Modeling Multimodal Communication as a Complex System

Kristinn R. Thórisson

Center for Analysis & Design of Intelligent Agents
and Department of Computer Science
Reykjavik University
Kringlunni 1, 103 Reykjavik, Iceland
thorisson@ru.is

Abstract. The overall behavior and nature of *complex natural systems* is in large part determined by the *number* and *variety* of the mechanisms involved – and the *complexity of their interactions*. Embodied natural communication belongs to this class of systems, encompassing many cognitive mechanisms that interact in highly complex ways, both within and between communicating individuals, constituting a *heterogeneous, large, densely-coupled* system (HeLD). HeLDs call for finer model granularity than other types of systems, lest we risk them to be not only *incomplete* but likely *incorrect*. Consequently, models of communication must encompass a large subset of the functions and couplings that make up the real system, calling for a powerful methodology for integrating information from multiple fields and for producing runnable models. In this paper I propose such an approach, *abstract module hierarchies,* that leverages the benefits of modular construction without forcing modularity on the phenomena being modeled.

Keywords: Multimodal realtime communication, computational model, complex natural system, heterogeneous large system, abstract module, theory of dialogue.

1 Introduction

A large number of mental mechanisms play a role in embodied dialogue, from task planning to sentence composition to control of eye gaze from moment to moment. How these are coordinated has to be explained if we want to claim that we understand communication. Examples of what a *complete* theory of dialogue should be able to explain is whether/how some eyeblinks seem to be related to the production of speech content, how it is that sometimes there are less than 50 msec gaps between speaking turns, how facial expressions or intonation can modify the meaning of utterances, and why people look at their hands when doing some types of gestures and not when doing others. Ideally it should also explain how and why communicative and social behaviors are affected in certain ways and not others by certain drugs such as alcohol.

A basic theory of human dialogue presupposes that humans communicate – it is not its role to explain why they communicate in the first place or how such communication systems came about; that is the realm of sociology, biology and evolution science. Suffice it to say that without communication a species has less

survival potential, and that the mechanism of reciprocity between individuals of a species must be present for such systems of communication to emerge. Assuming that an individual of a communicating species must have certain basic abilities to perceive and act in ways to communicate in a certain way, we must also postulate certain perceptual and representational capabilities related to communicative interaction.

Among the properties of minds that most greatly seem to influence the commucative apparatus and its operation in realtime dialogue is cognitive capacity. Turntaking is a necessary mechanism to manage the limited processing capacity of a normal human mind in normal operation when processing information-rich content: The reason that we take turns when communicating is that the mind has a limited capacity to generate speech while concurrently hearing and understanding the other speech – attempting to do so for longer than a few seconds is sure to significantly reduce a person's retention and understanding of what was said around her (cf. [1, 2]). Any model of realtime embodied dialogue must therefore take into consideration – in addition to the ability to generate speech, express concepts, gesture and take turns – attentional mechanisms and the inherent capacity of the various cognitive faculties. A necessary and sufficient model of dialogue should in fact be detailed enough that we could build an artificial system that can participate fully in human dialogue. Of course we are not going to build such a system on first attempt, or even second; as we must bring to bear on the task methodologies from several disciplines, this will take significant effort and time.

This paper has two main parts. In the first part I argue that multimodal realtime dialogue shares many characteristics with other complex sytems, such as economies and ecosystems – what we refer to as heterogeneous, large, densely-coupled systems (HeLDs) – and must therefore be studied using some of the same methods as are being developed in these fields. Recognizing this fact may have important implications, in particular, that we need to supplement our efforts with methods from simulation and computational modeling theory. We will start by discussing the claim that embodied multimodal dialogue is a *complex system*, in the sense used by e.g. Simon [3]. Such systems embody/express emergent properties that have been difficult to understand without resorting to large, detailed computational models.

In the second part of the paper I present ideas on the kinds of architectures that might implement such a complex system. In particular, we will discuss how a modeling approach called *abstract module hierarchies* that can overcome many of the difficulties associated with studying complex systems, and how modularity in *implementation* does not have to presuppose modularity at the *cognitive* or *brain* levels. We briefly present two systems exemplifying the use of abstract modules for modeling cognitive and neurocognitive mechanisms. Lastly we will look at arguments for why there might be reason to think that the brain is a modular computing substrate, and thus reason to expect isomorphism between cognitive and brain structures.

2 Complexity of Cognitive Mechanisms

Even the most casual analysis of human realtime communication reveals an intricate complexity of knowledge and behaviors that together define it. First I will briefly

review three important sources of complexity, namely interaction at multiple timescales, perception-action loop and multiple information types.

