Framing the ‘social brain’

Challenges and lessons from philosophy of mind and AI for cognitive neuroscience.

BSc AI Thesis

Thymen René Wabeke

Supervisors:
Mark Blokpoel & Iris van Rooij
Department of Artificial Intelligence,
Donders Institute for Brain Cognition, and Behaviour,
Radboud University Nijmegen

12th Sepetember, 2012

Abstract

There is a growing idea that our brain to a large extent has been shaped by the cognitive demands made by novel and complex social tasks. In this context, a common term used to refer to the structure or function that has evolved to reach such demands is the ‘social brain’. This term is used by scientists in several different research domains (e.g. evolutionary psychologists, neuroscientists and cognitive scientists). This thesis distinguishes four notions of the social brain and introduces a corresponding conception for each notion. It seems important that scientists are clear about which conception they are researching. Furthermore, the thesis investigates what cognitive architectures (modular versus abductive) are consistent with a functional social brain. It turned out that every type of perspective on the social brain either seems to be conceptually infeasible or computationally intractable. In the context of research on the social brain it seems important to be explicit about what kind of cognitive architectures one is adopting. After all, adopting one or the other architecture can have fargoing consequences for how one interprets existing findings. Also, adopting a specific architecture commits a cognitive scientists to addressing a different set of theoretical challenges that will need to be overcome to be descriptively adequate and computationally feasible.

Contents

1 Introduction

Many scientists are researching a ‘social brain’ (e.g. Baron-Cohen, Leslie, and Frith (1985); Dunbar (1998); Brothers (1990a)). However, the term is not always used in the same sense. Evolutionary psychologists often refer to the ‘social brain hypothesis’ when mentioning the ‘social brain’, whereas neuroscientists use the term as a reference to brain regions. In the context of research on the ‘social brain’ it seems important that scientists are clear about what they are researching. Before scientists can be clear about what notion they adopt in their research, clearness on the different meanings of the term ‘social brain’ seems necessary.

This thesis investigates in what senses the term ‘social brain’ is used in todays neuroscientific literature and in this way provides clearness on the different notions of the social brain. In Section 2 four possibly related though conceptually distinct notions of the social brain are distinguished. The functional notion of the social brain, which is a reference to all brain functions related to social cognition, is adopted as a basis for the conceptual analysis of this thesis.

The subsequent research is about cognitive architectures that may be consistent with a functional social brain. There has already been research on cognitive architectures (e.g. modularity and abduction).[1] Thinking about how this knowledge can possibly be applied to a social brain is an interesting topic, because different architectures have different assumptions and therefore other theoretical and explanatory benefits and drawbacks. The different assumptions of architectures may influence the conceptual and computational viability of a social brain perspective, so this thesis also investigates what cognitive architectures can support a functional social brain.[2] The research is based on two properties, namely informational encapsulation and domain specificity. Both properties are interesting, because they have influence on the conceptual and computational feasibility of a social brain perspective. Section 4 points out that each combination of properties seems to have its own conceptually and computationally benefits and drawbacks. Also, a cognitive architecture for each combination of properties that seems consistent is proposed. The two architectures that will be discussed —modularity versus abduction— are described in Section 3.

The research in this thesis distinguishes from Adolphs’ (2009) approach in two ways. Firstly, Adolphs tries to investigate whether social cognition is special or not, whereas this thesis focusses more on the conceptual and computational benefits and drawbacks of different types of perspectives on a functional social brain. However, the properties that Adolphs uses in his research, ‘selectivity’ and ‘functional specialization’, seem related to some of the properties used in this thesis, namely ‘specificity’ and ‘general purpose’. Selectivity is about “the level of the domain of information that is being processed”. This is close to specificity, because a specific system only responds to inputs of a particular domain. Adolphs uses the second property, functional specialization, as a degree to what extent modules are exclusive. This is about whether modules are used exclusively for social tasks, or whether the module is also used for several distinct tasks. A system is general purpose if it responds to many different inputs. General purpose seems related to specialization, because a system that responds to different kinds of inputs is perhaps not exclusively used for a specific task. Secondly, the subsequent research in this thesis differs from Adolphs’ approach, because it uses the functional notion of the social brain as a basis, whereas it seems that Adolphs takes into account the hardware notion, for instance, by mentioning “neural substrates of social cognition” (Adolphs, 2009).

This thesis has two major scientific contributions. Firstly, Section 2 provides clearness

[1] A modular architecture refers to an architecture based on the modularity theory proposed by Fodor (1983). An abductive architecture is non-modular in the sense of informationally unencapsulated and having an organization that allows it to fixate beliefs via abduction (e.g. Peirce, Weiss, and Hartshorne (1974); Haselager (1997); Fodor (2000) describe abduction). Both architectures are described in Section 3.

[2] In this thesis ‘support’ means that a particular cognitive architecture is consistent with a system. For example, suppose that a modular architecture is consistent with the auditory system. This means that modularity supports the auditory system.

on the different notions of the social brain by distinguishing four different senses of the term. Furthermore, the results in Section 4 show that it is important for cognitive neuroscientists to be explicit about which of the two architectures they are adopting, e.g., in the context of research on the social brain. After all, adopting one or the other architecture can have fargoing consequences for how one interprets existing findings. It is also shown that adopting a specific architecture commits a cognitive scientists to addressing a different set of theoretical challenges that will need to be overcome to be descriptively adequate and computationally feasible.

2 Scrutinizing the 'social brain'

Today there is much research on a 'social brain' (e.g. Baron-Cohen et al. (1985); Dunbar (1998); Brothers (1990a)). However, this term has different meanings and is used by scientists with diverse backgrounds (e.g. evolutionary psychologists, neuroscientists and cognitive scientists). Not all scientists are aware of these different senses and some of them even seem to use different senses interchangeably. For instance, Gallagher and Frith (2003) mention the Theory of Mind and several other "key brain regions believed to comprise the 'social brain' and their role in the development of this ability". By mentioning terms like 'brain regions' the authors seem to talk about the 'social brain' from a neural or hardware perspective. At the same time, they also suggest, by mentioning properties of modularity like 'possibly dedicated' and 'domain specific', that these regions correspond to functional mechanisms that are modular. In the context of research on the 'social brain' it is important to be clear about what is meant with this term. This section distinguishes four possibly related though conceptually distinct notions of the 'social brain'. The functional notion of the 'social brain' is adopted as a basis for the subsequent research in Section 4.

