Humboldt-Universität zu Berlin

Seminar 7010910:

History of Economic Thought in the 20th century

Coasian reasoning and mental models in D. C. North's theory of institutional economics around 1990

Author:
Paul Boës
Immatriculation Number:
576106

April 16, 2017

Abstract. In this essay, I am concerned with the evolution of Douglass C. North's theory of institutional economics. In particular, I focus on the way in which external ideas from North's intellectual environment have entered his thinking around the period 1990-93, when he published his most influential book, *Institutions, Institutional Change and Economic Performance*, and received the Nobel Prize in Economics. Following a historical survey of the general trend of his thinking during his professional career, I argue that two such ideas, the resource-oriented reasoning that underlies Ronald Coase's argument for his influential theorem about transaction costs, as well as the theory of mental models from cognitive science, have entered North's theory around that period both explicitly and, more interestingly, implicitly, in such a way that his view of institutions influenced his theory of cognitive agents, and also his knowledge of mental models influenced his view of institutions. I also briefly discuss the fact that the nature of this implicit influence on North's theory nicely illustrates a general dynamic of how researchers form new knowledge, that is of interest to the history of economic thought.

1. Introduction

With more than 120000 citations, Douglass Cecil North is one of the most influential economists of the last century (Google Scholar, 2017). Together with Robert W. Fogel, he received the Nobel Memorial Prize in Economic Sciences in 1993, "for having renewed research in economic history by applying economic theory and quantitative methods in order to explain economic and institutional change" (Nobel Media AB, 2017b). As an economic historian, he is strongly linked with the theory of institutions and has spent most of his professional career in this field. One remarkable fact about his work on the subject is that one can fairly clearly identify a general trend in the development of his theories of institutional economics. This development arguably culminated in the period around the publication of his *Institutions, Institutional Change and Economic Performance* (IIE) in 1990, with over 50000 citations by far hist most influential book, and his receiving of the Nobel Prize three years later.

In this essay, it is my aim to provide an improved understanding of the content of North's theory in that period.[1] I want to do so by analyzing his presentation of some of the central aspects of his theory in the context of the economic and scientific currents of the time. In particular, I will argue that his view of the relationship between two such central aspects -the behavioral and the institutional mechanism of an economy- was strongly influenced by two ideas that he took, at very different points in time, from his intellectual environment and then applied to his theory by means of "reasoning by

[1]When talking about "North's theory" without further qualification, I will in the following always refer to the theory of institutional economics that he presented in his writings between 1990 and 1993.

analogy": The first is the reasoning that underlies Ronald Coase's famous argument for the property right allocations in a world without transaction costs, i.e. the "Coase theorem". The second is North's acquaintance with the theory of mental models from cognitive science, some time between 1990-92.

The structure of this essay is as follows: I will briefly introduce the notion of an institution and its role in the new institutional economics, as well as survey the development of North's theory of institutional economics in the course of his professional career, focusing on the broad trend that led up to IIE (Sec. 2). Following a more detailed presentation of my thesis (Sec. 3), I will then argue for the latter, focusing first on the reasoning behind Coase's theorem (Sec. 4) and then on the theory of mental models (Sec. 5). Finally, I will use the findings of those sections with respect to their interest for the history of economic thought more generally (Sec. 6), before concluding.

2. Insitutions and the Coase Theorem

What are institutions? In the first sentence of IIE, North defines them as follows: "Institutions are the rules of the game in a society or, more formally, are the humanly devised constraints that shape human interaction" (North, 1990, p. 3) In his Nobel Prize lecture, he describes as them as "the incentive structure of a society and [...], in consequence, the underlying determinant of economic performance". From these characterizations, it becomes clear that institutions form a very broad class of things. Examples would be the basic codes of human conduct that can be found in every society ("Thou shalt not kill"), specific cultural rule sets or religious beliefs ("Never contradict an elder"), but also higher-level structures such as the different laws of custom or inheritance that exist in different countries ("When do you own a piece of land?") and specific traditions such as vassalage in feudalism. That institutions exist and affect people's daily lives is clear, but how are they, as North claims, the "determinant of economic performance"? The answer to this lies in the "Coase Theorem", which is not so much a formal theorem, but rather an argument that Ronald Coase 1960 made in his seminal article *The Problem of Social Cost*.[2] The argument considers the efficiency of property right allocations in the absence of transaction costs. A rough statement of it is the following:

> **Coase Theorem:** Given a set of producers and some set of properties, assume that different producers can extract different value from having the property right over some property. Moreover, assume a world of zero transaction cost, that is, bargaining for and exchange of property rights is free. Then, the final allocation of property rights is socially optimal regardless of the initial allocation of property rights. Conversely, if the world has non-zero transaction costs, then the final allocation of property rights does depend on the initial allocation of property rights and is, in general, not socially optimal.