2.1 Multiple Levels of Detail

The full range of dialogue behaviors can be affected by events on many space and timescales, from the emotional impact that a long-winded insult can have on the choice of one's words to the implied surprise of 30 msec noise bursts from clicks of the tongue; from the threat "emanating from" a large fist shaken in one's direction to the effects of Ethanol in alcoholic drinks on outbursts of "honesty".

The shortest and the longest behavioral event in a meaningful dialogue range from 10-40 milliseconds (e.g. tongue click or a quick glance) upwards of several hours, possibly days.[1] Meaningful behaviors thus span at least 4 orders of magnitude of time, even in the simplest cases of multimodal dialogue of meeting and saying "hi" in the hallway. Dialogue participants must keep track of events at the full range of timescales; various kinds of behaviors form clusters at certain timescales, or "bands": eye movements and gaze at the low end (from 40 msec upwards of 1 second for lingering gaze), head, hand and arm movements between 1-2 per second, body movements a bit less frequent, and so on. The perception of each participant of others' behaviors, and their alignment and reciprocation of such behaviors by others, is highly task-dependent, yet bounded by the natural limitations inherent in the body and cognitive capabilities.

Fine-grained analysis of the temporal nature of multimodal action during dialogue, from gaze and upwards, reveals significant repeated and mirrored patterns between dialogue partners [4], a clear sign that the perception and action of each participant is being coordinated to a very fine degree and that interaction behaviors happens at many timescales – gaze is met with gaze, verbal utterance with verbal utterance, gesture with gesture, topics are negotiated, abandoned and revisited. Clearly, tight coordination of such multi-dimensional events requires an intricate underlying architecture, where short- and long-term planning, powerful perception, reactive decisions and fine motor control all come together in a coherent way.

2.2 Perception-Action Loop

Dialogue, and especially multimodal dialogue, is inherently realtime. By *perception-action* loop we mean the continuous, "on-line" ability of living beings to react to something that comes in through their sensors and monitor their own behavior and the

[1] As the scale moves towards days, months and years, the category "meaningful dialogue" gets increasingly vague. We could use matching goals to classify a set of communicative events, such as utterances, gaze, facial expression, etc., into a larger communicative event such as interview, collaboration, etc. This method, however, has its limits, for as the size of the overarching goals increases the grouping becomes less obvious. To take an example, just because two people work at the same company, and thus share a positive attitude towards each other, we would hardly classify their greetings every day, over a period of e.g. 20 years, as "communicative event" lasting 20 years, even though their employer provides them with a shared goal and a context for greeting. In spite of its limitations, the method of using goals in the classification is useful in cases where the goals are fairly obvious.

environment in a continuous fashion. At the low end the shortest possible voluntary path through this loop in humans, from sensation to action, is bounded by the choice reaction time, 90 ms [5], and at the high-end by the patience of participating in a back-and-forth with someone about a topic or class of topics, as well as the speed at which new thoughts and associations can be generated in relevance to that topic.

An interval of 60 msec, from the time one senses something until they need to give a reply, or take turn, is not sufficient time to contemplate much at all, e.g. infer the major implications of a sentence such as "The Chinese have a vested interest in keeping business open to the West", yet people do show visible and meaningful responses even as the sentence is being spoken. Quick, semi-automatic and fully automatic behaviors, such as fixations and saccades, have been classified as "reactive"; behavior based on deliberate and consciously reportable effort have been termed "deliberative" (cf. [6]). While the distinction seems crude, it has some basis in brain structure [7]. Where reactiveness cannot provide sufficient responses, prediction kicks in: Using various features of speech, gesture, intonation, gaze and so on, people will anticipate what a person is going to do and will act accordingly, to respond, comply, give turn or hand over a tool that was requested. Even the simplest target tracking tasks seem to involve realtime (close-horizon) prediction [8] as does listening to speech [9]. It seems likely that our cognitive apparatus employs several interwoven mechanisms for producing such intricately timed behaviors.