One prominent use of the term social brain is in the context of the 'social brain hypothesis' (e.g. Dunbar (1998); L. Barrett and Henzi (2005)). This hypothesis is often used by evolutionary psychologists and explains the extraordinary size and complexity of the human brain. It is well known that primates have large brains compared to other mammals of equivalent size. Chimpanzees and bonobos, for instance, are species close to human in evolution, but have a brain that is only 25-35% of the size of the human brain (Adolphs, 2009). According to the 'social brain hypothesis' the enlarged brain size is the result of complex social environments. In his paper, Dunbar (1998) claims that "The social brain hypothesis implies that constraints on group size arise from the information-processing capacity of the primate brain, and that the neocortex plays a major role in this." In short Dunbar argues that when group size increases, the social complexity also increases and thus there must be more resources to manipulate information. This results in growth of the cortex. The 'social brain hypothesis' argues about the origin of a social brain, it explains why humans have a large brain size, rather than explaining how its functionality is performed. This thesis refers to this notion of the social brain as the 'Complexity-conception'.

In the literature there is also a sense of the social brain that is about difference in performance on false-belief tasks between people with autism and without (Baron-Cohen et al., 1985). A false-belief task is about recognizing that others can have beliefs about the world that are diverging. This ability to infer others' mental states is known as the Theory of Mind (ToM). Previous research shows that people with autism are not able to succeed in false-believe tasks. In experiments by Baron-Cohen et al. (1985) for instance, 80% of autistic children failed on false-belief tests. Baron-Cohen et al. suggest that the difficulties that autistic people have with false-belief tasks are the result of limitations of their ToM. However, this is not the only view on ToM. According to Gerrans (2002) "It is the absence of some or all these early abilities in autism which deprives autistic subjects of a crucial developmental resource and gives the misleading impression that the essential difference between autistic and normal subjects is at a higher level: a module concerned with social cognition." This brain function, Theory of Mind, is often conceived as the social brain. This thesis refers to this notion of the social brain as the 'Disabilities-conception'.

The social brain is defined by some scientists as all brain functions related to social cognition (Adolphs, 2009). ToM is often assumed to be a part of the social brain, but there are more brain functions related to social cognition. Examples of such functions are the processing of faces (Kanwisher, McDermott, & Chun, 1997) and the detection of people who cheat on social contracts (Cosmides & Tooby, 1992). Adolphs (2009), for instance, investigates whether social cognition is "in any sense specialized for processing social information or whether social cognition is just like cognition in general, only applied to the domain of social behavior." To some scientists, the social brain is a reference to all brain functions related to social cognition. This thesis refers to this notion of the social brain as the 'Functional-conception'.

According to Brothers (1990a) it is reasonable to believe that the growth and specialization of the human cortex —earlier described as the 'Complexity conception'— "stamped upon brain function and therefore accessible to investigation at the neural level: in effect, the attempt to relate growing knowledge about primate social cognition to neural activity opens up a new area for brain research." Brothers uses the term 'social brain' as a reference to brain regions whose activity has been found to correlate with social tasks, and she does not mean the functional conceptualization of social tasks (e.g. cognitive architectures). Research in this context has for instance been done by Happe et al. (1996). They focus on ToM and suggest that "a highly circumscribed region of left Medial Prefrontal Cortex is a crucial component of the brain system that underlies the normal understanding of other minds". This thesis refers to this notion of the social brain as the 'Hardware-conception'.

The previous paragraphs described four notions of the term social brain. The conceptions that were introduced are summarized in Table 1. Though the conceptions differ, they also seem to be related. Section 3 describes two cognitive architectures that may be applicable to the functional social brain. It is possible that adopting a particular architecture has an effect on the 'Hardware-conception' of the social brain. For example, if the social brain is functionally modular, it may also have an individuated physical architecture. This is an interesting topic, but it is not in the scope of this thesis.

The conceptual analysis of this thesis uses the 'Functional-conception' as a basis, because this is the most cognitive conception. Assuming that there is a functional social brain, what cognitive architectures can support it? This is investigated especially with respect to conceptually and computational feasibility. The next section provides background information about two architectures —modular versus abductive– that may be consistent with the social brain.

#	Name	Description of the conception	Consistent references
1	Complexity	The social brain as an explanation of the evolutionary growth of the cortex in primates related to group size (i.e. social complexity).	Dunbar (1998); Allman (1999); L. Barrett and Henzi (2005)
2	Disabilities	The social brain as an explanation of the difference in performance on false-belief tasks between people with autism and without.	Baron-Cohen et al. (1985); Gerrans (2002)
3	Functional	The social brain as a reference to all brain functions related to social cognition (often implicitly written about as dedicated or specialized, e.g., ToM-module).	Baron-Cohen et al. (1999); Adolphs (2009)
4	Hardware	The social brain as a reference to brain regions whose activity has been found to correlate with social tasks (e.g. Medial Prefrontal Cortex, Superior Temporal Sulcus, ACC, amygdala, anterior singulate).	Brothers (1990a); Baron-Cohen et al. (1999); Happe et al. (1996); Adolphs (2009)

Table 1: Four different conceptions of the notion 'social brain'.

3 Cognitive architectures of the social brain

Section 4 investigates cognitive architectures that may support a functional social brain. Before the outcomes of this research are presented, one needs to know something about the architectures that will be discussed. There has already been research on the question whether ToM is modular or not (e.g. Gerrans (2002); Leslie (1991)). Because of the discussion about a possibly modular ToM, this thesis investigates whether the social brain as a whole can be consistent with a modular architecture. If it turns out that this is not the case, a non-modular architecture should be adopted to the social brain. A prime example of a non-modular architecture is an architecture that has an organization that allows it to fixate beliefs via abduction. In this section the characteristics of modularity and abduction are discussed. This section also describes why the modularity properties informational encapsulation and domain specificity are used as a basis for the conceptual analysis of this thesis.

3.1 Modular architecture

In ‘The Modularity Of Mind’ Fodor (1983) describes a theory of perception and cognition that has had a lot of impact in cognitive neuroscience and started a discussion about modularity (e.g. Pinker (1997); Fodor (2000); Carruthers (2003)). Some scientists strongly believe that the whole brain consists of modules (i.e. the ‘Massive Modularity thesis’ of Cosmides and Tooby (1994)), while others doubt if Fodorian modules can exist at all (Prinz, 1998). The theory proposed by Fodor does not contain a strict definition of modularity (Coltheart, 1999). What he actually did was arguing that modularity is marked by a set of psychologically interesting properties described in Box 1. It is possible that a system is modular to the extent that it exhibits these properties. In the next paragraphs informational encapsulation and domain specificity are explained in detail, because these properties are used in the conceptual analysis of this thesis.