[2]Coase did not himself refer to his argument as a theorem, in fact it was Stigler in his 1966 textbook *The Theory of Price* who coined the term.

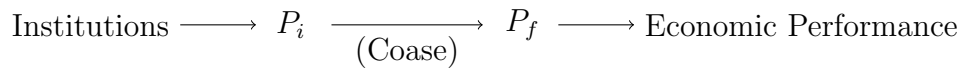

FIGURE 1. By the Coase Theorem, in a world with non-zero transaction costs, the institutions determine a state's economic performance, because there is no convergence in going from the initial property right allocations, P_i to the final ones, P_f.

The theorem makes two statements. One concerns the sufficiency, the other the necessity of zero transaction costs for optimatliy of the final property right allocation. Coase, in his 1960, only discussed the first statement. His argument was along the following lines: In a world without transaction costs, trading a property right from an inefficient producer to a more efficient producer is always a pareto-improvement, because by exchanging the right to the property in exchange for a sum that lies in between the value produced by the inefficient producer and that produced by the efficient producer, both parties always increase their wealth.[3] Moreover, it is easy to see that by iteratively exchanging property rights in this way, one necessarily reaches a final state that is both pareto-optimal and socially optimal.

The argument for the converse statement also is straightforward: If there is an additional cost associated to every transaction, then it might not be the case anymore that there is a price for the property right such that an exchange of property right from the inefficient holder of the right to the more efficient produces a net gain in wealth for both parties (in particular, this will be the case whenever the gap between the values created by the two producers is smaller then the transaction cost). Hence, in such a world the final property right allocation will, in general, not be socially optimal because the final property right do not lie with those producers that extract most value from that property.

It is this second, necessity relation between property rights and transaction costs, that provides the explanation for why North considers institutions to be the determinants of economic performance: *De facto*, we live in a world with (massive) transaction costs. Therefore, the initial distribution of property rights does matter a lot for the their final distribution. Now, the final distribution of property rights determines the economic performance of a state. At the same time, the initial distribution of property rights is determined by the institutions, in the above sense, of the society of this state. Hence, institutions determine the economic performance (see Fig. 1).

It is important for my argument in Sec. 4 to emphasize the "if and only if"- nature of the Coase Theorem: institutions matter *only* in a world with non-zero transaction costs. If there were no transaction costs, then there would still be institutions, and these institutions would still determine the initial property right allocations. However, because, by the Theorem, the final property right allocation would be independent of the initial

[3]Of course, this implies that the initial property right allocation matters very much for the final distribution of *wealth* but that is not the concern of the Coase Theorem

one, the economic performance of a state would be optimal *regardless of the institutions in that state.*

Of course, Coase's argument doesn't settle the details of the mechanisms through which institutions determine economic performance, nor which institutions are particularly important. For example, there can be both efficient and inefficient institutions. Efficient institutions are such that, *despite* the existence of transaction costs, they determine the initial property right allocations in such a way that the final allocation socially optimal. To answer questions of this type is the *raison d'être* of new institutional economics as a field, as well as North's works in particular.

I will now present a brief survey of North's theoretical development, because my discussion of this theory around 1990-93 makes sense only in light of this development.

2.1. **The road to non-rationality: A survey of North's development.** North received both his undergraduate and graduate education at the University of Berkeley, after which he took up his first position as an assistant professor at the University of Washington in Seattle. Two occasions that brought him to the East Coast have, according to himself, had a major impact on his professional development (Nobel Media AB, 2017a): The first was a Social Science Research Fellowship in 1951-52 for his dissertation, during which he attended a sociology seminar under Robert Merton and which lead to Joseph Schumpeter having a strong influence upon him. The second was a position as a research associate at the National Bureau of Economic Research in Boston in 1956-57, which brought him in contact with the leading economists of the time and put him at the center of the then nascent school of "cliometrics" in the field of history of economics. Cliometricians, or new economical historians, took a more quantitative approach to economic history, applying economic theory to historical data (see (Williamson *et al.* , 2008) for an introduction).

