), 233–261.