2.3 Multiple Information Types – Numerous Forms of Constraints

Computationally speaking, the data that enter into dialogue are of many types [10, 11], spanning the full spectrum from deterministic to stochastic, continuous to discrete. In fact, the biggest criticism of turntaking research in the last 2-3 decades can be said to be a level of simplification that has had a chilling effect on progress and bound researchers in the shackles of side-effects, arguing over details such as whether there is such a thing as a "turn-constructional unit" and if so, it being sentential [12], syllable-based [13], multimodal (cf. [11, 14, 15]) or content-driven [11]. The situation painfully reminds one of the story about blind men arguing about an elephant[2] – one touches the trunk and concludes that elephants are like snakes, another touches a leg and concludes that elephants are tree-like. The analogy has seldom been so appropriate, as a quick look at a few examples of successful turntaking shows: Efficient task-oriented communication on a noisy factory floor – no possibility to synchronize on syllables or full sentences; successful communication between the deaf (no sound at all); successful communication on the telephone – no multimodal information (although plenty of verbal paraverbal information) – the examples clearly argue against simplistic explanations of turntaking, as for example proposing that phoneme timing is its main perceptual driver [13]. As O'Connell et al. [11] point out, a proper turntaking theory should cover varied situations ranging from debates, to lectures, negotiations, task-oriented interactions, media interviews, dramatic performances, casual chats, formal meetings, etc. If such a theory is to provide an understanding of what drives turntaking, then it must provide a way to account for the (several) goals that will be in operation in any conversation – goals pertaining to the

[2] *Six Blind Men and the Elephant* – poem by John Godfrey Saxe (1816-1887).

individuals' disposition (e.g. a seller who wants to maximize profit), social norms (e.g. no introductions at a store counter), relationships between the participants (e.g. friends who want to stay friendly), purpose of the interaction – the aligned goals between the individuals, i.e. the purpose of the interaction, etc.), cognitive limitations, characteristics and "parameter settings" (e.g. average and maximum speed of understanding what is being said, speed of planning, motivation levels, vested interest), as well as a host of issues related to semantics – from confusing sentences to Freudian slips – which can affect emotions, attitude and other things that influence timing and events in turntaking (see Figure 3). To explain these, however, it does not

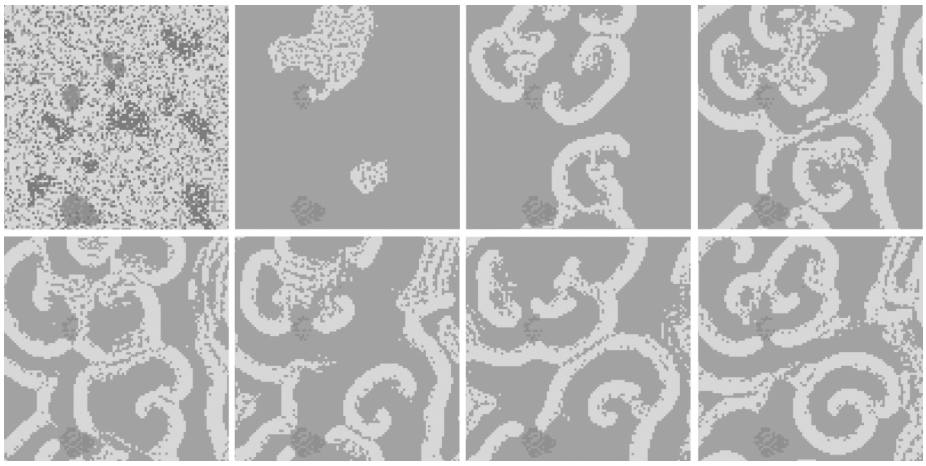


Fig. 1. A grid of 100x100 cells in a cellular automaton, each determining its own state as represented by a color (shown here as one of 4 shades of gray), according to the rules in Figure 2. Spiral patterns will emerge and persist over minutes and even hours of running. The spirals are emergent from the rules: It is difficult, if not impossible, to predict the emergence of such spirals looking only at the rules, partly because they only appear under certain initial conditions, but mainly because it is the interaction between the rules that produces them. Each snaphsot taken at the initial state (upper left), and then at 15 second intervals, ordered from left to right, top to bottom.

suffice to propose some simple mechanisms that can generate (some limited amount of) the surface phenomena observed in human dialogue – the mechanisms proposed must explain those and additionally how they interact with the various complexities of perception and direction of attention, the human ability to formulate coherent sentences, to understand them when spoken by others, and their ability to follow social etiquette.