Another important turning point for his research focus was his decision, in 1966-67, to move from working on American to European history. Of this, he says that "[r]e-tooling turned out to change my life radically, since I quickly became convinced that the tools of neo-classical economic theory were not up to the task of explaining the kind of fundamental societal change that had characterized European economies" (Nobel Media AB, 2017a). His search for those tools in the then still young field of new institutional economics led to two books, published in 1971 and 1973 respectively: *Institutional Change and American Economic Growth* (together with Lance Davis) and *The Rise of the Western World: A New Economic History* (together with Robert Thomas). These books mark the beginning of a trajectory in North's theoretical development that was to last for over twenty years. In them, North and his co-authors applied neo-classical models with the usual rationality assumption about economic agents to problems in European and American economic history. Moreover, institutions in North's theory were assumed to be efficient, in the sense that they would produce the economically optimal outcome.

This last assumption, however, clashed with the historical findings, in particular the historical stability of long-run poor economic performance, that is, the stability of inefficient institutions (ibid.).

Aware of this discrepancy, North continued to develop his theoretical framework, a new version of which got published almost ten years later as *Structure and Change in Economic History* in 1981. There, North abandoned the notion that institutions were efficient, without however dropping the neo-classical assumption of rationality from his theory. The result was a theory in which the existence and stability of inefficient institutions could be explained neo-classically. In short, in this model rulers devised property rights in their own interests, leading to inefficient but stable institutions. However, while this new model as such could provide a partial answer to the problems that confronted North's original model from the early 70s, what it could not explain was how the above inefficient institutions would not be driven out by competitive economic pressures: Yes, the political power of rulers could enable them to impose rules that were in their own interest, but from a neo-classical perspective even they would have, at some point, bow to the laws of the market.

North set out again to further develop his model, so that it could also accommodate this latter problem. The result, again almost ten years later, is IIE. Here, to solve the above problem, he drops both the efficiency of institutions, as well as the rationality assumption as one of the central tenets of neo-classical theory: "The analytical framework is a modification of neo-classical theory. What it retains is is the fundamental assumption of scarcity and hence competition and the analytical tools of micro-economic theory. What it modifies is the rationality assumption. What it adds is the dimension of time" (Nobel Media AB, 2017c). Within this framework, North's answer to the above problem then becomes the following:

> The answer hinges on the difference between institutions and organizations and the interaction between them that shapes the direction of institutional change. Institutions [...] determine the opportunities of a society. Organizations are created to take advantage of those opportunities, and, as the organizations evolve, they alter the institutions. The resultant path of institutional change is shaped by (1) the lock-in that comes from the symbiotic relationship between institutions and organizations [...] and (2) the feedback process by which human beings perceive and react to changes in the opportunity set. (North, 1990, 7)

It is the second aspect of this explanation, the feedback process between human beings and the changes in the opportunity set, that is responsible for North's decision to drop the rationality assumption. As will be discussed in more detail in the next section, North's idea here is that the differing perceptions of human beings, *qua* economic agents,

Year	Publication	Efficiency of Institutions	Rationality Assumption
1970-73	*Inst. Change. /Rise of WW*	Yes	Yes
1981	*Struct. & Change*	No	Yes
1990	*IIE*	No	No

TABLE 1. Development of North's thinking in his major books: Over the years, he increasingly weakened his assumptions about institutions and the rationality of agents.

prevents the economic pressures to enter the market in such a way that they could drive out inefficient institutions.

This, then, was the state of North's theory around the time that he got awarded the Nobel Prize. What this brief survey shows is that it took him almost three decades of research to find a model that balanced his ambitions to generality as a cliometrician, while also withstanding some basic empirical scrutiny. As we have seen, the trajectory of North's thinking led him to successively weaken the neo-classical assumptions of his theory, as summarized in Table 1. It is with this trajectory in mind that I now turn to my thesis.