Box 1: Modularity properties
Fodor (1983) proposed nine psychologically interesting properties that can be used to characterize a modular system. These properties are:

- Domain specificity: Modules only operate on certain kinds of inputs.
- Mandatory operation: Modules operate in an automatic way.
- Limited central accessibility: Higher levels of processing have limited access to the representations within a module.
- Fast processing: Modules are quick in generating outputs.
- Informational encapsulation: Modules do not need information of other psychological systems (for example from higher levels of processing) in order to operate.
- ‘Shallow’ outputs: Modules have relatively simple outputs.
- Fixed neural architecture.
- Characteristic and specific breakdown patterns.
- Characteristic ontogenetic pace and sequencing

Encapsulation is an interesting properties for various reasons. Firstly, encapsulation can have influence on the computational tractability (Fodor, 1983). The global inferences an unencapsulated system can make may increase the computational complexity. Secondly, this thesis shows that encapsulation also relates with other modularity properties like mandatoriness and speed. So the encapsulatedness has also influence on other properties. Finally, in more recent writings Fodor (2000) has treated encapsulation as a necessary property for a system to be modular. The fact that encapsulation is an important property —if you accept Fodors reasoning— is another reason why encapsulation is an relevant property. Specificity is a property of relevance as well, because the degree to which a social brain only responds to inputs from a social domain may have influence on the conceptual viability of that perspective. For example the question is considered whether one can call a completely domain general system a *social* brain at all.

An informationally encapsulated system cannot access information stored elsewhere when processing a given set of inputs (Fodor, 1983, 2000). Such a system can only use the information contained in those inputs, plus whatever information might be stored within the system itself (for example, grammar in the case of language). The opposite, unencapsulation means "having complete access to a person's expectations, beliefs, presumptions or desires" (Fodor, 1983). It is not necessary that the system uses all the information in the brain. The fact that it has complete access and thus potentiality can use all information already means that the system is unencapsulated.

Fodor believes encapsulation is a necessary property for modularity. He uses the persistence of perceptual illusions as an argument for this claim. The classic illustration of encapsulation is the Müller-Lyer illusion (Müller-Lyer, 1889) where the two lines continue to look as if they were of unequal length even after one has convinced oneself otherwise (Figure 1). If perception were not encapsulated, global inferences to background information could be made. These inferences would result in a corrective judgement and the illusion would go away. This is not the case, so according to Fodor the Müller-Lyer illusion is an argument why some modules are encapsulated.

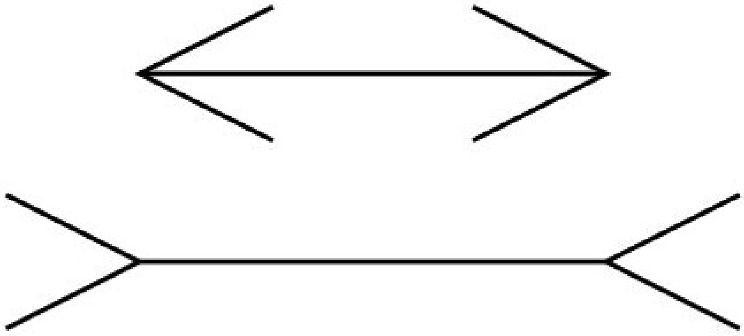

Figure 1: The Müller-Lyer illusion is a classical example of encapsulation where two lines do not seem to be of the same length. If perception were not encapsulated, then the illusion would go away as soon as the corrective judgement is formed. To Fodor this is an argument why some modules are encapsulated.

Other properties, such as mandatoriness and speed, follow from informational encapsulation. Mandatoriness means that a system is insensitive to the utilities of the organism and automatically starts operating when a relevant stimulus is represented. Recall that encapsulation limits modules in accessing information, they cannot make inferences to information in other systems. Because a mandatory process cannot be influenced by other processes (for example an explicit desire), the information that enters the system is also limited. This proves mandatoriness is in some way related to encapsulation. When there is no encapsulation and there are no other assumptions a system may become slow. This is because encapsulation limits the information that enters the system which could result in reducing the informational load. If there is no encapsulation a system has access to all information in the brain. The global inferences such a system can make may increase the computational complexity and may thus also influence the speed of processing.

Another important property of modularity is domain specificity. Domain specificity is hard to define, because the terms domain and specific are vague (Prinz, 1998). What do we call a domain? When is a module specific? Consider a language module. This cognitive system is domain specific in the sense that it has something to do with one particular domain, namely language. But inside this big module there are subsystems that are even more specific. Think for example of a subsystem for phonetic-analysis and lexical-form. Why do such subsystems not deserve to be called modules? Research on this topic has been done by Block (1995). He argues that a cognitive module (for example the language module) can be decomposed in a set of smaller modules. The question is when this decomposition stops. According to Block "Decomposition stops when all the components are primitive processors, because the operation of a primitive processor cannot be further decomposed into suboperations". However, this thesis uses the definition of Coltheart (1999) when referring to domain specificity. He proposed that a cognitive system is domain-specific if it only responds to inputs of a particular domain. Take for instance a face recognition module. If this module is domain specific it only responds to a visual representation of a face and not to, for example, a voice or a written word. General purpose refers to the opposite, namely a module that responds to (in theory) a boundless number of different inputs.

3.2 Abduction, a non-modular architecture

If the outcomes of the research presented in Section 4 show that the social brain seems inconsistent with a modular architecture, it needs to be supported by another type of cognitive architecture. The upcoming paragraphs describe an architecture that is non-modular in the sense of informationally unencapsulated and having an organization that allows it to fixate beliefs via abduction. Firstly, abductive reasoning in general is described. Later is shown how abduction is suitable to fixate believes and why abduction is unencapsulated.