2.4 Emergence

If scientists of an alien race were to land on Earth to analyze how automobiles operate, they would see numerous different behaviors of these entities: long rows moving slowly in packs, single cars zipping along dirt roads, cars stopped at a red light, single cars parked by houses and rows of cars parked by shopping centers, cars

slowing down when approaching intersections, and a few incidences of cars jammed into each other. But as most Earthlings know, such high-level behaviors (the behavior of the whole traffic system within its environment, cities) are all emergent from interactions between several components such as the car's owners, the human laws pertaining to what may and may not be done while operating a motor vehicle – even the long-term goals of the car's owners, e.g. not wanting to die. The observed behaviors of cars are an *emergent* property stemming from the interactions among these complex components. To the aliens' delight, the fact that these components and their interactions are lawful means that observable effects can be classified, labeled and reproduced through manipulations, and modeled at the observable features level – the system's gross-anatomy. To achieve accuracy and breadth in applicability of such

When Light Gray (LG)

1. Turn MG: If there are more than 20 LG cells around and lifetime exceeds 30
2. Turn MG: If there are less than 12 LG cells around and lifetime exceeds 20
3. Turn MG: If number of LG cells around equals 25
4. Turn MG: If lifetime under any circumstance exceeds 60

When Medium Gray (MG)

5. Turn LG: If there are more than 8 LG cells around and their lifetime combined exceeds 80 and there are more than 10 MG cells around

When Dark Gray (DG; only visible in the initial state)

6. Turn MG: If there are more than 3 DG cells around and the sum of their lifetime exceeds 2 and lifetime is greater than 8

Fig. 2. Rule set used for the states (grayscale) in the cellular automaton simulation in Fig. 1. Notice that LG and MG are responsible for the spiral patterns. (*Lifetime* means lifetime of the current cell, measured in simulation steps; *around* means the closest cells, in a 5x5 grid, surrounding the current cell.)

a model one must, however, go beyond the surface phenomena and uncover hidden factors. Unless we lived in a perfectly bijective reality, we have to infer the underlying causality and for that we must build on what lies at the next organizational sub-level.

Emergent phenomena have been extensively studied using cellular automata, where simple one, two or three-dimensional grids of cells, decides its own state based on the behaviors of its neighboring cells, according to a set of local rules (cf. [16]). Figure 1 shows an example of such a system: The spirals are an emergent pattern, stemming from a small set of identical rules (Figure 2) local to each square (there are 100x100 cells in this example). The spirals are highly persistent in light of significant disturbances, but are quite sensitive to initial conditions and will not appear in about 10% of cases with randomized initial conditions. These particular rules, however, need to *all* be present for the spirals to form: Taking out any single one will remove the appearance of spirals altogether.[3]

[3] An interesting exception appears when removing Rule 3 – it makes the spirals somewhat less spirally but does not remove them completely.

If these alien scientists happened to be extreme optimists they might spend a lifetime studying only the observable features of automobiles, hoping to unravel the whole story that way. In the process they would know a lot about the high-level behaviors of cars but very little else. They would be able to describe and classify all of the cars' behaviors in fine detail, but they would never be able to explain all of them, because that is simply impossible without an underlying model of how cars are operated, have owners, are made of metal, must stay on roads, collide as a consequence of operator and mechanical problems, etc. For example, predicting that more collisions happen when the sun sits low on the horizon and hits the cars from the front is quite easy, based on prior observations, but explaining *why* it happens requires nothing short of an understanding of drivers and their perceptual apparatuses. Or consider explaining why the cars move without the concept of an engine; or of how an engine operates without some idea of fossil fuels, electricity and flammability, density and strength of aluminum and iron, etc. In other words, a model of their components and all elements related to the automobiles, constitutes a complete understanding of automobiles; social convention and cognitive limitations are needed for understanding traffic – the behavior of groups of cars in cities. The high-level behaviors of automobiles *are an emergent property of the interactions between all the elements that matter to their behavior*, nothing more and nothing less.

2.6 Intermediate Conclusion

As we have argued above, multimodal natural realtime dialogue is likely to be most adequately described as a complex system. Figure 3 gives a list of many factors and phenomena that can affect the way observable features in a dialogue turn out. The enormous complexity involved in even a single item on this list makes it ever clearer that embodied, realtime face-to-face communication is a complex system involving a large number of functions,[4] the result of many interacting subsystems, none of which has clear domination over the system's characteristics – each element contributes to some part of the system's operation through its local operation and interactions with other elements. Evidence from evolution points in this direction too; in particular, evolution is likely to have come up with a tangled web of mental mechanisms that serve many purposes in many ways, because once one mechanism is in place it is more likely to be modified in subtle ways and reused than for another mechanism to evolve from scratch – a phenomenon called *exaption*. The result of such processes is systems with large amounts of structural dependencies – mixed heterogeneous systems.