3. Question and Thesis

As stated in the introduction, in this essay I want to analyze North's theory around the period 1990-93 in the context of the economic and scientific currents of the time. My question is: Can we identify ideas stemming from those currents, that have found entrance into his theory, in order to better understand the content of this theory? My answer here will be in the affirmative. I want to argue that one can identify at least two such ideas and that both of them have entered his theory both explicitly and implicitly. By explicit, I here mean that he states and uses them under their name as part of this theory. By implicit, I mean that he alters his thinking and presentation of *other* elements of the theory by drawing analogies between those elements and the ideas.

Let me be more specific: In the following section, I will argue that not only does North explicitly use Coase's theorem to argue for the relevance of institutions, he also uses it implicitly, namely when he applies the exact same argumentative structure that was used in the "derivation" of the theorem, to argue for the relevance of the perceptions and values of individual human agents in order to explain the stability and longevity of inefficient institutions. Here, tit is the reasoning behind Coase's theorem that is the first of the two ideas that North has picked up from his intellectual environment to apply them, implicitly, to his theory. Then, in the next section, I will argue that North's acquaintance with the theory of mental models from cognitive science, some time 1990-1992, has not only found explicit entrance into his work, but has moreover entered his theory implicitly, by altering his stance on the nature of *institutions* themselves. The notion of a mental model then is the second of the above ideas.

Given the evident influence of North's ideas as he held them around that time, I believe that the identification of such implicit usages of external ideas is instructive and valuable both for an evaluation of North's theory as well as from the perspective of the history of economic thought.

4. From institutions to cognition: Coasian thinking in North's 1990 theory of cognitive agents

In IIE, North drops the rationality assumption from his theory. He does so in order to be able to explain the empirical fact that economic competition does not force the extinction of inefficient institutions that are devised by political rulers (and other stakeholders in power positions) in their own interest. It is the aim of this section to argue that North makes implicit use of the reasoning in Coase's theorem to argue why it is necessary to drop the rationality assumption to be able to explain this fact. My argument will proceed as follows: First, I will reiterate North's explanation of the stability of inefficient institutions and clarify how the assumption of non-rational behavior of human agents plays into it. Secondly, I will show that North considers the dropping of rationality *necessary* to understand the stability of such stable inefficient institutions and present his argument for this view. Thirdly, I will show that this argument is exactly analogous to the one given by Coase, hence making my point.

Let us first recall from the survey in Sec. 2, that in his (1981), North had dropped the assumption that societies would produce efficient institutions, in order to explain the existence of inefficient institutions by means of a neo-classical model, in which inefficient institutions would exist because agents in power positions could enforce inefficient institutions that worked to their advantage. What had left him unsatisfied, however, was the fact that his model could not explain the *stability* of these inefficient institutions, because according to his original model the competitive economic forces would have to force away such inefficient institutions.

As we have heard in the long quote of that section, North's answer in IIE to that challenge are two dynamical mechanisms that are the result of the fact that a given set of institutions always determines an opportunity set, that is, a set of problems or resources that are induced by the institutions of a state. These dynamical mechanisms are: (a) The opportunity set drives the creation of organizations that are specialized to exploit this opportunity set (If you introduce red traffic lights, there will be artists specialized to perform at those traffic lights, etc.). According to North, since those organizations depend, for their existence, on the persistence of the opportunity set that they are created to exploit, their reluctance to allow for the change of institutions will introduce frictions to the institutional dynamics and consequently stabilize inefficient institutions; (b) Human beings perceive the changes in the opportunity set differently. For instance, they might not always agree on the direction or amount of change. Consequently, this lack of agreement between human agents makes it difficult for the economic forces that

are imposed on institutions by means of the activity of those human beings *qua* economic agents, to exert the pressure that would be necessary to drive inefficient institutions out.