Abduction was introduced by Peirce (e.g. Peirce et al. (1974)) and differs from two other well-known kinds of reasoning, namely induction and deduction. Induction means determining the rule based on numerous examples. For example, 'My pants are always dirty when I spill milk on it; Therefore, when I spill milk, my pants get dirty.' Deduction means applying a rule to a case. For example, 'All pants get dirty when you spill milk on them; I am wearing pants and spilled milk; Therefore, my pants are dirty.' Deductions do not determine new rules, they are tautological. Abduction means determining the precondition. Abduction infers the precondition that best explains a case. For example, 'When I spill milk, my pants get dirty; My pants are dirty; Therefore, I may have spilled milk.' The precondition 'I spilled milk' is not the only possible explanation. Abduction is often also referred to as inference to the best explanation (Haselager, 1997; Thagard, 2000), because the hypotheses generated by an abductive inference are such that they provide good explanations of the observations available.

Now, 'inference to the best explanation' will be explained using an example. Imagine that you wake up and notice that the weather is very windy. The night before you have parked your bike at the sidewalk, but unfortunately it has fallen on the ground. You conclude that the bike has fallen due to the wind. This hypothesis best explains the scene you are facing. However, it could also be a prank of the boy next door. Or a drunk person that bumped into your bike. The hypothesis that your bike has fallen due to the wind does cannot be concluded logically from the premises. Given partial information about the world (e.g. the weather is windy) and a set of hypotheses or candidate explanations (e.g. bike has fallen due to the wind), the hypothesis is inferred that best explains the information about the world. This type of inference is also called abduction.

In this thesis the term abduction can be understood as referring to "an inferential process that takes as input partial information about the world and generates as output hypotheses about which states of the world are believed to currently hold and which ones not" (Haselager, Dijk, & van Rooij, 2008). When for example the 'states' being hypothesized can include 'mental states' of other agents in the world, the abductive process is suitable to fixate believes and can be seen as implementing Theory of Mind in a non-modular way.

Abduction is non-modular, and in particular unencapsulated, in the sense that the information that is relevant to abduce the best hypothesis can, in principle, come from anywhere (Fodor, 2000). Inferences to expectations, beliefs and desires seem necessary when abducing the best hypothesis, because there is no a priori boundary to what information is relevant. Recall that unencapsulation means "having complete access to a person's expectations, beliefs, presumptions or desires" (Fodor, 1983). So, unless there is a local way without inferences to other cognitive systems to know what is relevant, abduction seems unencapsulated.

This thesis uses the term 'abductive architecture' when referring to an architecture that is non-modular in the sense of informationally unencapsulated and having an organization that allows it to fixate beliefs via abduction. There might be other ways of implementing a non-modular system apart from abduction. However, this thesis uses abduction as a prime example of a non-modular architecture, because to my knowledge there is no other non-modular architecture that may be consistent with a functional social brain.

4 Conceptual analysis

In Section 2 four notions of the social brain were distinguished. The conceptual analysis of this thesis is on the 'Functional-conception', which is a reference to all brain functions related to social cognition. This section investigates what cognitive architectures may support a functional social brain. This research is based on two properties, namely informational encapsulation and domain specificity. Recall from the previous section that both properties are interesting with respect to the computational complexity and conceptual viability of the social brain. For example, assuming an unencapsulated social brain means that the systems has access to information in other cognitive systems. This may result in computational intractability. On the other hand, one can doubt about the conceptual viability of conceptualizing the social brain as general purpose; What is *social* about a social brain that responds to many non-social inputs? Encapsulation and specificity lead to four combinations of properties (see Figure 2). This section shows that each combination of properties —'type of perspective' is used as a reference to a combination of properties— has its own conceptual and computational benefits and drawbacks.[3] Furthermore, for each combination of properties a cognitive architecture that seems consistent is proposed.

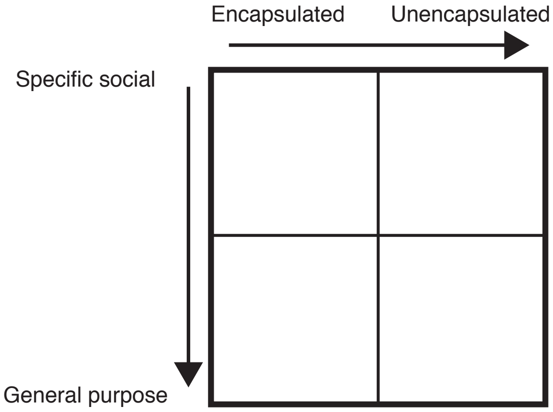

Figure 2: Every perspective on the social brain can be characterized by its relation to the modularity properties informational encapsulation and domain specificity. This section discusses the benefits and drawbacks of each combination of properties with respect to the conceptual and computational feasibility.

4.1 Encapsulated & Specific

Perspectives that are characterized by this combination of properties assume that the social brain only responds to inputs from a social domain (specific social) and that the social brain cannot make inferences to information stored in other cognitive systems (informational encapsulated). This combination of properties is consistent with a modular system described in Section 3.1. The benefit of modular systems is that they seem computationally feasible (Fodor, 1983). However, the following paragraphs show that there seem to be (at least) three theoretical issues assuming the social brain is encapsulated and specific.

To be modular, and thus to take benefit of the computational feasibility of the modularity theory, a system must either be one module or consist of a set of modules. When conceptualizing the social brain as encapsulated and specific it is not necessary that the system is one big module. In fact, it is likely that the social brain consists of a large number of modules since they probably perform more effectively and efficiently than one big module with more general

[3] The term 'perspective' is used to refer to a view on the social brain. Perspectives that share a combination of properties belong to the same 'type of perspective'.

functions (Pinker, 1997; Tooby & Cosmides, 1992).[4] If at least one 'part' of the cognitive architecture is non-modular the social brain as a whole cannot be modular any more. Now, an example of a part of the social brain that is non-modular is given.

The 'Theory of Mind' (ToM) is an example of a function that is part of the social brain (described in Section 2 as the 'Disabilities-conception'). Assuming that the social brain is modular, means that all parts must be modular as well. But actually ToM seems to be non-modular. According to Fodor (1983) only input and output systems are candidates for modularity. The title 'Modularity of Mind' might be a bit misleading but Fodor believes systems that are responsible for higher cognitive processes such as reasoning and ToM are non-modular. Gerrans (2002) also disagrees with the hypothesis of a modular ToM. He argues that ToM itself is an unencapsulated system that uses 'early' modular input systems that process social information. Examples of such inputs systems are mechanisms to detect emotional expression and goal directed behaviour. Gerrans argues that ToM makes global inferences by using involuntary signalling via facial expression and bodily posture as an example. Imagine a deceiver that is producing signals. A receiver that has a completely encapsulated ToM cannot make global inferences in order to, for example, override his trusting responses to signals faked by the deceiver. This would probably be unviable, because the receiver could be tricked over and over. To make this detecting process less vulnerable to deception, ToM is likely a central system that is unencapsulated and can make global inferences. According to Gerrans ToM is non-modular and learns to synthesize 'early' social inputs (e.g. detection of emotional expression and goal directed behaviour).