[4] The term *function* is used here as in psychology, anatomy and biology, as *functioning*, *ability*, *role*, etc., akin to the concept of *structure* in anthropology (cf. [17]): The “family” is a structure encompassing more than its parts, but yet can only be pointed at by naming its constituents or from the exterior, by naming its connections to the tribe, as it exists at a different level of organization. This kind of metonymy shows in fact that it’s beneficial to clearly separate the substrate and the emergent functions (i.e. to abstract based on *organizational levels*) when wanting to identify structuring feedback loops, to run them and validate (or dismiss) abstraction hypothesis.

1. Timespace
  1.1 Physical constraints
    1.1.1 Body can only be in one place at a particular time
    1.1.2 Sensory organs limited area coverage
    1.1.3 Manipulators of limited number (arms-hands 2, fingers 2x5, typically)
  1.2 Temporal constraints
    1.2.1 Body takes time to move (especially important for sensory apparatus)
    1.2.2 Body only exists a particular period in time (hence the need to communicate across time)
    1.2.3 Sensory uptake takes time
  1.3 Cognitive apparatus
    1.3.1 Variable time for processing different types of information from senses

2. Information-carrying capacity of our communicative apparatus (body)
  2.1 Arms and hands
    - Placement, speed, shape, manner of movement may all matter
  2.2 Face
    2.2.1 Gaze direction, fixations
    2.2.2 Head direction, movement
  2.3 Mouth
    - Speech, non-speech sounds/paraverbals
  2.4 Body
    - Stance, direction, shape

3. Cognitive capacity
  3.1 Perceptual integration: Hearing and vision are different types of data
  3.2 Attentional control
    - To understand well we have to focus our attention on a single individual's communicative acts; this is perhaps the single biggest reason for the existence of turntaking
    3.2.1 Visual attention
    3.2.2 Auditory attention
  3.3 Knowledge (this is big)
    3.3.1 Individual differences
    - Individuals have different amounts and types of knowledge, hence a need for grounding
    3.3.2 Knowledge of social convention
    - Various types of behavior may be inhibited or expected by social rules of conduct
    3.3.3 Situation recognition
    - A situation needs to be classified correctly in order to be acted upon with the intended effect
  3.4 Memory
    3.4.1 Memory types
    - We have different memory systems for events, words, concepts; these have various limitations
  3.5 Goals & Intentions
    - Various goals may come into play; this is a list in and of itself. These factors are closely related to and interact strongly with knowledge.
  3.6 Planning
    3.6.1 Planning of body movement
    3.6.2 Planning of words
    3.6.3 Synchronization of various bodyparts
    - For sensing (e.g. fixate on the right place) and for information production
  3.6 Learning

Fig. 3. These are only some of the constraints that a communicating system must take into account; most of them may influence, in one way or other, the way participants in dialogue behave.

So how complex is natural multimodal communication? Is the complexity greater than that of an automobile (minus its human operator)? Surely. Is it more complex than the example cellular automata world depicted in Figure 1? Most certainly, as the preceding sections clearly hint at. How about an ecosystem? Probably not; besides being dependent on very complex energy transfers, many of the functional elements in a (human-less) ecosystem contain cognitive perceptuo-motor systems that rival human ones. Therefore we can assume that the complexity of multimodal communication, as a system of systems, lies somewhere between an automobile engine and an ecosystem. When trying to formalize systems with a large number of functions and inter-structural dependencies, the requirement for a high level of model detail is thus likely to be very strong, as no single factor explains a significant part of the whole system's operation, just as the rules in the spiral world example above.

We can now make the following summary about multimodal realtime dialogue:

(1) Observable behaviors of dialogue participants – glances, manual gestures, choice of words, intonation and prosody, etc. – are not any more sufficient for explaining the phenomena of communication than the movement of automobiles is sufficient to explain their operation.
(2) In order to be adequate, our human communication models may very likely have to encompass most (if not all) the components and couplings that make up the system; anything less is likely to be both incomplete and incorrect. Leaving out a large set of phenomena tightly integrated with, and observed to affect, dialogue behaviors, such as e.g. gesture, prosody and intonation – even breathing – is very likely to leave us with an incomplete model of dialogue, quite possibly a model that is also incorrect.

In face of this conclusion we need to answer several questions. The main one, the one we will address in the next section, is *What methods can be employ to build a model that can take the part of a human dialogue participant and thus explain sufficiently how embodied multimodal dialogue works*?