It is the second part of North's explanation, (b), that relates to the dropping of the rationality assumption. I will therefore here focus on it. How exactly is (b) related to the rationality of agents? To North, the rationality assumption of neo-classical theory involves not only the assumption of Bayesian rationality, to always make the best decisions based on the prior distribution that one holds, but also the (implicit) assumption that the human agents hold full information of the situation, i.e. that their priors coincide with the actual distribution. In North's words:

> More controversial (and less understood) among the behavioral assumptions, usually, is the implicit one that actors possess cognitive systems that provide *true* models of the worlds about which they make choices, or, at the very least, that the actors receive information that leads to convergence of divergent initial models. (North, 1990, 17, emph. original)

It is the absence of this convergence in the last sentence of this quote, that provides the crucial link between the rationality assumption and North's explanation in (b): By definition of what North considers to be the rationality assumption, dropping it is necessary for (b) to obtain, because two rational agents could never differ in their perceptions of the changes in the opportunity set.

Now, even if one grants that (b) does provide an explanation of the stability of inefficient institutions, one can ask whether North could not do without it. That is, is an inclusion of the differing perceptions of human beings really necessary to allow the existence of such institutions? According to North, this is the case (ibid., Ch. 3). It is essentially because, if one maintained the rationality of agents, and with it the idea that the models that human agents have of their environment would converge to a single, objective model, then this agreement of all the models of all humans in a society would effectively implement exactly *that* competitive economic power that drives out inefficient institutions and that North, following his (1981), did *not* find to pertain empirically. In other words, North argues that human beings can exert the economic pressure that is necessary to drive out inefficient institutions *if and only if* they are taken to be rational in the above sense. In making this argument, he quotes the following passage from (Simon, 1986, S210f.) in both IIE and his Nobel Prize lecture, of which I here reproduce those passages central for my argument:

> [I]f we postulate an objective description of the world as it really is, and if we assume that the decisionmaker's computational powers are unlimited, then two important consequences follow. [...] Second, we can predict the choices that will be made by a rational decisionmaker entirely from our knowledge of the real world and without a knowledge of the decisionmkaer's perceptions or modes of calculation [...]

> If, on the other hand, we accept the propositions that both the knowledge and the computational power of the decisionmaker are severely limited, then we must distinguish between the the real world and the actor's perception of it and reasoning about it [...].
>
> The rational person in neo-classical economies always reaches the decision that is objectively, of substantively, best in terms of the given utility function.

With respect to the existence of stable but inefficient institutions, the implication here is that, if agents were rational, then regardless of the existence of transaction costs, no inefficient institutions could withstand the rational decision making power of all the economic agents. As such, dropping rationality is not merely a convenience, but a necessity if one aims to explain the existence of stable but inefficient institutions.

It is at this point, that I want to argue for the presence of implicit Coasian thinking in North's argument. Recall from Sec. 2, that this theorem, in brief, said that the final property right allocation will be unique if and only if the transactions costs are zero. In exact analogy, we can read off of Simon's quote and the way North uses it, the following argument: For a cognitive agent, her final internal representation of the external state of the world will be unique (namely the true representation of this world) if and only if the computational power is infinite (or, equivalently, the computational cost is zero). The schema of the statement here is the following: "A final state will be unique and independent of an initial state if and only if the available resources for the intermediate process are infinite". In this schema, the resources are wealth (if you have infinite wealth, then effectively your transaction costs are nil) and cognitive processing power, respectively. The initial states are initial property rights and input information, the final states final property rights and the final knowledge of the external world. Neither North (nor Simon) make the parallel of their arguments to Coase clear at any point, but, stated in the above fashion, the parallel, I take it, undeniable.

It is interesting to return to Fig. 1 in light of this parallel. Recall that this figure shows how, given the Coase theorem, institutions determine economic performance by determining the initial property right allocations within a society. We can understand, in light of the above analogy, North's argument for the necessity of dropping the rationality assumption in order to explain the existence of stable but inefficient institutions, also as a further qualification to the *Coasian* argument why institutions matter in the first place. What I mean here is the following: We have not in fact yet clarified the mechanism by which institutions determine the initial property right allocations, that is, the leftmost arrow in the figure. Instead, we have only said that they do. Indeed, with respect to this particular mechanism, the neo-classical economist might smell the chance for two possible arguments to argue for the re-instantiation of the neo-classical economical results, the ones that predict the establishment of the socially optimal final allocation, in spite of the