The doubt about a modular Theory of Mind perhaps already raises questions about assuming a modular social brain. But even when ignoring this argument, it is not evident that the social brain is encapsulated. The encapsulatedness of a system prevents it of having access to all information in the brain. However, in the case of a social brain this might be difficult, because one can talk about a lot of things in different subject domains (for example colour, physics, the taste of food, the meaning of life, etc.). Talking is generally considered to be a social task, but probably not all subjects domains are inside the social domain. According to Carruthers (2003) we can use cross-modular content, because we are capable of freely combining concepts across different domains. The problem is, however, how to access this content (Rice, 2011). An encapsulated system cannot make global inferences, so it seems that the social brain must at least have one part that is unencapsulated in order to retrieve concepts from other domains (and/or cognitive systems). The 'combining argument' is used as a reference to the apparent unencapsulatedness of the social brain when using corss-modular content.

Modular systems only respond to inputs from a particular domain (Figure 4). An example of such a system is the auditory system. This low-level system only responds to inputs in the sound domain —waves within a certain frequency range. The system does not respond

[4] An encapsulated system does not necessary have to consist of one module. In fact, there are many possible architectures that are encapsulated. For instance the two architectures in Figure 3. The boxes represent subsystems for which holds that they are informational encapsulated. A set of boxes form an encapsulated system (e.g. a social brain).

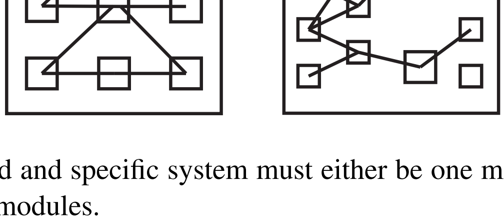

Figure 3: An encapsulated and specific system must either be one module or, as shown in the figure, consist of a set of modules.

to other inputs, visual stimuli for instance. Modular systems require some information about their inputs in order to filter inputs. In the case of the auditory system inputs are being filtered on the physical properties of the input itself. When the social brain is a modular system this means that by definition it only responds to inputs in a 'social' domain. But what is a social domain? Can it exist at all? Remember for example the combining argument, the fact that we can talk about anything. Furthermore, the reasoning about the auditory system cannot apply to the social domain. This is because in order to classify a social stimulus global inferences are necessary, since these social stimuli cannot be in- or excluded from the domain based on physical properties. For instance, when you hear someone talking you consider this as a social input. On the other hand, the sound of leaves in the wind are not interpreted as social inputs because they are not produced by an agent. In the case of the auditory system these two inputs belong to the same domain since the are both sound waves. In the case of a modular social brain the filtering of inputs seems to be a problem. Social inputs do not have specific (physical) properties that tag an input as social, so a social tag needs to be based on different information. One way to tag such input is to infer from its content and other knowledge and beliefs —because context is important for social relevance– whether the input is social or not. This inference, however, is not informationally encapsulated. Therefore, unless there is an alternative way of tagging inputs as social that is not based on inferences, this causes a conceptual problem, viz., a modular social brain seems to require non-modular inferences to filter inputs that do not belong to its domain.

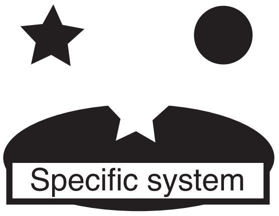

Figure 4: Specificity means that the system only responds to inputs from a particular domain. This can be explained using a receptor. The system represented above only responds to the star on the left. The circle on the right is an input from another domain and does not 'fit'. In the auditory system inputs are being filtered on the physical properties of the input itself. Inputs with other properties —for example visual stimuli— do not fit and cannot enter the system. Social inputs do not have specific properties, or more generally have some kind of class tag. In the case of a modular social brain the filtering of inputs seems to be a problem, because the tagging process seems to require global inferences.

As argued above, there seem theoretical issues (viz., the combining argument and defining social inputs) and inconsistencies (non-modular ToM) with assuming an encapsulated and specific social brain. Because of these issues, it seems that this type of perspective (top-left cell in Figure 2) is conceptually infeasible. How to resolve the issues? The following paragraphs show that choosing another combination of properties seems to solve the theoretical issues described above, but at the same time raises other conceptual or complexity questions.

4.2 Unencapsulated & Specific

Global inferences seem necessary in defining social inputs and the use of cross-modular content. These inferences are allowed when conceptualizing the social brain as unencapsulated. In this way the social brain has access to information stored in other cognitive systems. The following paragraphs show that conceptualizing the social brain as unencapsulated and specific seems to overcome explanatory limits of assuming the social brain is modular. However, the paragraphs also show that this type of perspective seems to have drawbacks with respect to computational tractability.

Firstly, the question is considered whether this type of perspective is conceptually feasible. To answer this question a face-recognition system is used as an example. We can derive from empirical observations that such a system is specific, because people with different visual impairments can, for instance, not recognize faces, but do recognize written words and vice versa (Coltheart, 1999). A face-recognition system is thus "not responding to inputs except those of a particular class", but is it encapsulated as well? Recall that unencapsulation means having "complete access to a person's expectations, beliefs, presumptions or desires" (Fodor, 1983). It is not explained here in detail how face recognition is performed. Although, it is not unlikely that during this recognition process some inferences are made to information represented by our expectations, beliefs, presumptions or desires. Because a domain specific system can at the same time be informationally unencapsulated, it seems that this type of perspective is thus conceptually feasible.