3 Models & Methods

If there is one thing clearer now than it was 50 years ago regarding natural language and dialogue, it would be that cognition related dialogue is more complex than had we dared to imagine. It has been said that biological research is difficult because in living systems everything is causally connected to everything else. Luckily this is unlikely to be true of cognitive mechanisms (and probably also biology), but we can be sure that any subsystem we may identify in multimodal dialogue is bound to have multiple connections to other subsystems in the human cognitive system.

Historically, an important tool for studying human behavior in psychology has been hypothesis refutation. Based on Popper's (in)famous argument that hypotheses can never be proven, only refuted [18], much psychological research today addressing cognitive architecture proceeds by experimentation based on fairly broad-stroke generalizations about its information structures. However, as eloquently argued by Newell [19], “you can't play 20 questions with nature and hope to win”, meaning that a coarse-grain approach through hypothesis testing through human subject

experimentation must be supported by other research methods.[5] The general idea behind the information processing view of intelligence, as introduced by Turing [21, 22] and others, has taken hold in many parts of psychological research. While strong versions of the thinking-as-computation stance have led to in-fights among researchers, modeling with structures does not imply isomorphism – that the modeled reality is modular – nor does it imply that the modeled object has to be computationally reducible.[6] We will come back to these issues shortly.

Simulation models vary widely depending on the phenomena under study. For example, the behavior of a homogenous system, e.g. a liquid consisting of one type of molecule in large numbers, behaving according to the laws of physics, can be described by relatively simple equations. Equations that take into consideration large-scale indicators of monetary inflation can be used to model large-scale movements of a market. But in neither case can these equations be used to describe individual molecules or currency transactions, respectively. Not so for many other systems. Consider the example of a car engine: trying to understand how it works by only looking at the carburetor and the battery is not likely to get us very far. The automotive engine is composed of a large number of heterogeneous components, each responsible for only a small part of its total operation, yet ignoring any one of them will likely leave us with an incorrect model. To take a hypothetical example from the brain, we might be able to model spatial hearing sufficiently abstractly for certain tasks that the human auditory system needs to perform, but typically that (limited) model will break down in many other contexts and for many other tasks. If we want to have a finer granularity of the spatial hearing faculty, the only solution would be to model it in more detail, because what defines it at those other tasks may very likely be its composition at lower levels of detail, which interact in complex ways with *other systems* needed for *other tasks*. As the list in Figure 3 shows, a model that can take the vast amount of relevant systems into account, and produce the kinds of patterns observed in multimodal dialogue, is not going to be simple. The architecture of such a system will have much more in common with the global telephone network and Internet than with the mathematical models of physics, that is, it will most likely be composed of heterogeneous interacting systems that are "nearly decomposable" but not quite, and it will be highly detailed. Furthermore, these models will be highly dynamic. The only (presently known) way to make such models is to implement them as information structures, in the form of programs, and run them on computers, monitoring their performance and comparing it to the natural systems they are supposed to represent. This has been the conclusion in many other fields studying

---

[5] Kosslyn [20] has taken this argument further and argued that binary decision making in researching complex systems can be done provided that the hypotheses are (a) anchored in detailed processing models and that (b) they are formulated from the viewpoint of multiple levels of analysis within a processing system. This is in accordance with the view argued here (see below).

[6] The computational stance is nevertheless an efficient framework for the construction of experimental (mathematical) models of the mind (cf. [23, 24, 25, 26]); it has advanced our understanding of the mind in several aspects, in many cases with superior results over other approaches, a good example being how neural impulses collect from the ears in the form of information that encodes position and orientation of sounds, directly in support of the survival of a species.

complex systems and is recognized as a powerful methodology for studying weather systems, evolution of galaxies, physical processes and more (cf. [27, 28, 29, 30, 31, 32]).

3.1 Large Heterogeneous Systems & Model Validity

Complex models with heterogeneous components call for a heightened need of thorough verification. One difficulty is that in such systems any subset of the observed behaviors can be mapped onto an infinitely large set of underlying hypothesized mechanisms, which are a challenge to verify. To take an analogy, uncovering the 8 rules of the spiral world (Figure 2) would be quite complicated simply by studying the emergent surface forms of the spirals. Numerous rule sets could undoubtedly be concocted that would generate similar, perhaps even identical spirals. But uncovering the *actual* rules would necessitate digging deeper, probably building a simulation of the world where one could try out different rule combinations running inside the logic believed to be responsible for their execution. Our human communication models might contain a high level of detail, but if it only addresses a limited level of detail it might be correct or it might be incorrect – in fact, there would be no way to tell.