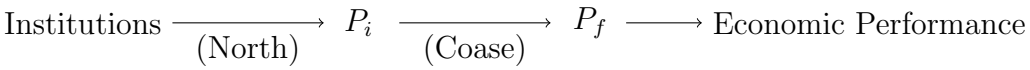

FIGURE 2. North's reasoning as re-evaluating the question of how institutions determine economic performance (compare with Fig. 1): By arguing in exact analogy to Coase's own original argument, North argues that there cannot be a convergence in the initial property right allocation, P_i, from institutions, because human beings are in fact non-rational, that is, they do not have access to infinite computational resources.

existence of transaction costs. The first of those would be that institutions are efficient, that is, they develop in such a way that they produce the optimal allocation even if there are transaction costs. This argument had been rebutted by North already in his 1981, as we saw. The second is that, while institutions are not efficient, it may be that the human agents on the basis of whose decisions property right allocations are determined, are rational, so that there would be a convergence to the optimal set of allocations, even if the underlying institutions themselves are not efficient. North's argument for the necessity of dropping the rationality assumption in IIE then effectively is a rebuttal of this second argument!

In light of this parallel, the analogy between the Coase's and North's argument becomes even stronger. This is because the analogy turns from an analogy of argumentation to an analogy of *usage*. North uses his argument, which works along the lines of Coase's to also analogously argue for the same thing: Institutions determine economic performance because we live in a world in which *both* computational power is costly and so is transacting (see Fig. 2).[4]

5. From cognitions to institutions: Mental Models in North's thinking after 1990

The second part of my thesis is completely independent of the first. It is the aim of this section to argue that North's acquaintance with the theory of mental models, some time 1990-92, has found explicit and implicit entrance into his theory. The explicit part of this is not up to debate: He references and discusses the theory of mental models in several publications after 1992. Indeed, he even has a publication devoted exclusively to it

[4]There is an interesting research question, that is closely related to the present one: How did the notion of there being limits to human cognition enter into the economic thinking of the new institutional economists at all? One possible explanation is that it happened in the course of early computer science, where the metaphor of human brains as machines was of course very common. Given that many of the founding fathers of computer science also worked on economical problems, maybe most eminently John von Neumann and Simon, it would be easy to see how this metaphor should have entered the field. A completely different explanation, in contrast, would consider the idea of limited cognition as stemming from a Kantian theme. Here, looking at the influence of Kantian ideas in the Austrian School, for instance through Schumpeter and Hayek, could be a promising starting point. For instance, North admitted that Schumpeter had been a strong influence on him (Nobel Media AB, 2017a), although IIE contains no references to either in the relevant passages. Of course, this question goes far beyond the scope of this essay.

(Denzau & North, 1994). The less trivial part of my thesis is to argue for the implicit part. Here, I will argue that his acquaintance with the theory of mental models has enabled North to draw explicit analogies between institutions and the cognitive structures of human beings in a way that he could not have done without knowing about the theory of mental models. The resulting analogy, in turn, has altered North's understanding of what institutions are.

I will proceed as follows: First, I will briefly introduce the notion of a mental model, and provide some evidence that North has come across the notion only after the publication of IIE, however, before 1992, when he first made use of it. Then, I will show how North uses the notion of a mental model in his Nobel Prize lecture to characterize institutions in a novel way, hence making my point.

Mental models are a conceptual tool from cognitive science with which the thought processes of cognitive agents are modeled. In particular, they are meant to describe the internal representations that cognitive agents formulate of their external environment in order to solve tasks that require their interaction with this environment. The basic notion of a mental model originated in the context of early research in behavioral child psychology in the 1920s in Paris, in particular with the work of Georges-Henri Luquet (1927), however they experienced a revival during the wave of behavioral psychology of the 80s, also in the USA, in particular the influential (Johnson-Laird, 1983).

North himself probably came across the theory of mental models in (Holland *et al.* , 1986), some time between 1990 and 1992. This follows, because the term "mental model" occurs, with an explicit reference to the latter book, in his *Transaction Costs, Institutions and Economic Change* from (1992). That he knew about the theory before 1990, when IIE got published, is unlikely, because he would surely have used the concept in that book, which he does not. Instead, there, North uses the term "mental construct" which he also at no point defines properly. This contrasts with a much more frequent use of the term "mental models" in publications after 1992.[5] It is difficult to not interpret this change in frequency as expressing the fact that North felt that his theory of cognitive agents had gained scientific vindication by its proximity to the theory of mental models and that this gave him more confidence in using the idea also in his theory.