However, the important characteristic of an unencapsulated system is that its processes cannot be realized in an informationally encapsulated way and therefore central systems cannot be modular (Fodor, 1983). Because of the non-modularity perspectives that use this combination of properties have to engage another type of cognitive architecture. An example of a supporting architecture is abduction. In Section 3.2 is described that abduction is a reference to "an inferential process that takes as input partial information about the world and generates as output hypotheses about which states of the world are believed to currently hold and which ones not" (Haselager et al., 2008). When for example the 'states' being hypothesized can include 'mental states' of other agents in the world, the abductive process can be seen as implementing Theory of Mind in a non-modular way. Abduction overcomes many of the explanatory limits of a modular social brain. Assuming the social brain does abduction, this can explain how global inferences are made. For example when defining social inputs (recall the 'tagging process') and the use of cross-modular content. In the case of ToM it can explain the apparent informationally unencapsulatedness of the inferences people can make about other people's mental states, such as beliefs, desires and intentions. That being said, the abductive architecture of the social brain also has its own theoretical challenges. Notably, by virtue of being unencapsulated the possible inferences afforded by data are essentially boundless when the system tries to make inferences about other agents' mental states (H. C. Barrett & Kurzban, 2006). The boundless inferences seem computationally infeasible (i.e., intractable or more formally NP-hard) for minds/brains with bounded computational resources (Kwisthout, 2011; Ford & Pylyshyn, 1996; Haselager et al., 2008). These theoretical problems are notorious in the philosophy of AI and collectively referred to as the 'frame problem' (e.g. Fodor (2000); Haselager (1997)).[5]

Conceptualizing the social brain as unencapsulated and domain specific thus seems to solve some conceptual issues of assuming a modular social brain, but raises questions about the computational tractability.

4.3 Encapsulated & General Purpose

Assuming an encapsulated and general purpose social brain means that the system not only responds to social inputs, but to inputs from (in theory) a boundless number of domains. At the same time, the social brain cannot make inferences to information in other cognitive systems

[5] Historically the frame problem meant something much more specific, but it has come to take on a wider meaning as 'the problem of relevance' (Fodor, 2000; Haselager, 1997) Originally the frame problem was about the difficulties describing the effects of action in logic without explicitly specifying conditions that are not affected by an action or are intuitively obvious non-effects (Mccarthy & Hayes, 1969). A solution to avoid the frame problem is to take only into account the properties that are *relevant*. The difficulty is to define what is relevant. Also, a system that is informationally unencapsulated has access to all information in the brain. This means that there is no a priori boundary to what information is relevant. The possibly boundless inferences seem computationally infeasible (i.e., intractable or more formally NP-hard) for minds/brains with bounded computational resources (Kwisthout, 2011; Ford & Pylyshyn, 1996; Haselager et al., 2008).

because it is conceptualized as encapsulated. At first sight conceptualizing the social brain as general purpose may sound a bit peculiar. This is because a system that is defined as a system for brain functions related to social cognition appears to be contradictory if it also responds to inputs that are non-social. It is, however, an interesting question whether one can call a completely domain general system a *social* brain at all.

There are various arguments why it is likely that the social brain is independent, or dissociable, from general intelligence (Adolphs, 2009; Baron-Cohen et al., 1999). There are, for instance, people that are good in social tasks, but perform bad on non-social task (and vice versa). Secondly, Adolphs (2009) suggests that the social brain evolved independently because social behaviour makes demands that are very unique. Furthermore, disabilities like autism "can cause selective impairment in social judgement without any necessary loss to general problem-solving ability" (Baron-Cohen et al., 1999).

The arguments in the previous paragraph show that there are processes that are independent or dissociable from general intelligence. But to what extent are these processes really social? Recall the difficulties in defining social inputs described earlier. Social inputs can, for example, not be characterized by physical properties. A sound wave itself, for instance, does not say anything about who produced it (e.g. the wind or a social agent like a truck driver). Maybe this is because *social* is something from the outside. Something we, humans, invented to classify a type of behaviour. It is possible that inside the brain there functionally is no such thing as *social*. This might explain the difficulties in defining a social domain, because these inputs perhaps do not exist in the brain.

A general purpose system does not need explicit social inputs, because such a system responds to inputs from (in theory) a boundless number of domains. A general purpose system is thus more flexible since it responds not only to inputs from a social domain, but also to inputs from other domains. The question here is how such a system can deal with many different inputs. Now, is investigated what architecture can be consistent social brain that is conceptualized as encapsulated and general purpose (bottom-left cell in Figure 2).

Imagine a social brain that is general purpose and encapsulated as well, what does it look like? To my knowledge the only architecture that seems applicable to this combination of properties is the 'adaptive toolbox' (Prinz, 1998; Gigerenzer, Todd, & ABC Research Group, 1999; Newell, 2005; Gigerenzer, Hoffrage, & Goldstein, 2008). The adaptive toolbox is a set of fast and frugal heuristics suited to different problems (see Figure 5). Frugal means that the heuristic uses the least necessary amount of information. It is not the heuristics themselves that makes us smart, but the fact that a good heuristic is chosen every time. Depending on the input, and possibly some other assumptions, the good heuristic is chosen. In the case of a social brain that is conceptualized as general purpose, a particular set of heuristics can for example be used to infer others' mental states (ToM) while another heuristic is used to calculate the speed of a passing train. The question is how to chose the best heuristic.

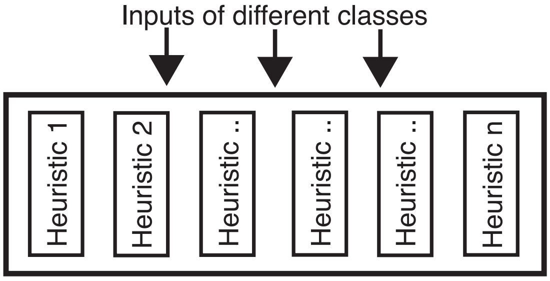

Figure 5: An'adaptive toolbox' consists of a set of fast and frugal heuristic. When making a decision the best heuristic is chosen. Such a system can be seen as general purpose in the sense that it is flexible and it is informationally encapsulated in the sense that each heuristic is 'frugal'. For different kind of inputs another heuristic can be chosen.

It is not unlikely that what is the best heuristic depends on the input. Another type of input will probably use another heuristic that is inside the toolbox. The relation between an input and the corresponding heuristic may be hardwired within the system. However, this does pose a conceptual challenge. If the relation between inputs and heuristics is hardwired the adaptive toolbox is not so flexible any more. One can even argue that such a toolbox has become specific, because a particular heuristic is always —not necessary exclusive— used for a particular domain of inputs.

There seems also another challenge that applies to both the hardwired and flexible variant of the adaptive toolbox. Since the best heuristic depends on the input, the system has to know something about it in order to chose the best heuristic. With respect to the social brain, this results once again in the difficulties of defining social inputs. In order to filter social inputs and in this way select the best heuristic it is not unlikely that global inferences are necessary. Due to the encapsulation these inferences seem to cause a conceptual problem.