Part of the problem thus lies in the fact that most current models, produced by the standard divide-and-conquer approach, only address a subset of a system's behaviors; yet for most complex systems, if we were to attempt to create a model that addresses *all* of the system's behaviors, the set of possible underlying mechanisms would be greatly reduced [33] – quite possibly reducing the probable mechanisms behind it to a small finite set. A way to address this problem is thus to take an interdisciplinary approach, employing results from various levels of abstraction to bear on the modeling efforts. Use of such hierarchical approaches is common in e.g. physics, as all physicists know, for example, that behind the science of optics lie the more detailed models of electromagentic waves [34]. Thus, when dealing with heterogeneous, large, densely-coupled systems (HeLDs) it is important that we try to constrain the search space for possible designs, and one powerful way to do this is to build multilevel representations (cf. [33, 35, 36, 37]); indeed, in understanding natural HeLDs this may be the only way to get our models right. Notice that the thrust of the argument is not that multiple levels are "valid" or even "important", as that is a commonly accepted view in science and philosophy, but rather, that to map correctly to the many ways subsystems interact in HeLDs they are a *critical necessity*, lest we chase variations on our altogether incorrect models ad infinitum.[7] Unless simulations are built at fairly high levels of fidelity it is not possible to experiment with changes and modifications to the architecture at various levels of detail. Without this ability we cannot differentiate between a large set of models that, on paper, look like they might all work. To quote Simon [32] on this subject, for much simpler phenomena: " Even a few particles, three or more, reacting in classical Newtonian fashion, create the notorious three-body problem, which is usually not solvable in closed form, and which, under many circumstances, leads to chaotic system behavior."

[7] A short overview of the importance, as well as pitfalls, of multiple levels of description in science is given by Bakker & Dulk [38].

So the solution to the problem of model validity, as well as the solution to increased model detail, is to attempt to anchor current models in a theory about phenomena at a higher or lower level of detail, assuming those theories have been experimentally grounded.

Another useful weapon in the fight for complete and accurate models is modular construction. Modular approaches, in contrast to monolithic designs, have been shown to speed up the development of large, complex robotic and simulation systems, and to facilitate the collaborations of large teams of researchers [39, 40, 41, 42). To take some examples, Martinho et al. [43] created an architecture designed to facilitate modular, rapid development of theatrical agents for virtual worlds and modularity played a large role in the construction of Bischoff et al.'s HERMES robot [44]. Simmons et al's robot Grace [45], with over 20 collaborators from 5 institutions, is another great example of a project that has benefited from a modular approach. Of course, whether the mind/brain can be modeled in a modular fashion is still debated in the research community and not all are convinced of its merits. However, in the software engineering sense, this claim in its essence simply represents a practical solution to a highly challenging problem: it does not force us into – or even in the slightest sense imply – the view that the brain is literally a set of components. Rather, the claim of modular construction is that our understanding of the brain/mind can be fruitfully formalized that way when implemented as computer models. Of course, we'd like to incorporate as many faculties of the mind as possible when modeling cognition, but this is impossible to do all at once; unlike monolithic approaches, a modular approach enables us to do this incrementally and to capture many of its aspects in many ways, thus preserving their richness under various perspectives. The trick is to realize that a modular construction does not have to imply a *theory of modularity*. To see how this could be so, we need to look at some theoretical building blocks that can lead the way. And so the next question for our modeling efforts arises, *What kinds of modularity*?

3.2 Abstract Modules & Near-Decomposability

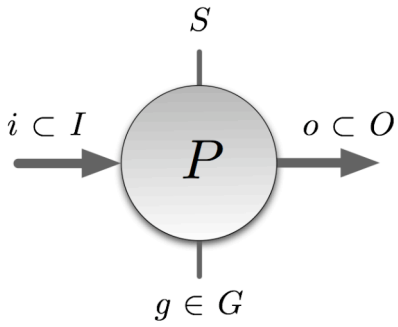
The concept of an *abstract module* builds on Simon's [3, 46, 47] concept of "near-decomposability" (ND): Systems that are divided into subsystems of interacting elements at multiple levels, where interactions between elements within a subsystem are an order of magnitude or two higher than interactions between subsystems. It can be found everywhere in nature, from the universe as a whole to biological to sub-atomic systems. A module in this sense is a theoretically motivated or practically motivated subcomponent or building block of a larger system, with causal relationships (couplings) to other such subcomponents (Figure 4). Together the subcomponents and their couplings define the system in question. We will look at examples of this in section 3.3 below, but first we will provide a general account of the idea and its benefits.