What exactly is the relationship between the two theories? The key passage that North seems to have taken from Holland *et al.* is the following, which he also quotes in (Denzau & North, 1994) to introduce mental models:

> "[w]e believe that cognitive systems construct models of the problem space that are then mentally "run" or manipulated to produce expectations about the environment. (Holland *et al.* , 1986, 12)

[5]Publications, in which I have found North to use the mental models terminology are (North, 1992, 1993, 1996a,b; Denzau & North, 1994; Nobel Media AB, 2017c)

The aspect of mental models that seems to have been of particular importance to North, is that, in the theory of mental models, these models are used to help cognitive agents deal with uncertainty about the actual state of the external environment (ibid., 13ff.). This obviously ties in very well with North's idea that the central conceptual novelty in dropping the rationality assumption from economic theories is that economic agents are not in possession of infinite computational resources and, consequently, true models of the environment. To North, one consequence of the considerations that were discussed in the last section is that the particular mechanisms by means of which agents deal with uncertainty have a direct impact on the economic performance and dynamics of a country, and hence need to be understood in order to build better theories of the latter: "The analytical framework we must build, must originate in an understanding of how human learning takes place" (Nobel Media AB, 2017c). The theory of mental models, with its explicit distinction between an external state of affairs and their internal mental representation, must have felt, to North, like a great discovery in the process of building this understanding.

I now turn to provide some evidence that learning about mental models not only gave North a particular scientific theory to reference in his works, but also altered his thinking about institutions themselves. The clearest exhibition of this is in the following quote from his Nobel Prize lecture:

> The relationship between mental models and institutions is an intimate one. Mental models are the internal representations that individual cognitive systems create to interpret the environment, institutions are the external (to the mind) mechanisms individuals create to structure and order the environment (ibid.).

North here rather strikingly offers a direct analogy between mental models and institutions. But how do mental models alter his view of institutions, as I claim? It is not, that North starts to think about institutions as mechanisms that reduce uncertainty, as can be seen, for instance by the following quote from IIE, that is, before he learned about mental models: "Institutions exist to reduce the uncertainties involved in human interaction" (North, 1990). The novelty here is rather of a more subtle nature: Institutions appear, in this quote, as the "mental models" of the external world, that is, as literally *organically* growing instruments of humans, devised to reduce uncertainty about economic and social activity, in just the same way in which actual mental models reduce uncertainty about reaction of the external world to my own actions. Of course, this literal metaphor goes beyond the institutions themselves: Economic activity, for instance, becomes analogous to the neural activity of the brain, societies one big brain. North admittedly does not explicitly carry out this analogy that far, but it is inviting, and one thing is clear: Without his acquaintance of the theory of mental models, North would not even have had the vocabulary to articulate the analogy in the fashion of the above quote. It is in this way,

I submit, that learning about mental models has altered North's view of the nature of institutions.

6. The evolution of North's thinking: How new ideas are incorporated into existing knowledge

In the preceding two sections I have argued that two different ideas, one from economics, the other from cognitive science, have entered North's theory and played an important role in determining the content of this theory, both explicitly and also, more interestingly, implicitly. While the two ideas, the resource-oriented reasoning from Coase's Theorem on the one hand and the theory of mental models on the other, are completely independent of another, the order in which they entered North's thinking and the way in which they have been used is interesting, I think, from the point of view of the history of economic thought: Chronologically, it is clear that the resource-oriented reasoning behind Coase's theorem, in terms of transaction costs, was known to North (as well as H. Simon, whose quote also included the computational power metaphor of human cognitive processing) for more than two decades when he used it in IIE to argue for the necessity of dropping the rationality assumption from economic theory. It must thus have been deeply ingrained into his thinking. It is, as such, not very surprising that he (as well as Simon, who was also trained as an economist (Nobel Media AB, 2017d)), should have had a predisposition to read and interpret the new ideas from cognitive science, when they entered into the arena of economics during the 80s, through the lens of this resource-oriented vocabulary.