At this point it is not evident that the social brain can be an encapsulated and general purpose system (bottom-left cell in Figure 2). Firstly, because on the 'Functional-conception' there does not seem to be anything *social* about a general purpose architecture. But even when ignoring this doubt it is hard to think of an applicable architecture. The one proposed, an 'adaptive toolbox' seem to raise conceptual issues. The research showed that assuming that the toolbox is hardwired means that particular input always uses the same heuristic. This will make this type of perspective specific instead of general purpose (top-left cell in Figure 2). On the other hand will a flexible toolbox result in an unencapsulated instead of encapsulated type of perspective (bottom-right cell in Figure 2). An unencapsulated and general purpose system is the only type of perspective that is not discussed yet.

4.4 Unencapsulated & General Purpose

A system that is unencapsulated and domain general (bottom-right cell in Figure 2) is the opposite of a modular system. The system has potentially access to all information in the brain and is not restricted to a particular domain of inputs. A social brain that is conceptualized as unencapsulated and general purpose has thus to engage a non-modular architecture, for instance abduction. Because of the unencapsulatedness and the global inferences such a system can make, it has to deal with the same complexity issues like perspectives that are unencapsulated and specific do (top-right cell in Figure 2). But apart from these complexity issues, there may also be some conceptual questions.

I already mentioned the question, what is *social* about a social brain that responds to non-social inputs. From a functional perspective, one may argue that a general purpose social brain is conceptually unviable. The 'Functional-conception' of the social brain is a reference to all brain functions related to social cognition. However, if it seems that the cognitive architecture(s) of these functions are not specificly social, there does not seem to be anything *social* about this type of social brain, at least from a functional (or architectural) perspective. In this case social cognition seems to be the same as cognition in general. Assuming the social brain is general purpose seems thus inconsistent with the 'Functional-conception'. A general purpose social brain is, however, consistent with the 'Complexity-conception'. Recall that this notion is a reference to the evolutionary growth of the cortex in primates related to group size. Because this conception explains why humans have a large brain size, rather than explaining how its functionality is performed, it is still consistent with the idea that the cognitive complexity that humans have evolved under social pressures is a general purpose system. This divergence — i.e., the 'Complexity-conception' being consistent with a general purpose social brain, whereas the 'Functional-conception' being inconsistent— furthermore highlights the importance that neuroscientists are clear about what conception they are researching.

5 Discussion

The first research question of this thesis was about investigation in what sense the term ‘social brain’ is used in today’s neuroscientific literature. The thesis distinguished four notions of the ‘social brain’ and introduced a corresponding conception for each notion. The thesis showed that in the context of research on the ‘social brain’ it seems important that scientists are clear about which conception they are researching. Though the conceptions differ, they also seem to be related. Adopting a particular kind of cognitive architecture —this is the scope of the ‘Functional-conception’— can have consequences for the neural structure. This thesis did not focus on this question. Further research on the ‘Hardware-conception’, which is about neural structures, may be an interesting topic.

Secondly, this thesis investigated what cognitive architectures can support a functional ‘social brain’. Four types of perspectives were discussed (based on the modularity properties informational encapsulation and domain specificity). It turned out that each combination either seems to be conceptually infeasible or computationally intractable (see Figure 6). Although perspectives with conceptual problems seem fundamentally unviable, the perspectives with computational issues seem to pose considerable challenges that we may yet solve by scientific progress on methods from artificial intelligence (e.g. Kwisthout and van Rooij (2012); Blokpoel, Kwisthout, van der Weide, Wareham, and van Rooij (Submitted).

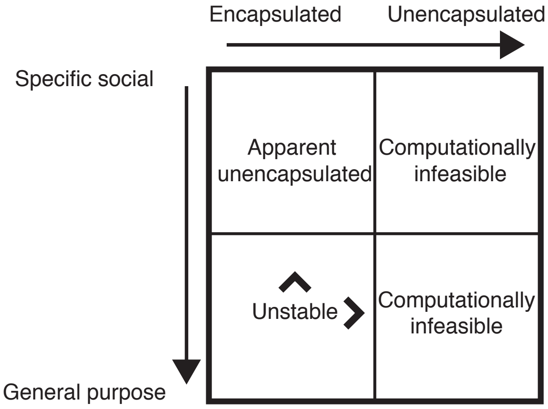

Figure 6: The conceptual analysis of this thesis showed that every combination of properties has its own conceptual or computational issues. Assuming the social brain is encapsulated seems conceptually infeasible, whereas conceptualizing the social brain as unencapsulated seems to result in computational issues.

Furthermore, this thesis proposed a cognitive architecture for every type of perspective. Given that each of the two described architectures —modular versus abductive— has its benefits and drawbacks, it seems important for cognitive neuroscientists to be explicit about which of the two architectures they are adopting, e.g., in the context of research on the ‘social brain’. After all, adopting one or the other architecture can have fargoing consequences for how one interprets existing findings. Also, adopting a particular architecture commits a cognitive scientist to addressing a different set of theoretical challenges that will need to be overcome for the account to be descriptively adequate and computationally feasible.

Apart from the properties used in the analysis of this thesis one can probably come up with other characteristics that may be interesting as well when investigating cognitive architectures that can support a social brain. Also, the definitions of the properties encapsulation and specificity are themselves a popular subject of discussion. In the literature there is no clear definition of these terms. Specificity can, for instance, be interpreted as something exclusive. A module can only be used for a specific cognitive task. Specificity has also a less strict sense, namely that a task specifically uses a particular module and that other tasks may also use that module as long as the inputs are from the same domain. This is the definition of specificity that was used in this thesis. In the same context Stone and Gerrans (2006) argue: "However, from the fact that a set of neurons is necessary for performance of a particular cognitive function it does not follow that the neural circuit is specific to that function." Furthermore, four distinctly different types of perspective were discussed ignoring the degrees to what extent perspectives are encapsulated and specific. By basing the analysis on other properties, by using other definitions of encapsulation and specificity, or by including the degrees the outcomes could differ from the ones presented in this thesis.

In this thesis only a modular and abductive architecture were discussed. Assuming a modular social brain seems to pose conceptual challenges. Remember that abduction was used as a solution to these challenges, but at the same time seems to increase the computational complexity to an unviable level. To my knowledge there is to date no non-modular architecture that is both conceptually and computationally feasible. If such a non-modular architecture exists and is applicable to the social brain, this could have a major impact on research on the social brain.