At the same time, we have seen that his acquaintance with mental models, in the process of developing a deeper understanding for the contents of the theories of cognitive science, ultimately lead to a re-thinking of his *economics*, i.e. his view of the nature of institutions. This, I think, is a nice illustration of how input from external fields is being processed in the course of research more generally: At the beginning, when novel ideas form outside one's field enter the arena, they are interpreted in light of the existing knowledge structures that have been created in the course of developing expertise in one's original field. However, with increasing understanding and familiarity of those novel ideas, they will themselves alter the knowledge structures of the core expertise. In any case, a more detailed of this process for North's thinking lies beyond the scope of this essay and it is intended here rather as a closing remark.

7. conclusion

In this essay, I have been concerned with the evolution of D. North's theory of institutional economics. In particular, I have focused on the way in which external ideas from North's intellectual environment have entered his thinking around the period 1990-93, when he published his most influential book, *Institutions, Institutional Change and Economic Performance*, and received the Nobel Prize in Economics. Following a historical survey of the general trend of his thinking during his professional career, I have argued

that two such ideas, the resource-oriented reasoning that underlies Ronald Coase's argument for his influential theorem about transaction costs, as well as the theory of mental models, have entered his theory around that period both explicitly and, more interestingly, implicitly. Finally, I have briefly discussed the fact that the nature of this implicit influence on North's theory nicely illustrates a general dynamic of how researchers form new knowledge, that is of interest to the history of economic thought.

References

Coase, R. H. 1960. The Problem of Social Cost. *The Journal of Law and Economics*, **3**(10), 1–44.

Davis, L. E., & North, D. C. 1971. *Institutional Change and American Economic Growth*. Cambridge: Cambridge University Press.

Denzau, A. T., & North, D. C. 1994. Shared Mental Models: Ideologies and Institutions. *Kyklos*, **47**(1), 3–31.

Google Scholar. 2017. *Douglass C. North - Google Scholar Citations*. `https://scholar.google.com/citations?user=-LcMZqMAAAAJ&hl=en` (last access: 12.4.17).

Holland, J. H, Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. 1986. *Induction: Processes of Inference, Learning, and Discovery*. Cambridge, MA: MIT Press.

Johnson-Laird, P. N. 1983. *Mental Models*. Cambridge, MA: Harvard University Press.

Luquet, G.-H. 1927. *Le Dessin Enfantin*. Paris: Alcan.

Nobel Media AB. 2017a. *Douglass C. North - Biographical*. `http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/1993/north-bio.html` (last access: 12.4.17).

Nobel Media AB. 2017b. *Douglass C. North - Facts*. `http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/1993/north-facts.html` (last access: 12.4.17).

Nobel Media AB. 2017c. *Douglass C. North - Prize Lecture: Economic Performance through Time*. `http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/1993/north-lecture.html` (lase access: 13.4.17).

Nobel Media AB. 2017d. *Herbert A. Simon - Biographical*. `http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/1978/simon-bio.html` (last access: 15.4.17).

North, D. C. 1981. *Structure and change in economic history*. Norton.

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

North, D. C. 1992. *Occasional Papers: Transaction Costs, Institutions, and Economic Performance*. 30 edn. San Francisco: International Center for Economic Growth.

North, D. C. 1993. What do we mean by rationality? *Pages 159–162 of: The Next Twenty-five Years of Public Choice*. Dordrecht: Springer Netherlands.

North, D. C. 1996a (Dec.). *Economic Performance Through Time: The Limits to Knowledge*. Economic History 9612004. EconWPA.

North, D. C. 1996b (Dec.). *Economics and Cognitive Science*. Economic History 9612002. EconWPA.

North, D. C., & Thomas, R. P. 1973. *The Rise of the Western World*. Cambridge Books, no. 9780521290999. Cambridge: Cambridge University Press.

Simon, H. A. 1986. Rationality in Psychology and Economics. *The Journal of Business The Behavioral Foundations of Economic Theory*, **59**(2), 209–224.

Stigler, G. J. 1966. *The theory of price*. New York: Macmillan.

Williamson, S. H., Lyons, J. S., & Cain, L. P. (eds). 2008. *Reflections on the cliometrics revolution : conversations with economic historians*. London; New York: Routledge.