Acknowledgements

I would like to gratefully acknowledge my supervisors Mark Blokpoel and Iris van Rooij. Also, I would like to thank Arjen Verhulst and Arne Wijnia for reading and commentating a preliminary version of this thesis.

References

Adolphs, R. (2009). The Social Brain: Neural Basis of Social Knowledge. *Annual Review of Psychology*, *60*(November 2008), 693–716.

Allman, J. (1999). *Evolving brains*. Scientific American Library.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985, Oct). Does the autistic child have a ”theory of mind”? *Cognition*, *21*(1), 37–46.

Baron-Cohen, S., Ring, H. A., Wheelwright, S., Bullmore, E. T., Brammer, M. J., Simmons, A., et al. (1999, Jun). Social intelligence in the normal and autistic brain: an fMRI study. *Eur. J. Neurosci.*, *11*(6), 1891–1898.

Barrett, H. C., & Kurzban, R. (2006, Jul). Modularity in cognition: framing the debate. *Psychol Rev*, *113*(3), 628–647.

Barrett, L., & Henzi, P. (2005, Sep). The social nature of primate cognition. *Proc. Biol. Sci.*, *272*(1575), 1865–1875.

Block, N. (1995). The mind as the software of the brain. In D. Osherson, L. Gleitman, & S. Kosslyn (Eds.), *New york* (Vol. 3, pp. 377–425). MIT Press.

Blokpoel, M., Kwisthout, J., van der Weide, T. P., Wareham, T., & van Rooij, I. (Submitted). A computational-level explanation of the speed of goal inference. *Submitted to Journal of Mathematical Psychology*.

Brothers, L. (1990a). The social brain: A project for integrating primate behavior and neurophysiology in a new domain. *Concepts in Neuroscience*, *1*, 27–51.

Carruthers, P. (2003). On fodor’s problem. *Mind & Language*, *18*(5), 502–523.

Coltheart, M. (1999). Modularity and cognition. *Trends in Cognitive Sciences*, *3*(3), 115–120.

Cosmides, L., & Tooby, J. (1992, 1992). Cognitive adaptations for social exchange. In *The adapted mind: Evolutionary psychology and the generation of culture.* (p. 163 - 228). New York/Oxford: Oxford University Press.

Cosmides, L., & Tooby, J. (1994). Origins of domain specificity: The evolution of functional organization. In R. Cummins & D. D. Cummins (Eds.), *Minds, brains, and computers* (pp. 523–543). Blackwell Publishers.

Dunbar, R. I. M. (1998). The social brain hypothesis. *Evolutionary Anthropology*, *6*(5), 178–190.

Fodor, J. A. (1983). *The modularity of mind* (Vol. 341) (Nos. Book, Whole). MIT Press.

Fodor, J. A. (2000). *The Mind Doesn’t Work That Way* (Vol. 22) (No. 4). MIT Press.

Ford, K., & Pylyshyn, Z. (1996). *The robot’s dilemma revisited: The frame problem in artificial intelligence*. Ablex Pub.

Gallagher, H. L., & Frith, C. D. (2003, February). Functional imaging of ’theory of mind’. *Trends Cogn Sci*, *7*(2), 77–83.

Gerrans, P. (2002). The theory of mind module in evolutionary psychology. *Biology and Philosophy*(1994), 305–321.

Gigerenzer, G., Hoffrage, U., & Goldstein, D. G. (2008). Postscript: Fast and frugal heuristics. *Psychol Rev*, *115*(1), 238-9.

Gigerenzer, G., Todd, P., & ABC Research Group. (1999). *Simple heuristics that make us smart*. Oxford University Press.

Happe, F., Ehlers, S., Fletcher, P., Frith, U., Johansson, M., Gillberg, C., et al. (1996, Dec).

’Theory of mind’ in the brain. Evidence from a PET scan study of Asperger syndrome. *Neuroreport*, *8*(1), 197–201.

Haselager, W. (1997). *Cognitive science and folk psychology: The right frame of mind.* Sage Publications.

Haselager, W., Dijk, J. van, & van Rooij, I. (2008). Handbook of cognitive science. an embodied approach. In (chap. 14). Elsevier Science.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997, Jun). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.*, *17*(11), 4302–4311.

Kwisthout, J. (2011). Most probable explanations in bayesian networks: Complexity and tractability. *Int. J. Approx. Reasoning*, *52*(9), 1452–1469.

Kwisthout, J., & van Rooij, I. (2012). Bridging the gap between theory and practice of approximate bayesian inference. In *Proceedings of the 11th international conference on cognitive modeling)* (p. 199-204).

Leslie, A. M. (1991). The Theory of Mind Impairment in Autism: Evidence for a Modular Mechanism of Development. In A. Whiten (Ed.), *Natural theories of mind: Evolution, development and simulation of everyday mindreading.* Oxford: Basil Blackwell.

Mccarthy, J., & Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. In *Machine intelligence* (pp. 463–502). Edinburgh University Press.

Müller-Lyer, F. C. (1889). Optische Urteilstäuschungen. *Archiv für Physiologie*, 263–270.

Newell, B. R. (2005, Jan). Re-visions of rationality? *Trends Cogn. Sci. (Regul. Ed.)*, *9*(1), 11–15.

Peirce, C. S., Weiss, P., & Hartshorne, C. (1974). *Collected papers of charles sanders peirce / edited by charles hartshorne and paul weiss* [Book]. Harvard University Press, Cambridge, Mass. :.

Pinker, S. (1997). *How the mind works.* New York: W. W. Norton & Company.

Prinz, J. J. (1998). Is the mind really modular ? *Contemporary debates in cognitive science*(1983), 22–36.

Rice, C. (2011). Massive modularity, content integration, and language. *Philosophy of Science*, *78*(5), pp. 800-812.

Stone, V., & Gerrans, P. (2006). What’s domain-specific about theory of mind? *Social Neuroscience*, *1*(3-4), 309-19.

Thagard, P. (2000). *Coherence in Thought and Action Paul Thagard.* Cambridge, MA: MIT Press.

Tooby, J., & Cosmides, L. (1992, 1992). The psychological foundations of culture. In *The adapted mind: Evolutionary psychology and the generation of culture.* (p. 19 - 136). New York/Oxford: Oxford University Press.