An abstract module represents *abstracted system functionality*. It has a goal or purpose g, an input i, an internal state S, a transformation process P, and an output o. In the tradition of many multi-agent systems, the goal can be a human-imputed justification for the module's existence – in other words the module's role in the architecture – and need not reflect an underlying theory (just like the existence of the

module itself). The transformation process transforms the input to an output according to some rules; if *i* is a continuous physical force and *P* is a damping mechanism, *o* will be some derivative of *i* according to mechanical laws; if *i* is some discrete information packet and *P* a routing mechanism *o* may be *i* in unmodified form but with a new destination.

Figure 5-I shows causal relations between six physical (or hypothetical) entities with particular causal relationships. To model these using abstract modules, several approaches can be taken. In II, the relationships have been implemented as three simulation models, with messages taking the place of hypothesized (or real) causation; left-hand side in II represents transmitting modules and right-hand side receiving modules. In III, the structure in I has been implemented as two alternative modular models, X and Y. In X, two modules are used to represent all causal relationships of I. In Y some modules from II have been merged. Notice that even in this implementation, where e.g. each module is running on its own computer or as its own thread, the original physical/theoretical relationship between the causes and effects has not changed (except insofar as in this example their effects on each other may not have the same resolution as reality would have it). In both II and III the modules' internal state (see Figure 4) represents the state of the causes and effects in I. Looking at Figure 5-I as physical reality, or a theoretical model of physical reality, nothing in II or III has changed in our modularization of the physical or hypothetical systems in I.

Fig. 4. An abstract module is composed of a process *P*, input *i*, output *o*, a state *S* and a goal *g*.



Because the approach can be employed purely for the practical purposes of getting a handle on excess complexity, it follows that the cognitive modules proposed by Fodor [48], for example, can be modeled as a single abstract module in which the module's state and goal is not shared with other modules, only its input and output. But if the modular model is in this way completely independent of the theory, what then is the benefit of the modularization? Doesn't it get in the way? The short answer is *no* – even if the modules in our implementation are completely orthogonal to the actual theory the system implements they will allow for the construction of larger, more detailed models, and help relate the work to related fields. Additionally they will help anchor a given level of organization in tangible, physical structures wherever this is appropriate. There are significant benefits to modularization:

- *Modular systems are easier to expand than monolithic ones*. This is a well-known fact in software engineering and computer science.
- *A modular model of a complex system is easier to simulate, as modules can be moved between processors*. The primary reason why this is important in cognitive research is that so much of cognition has no serial dependencies and can (and should) therefore be executed in parallel. In addition, computing power is becoming increasingly available prices continue to drop and advances keep being

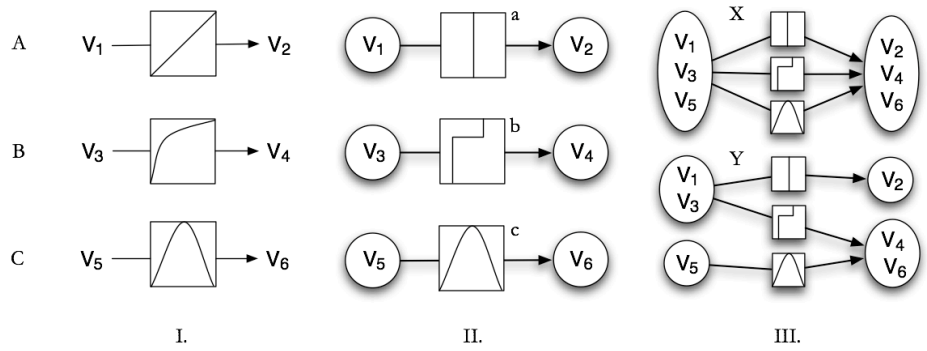


Fig. 5. Causal relations between variables [V_1, V_6]. I, II and III, left-hand side: Causes; right-hand side: Effects. Part I depicts physical causal relationships between variables (A – linear relation; B – logrithmic relation; C – hyperbolic relation). Alternatively, part I may represent theoretical models of physical or hypothetical constructs. In II, these relationships have been implemented as three modular simulation models, one module per causal factor and one per measured effect, with messages taking the place of physical relationships. The functions *a* and *b* connecting the modules have also been quantized from what they were in I. The left-hand side represents transmitting modules and the right-hand side receiving modules. In III, two modules are used to represent all causal relationships of I. In both II and III the modules' internal state (see Figure 4) represents the state of the causes and effects in I, respectively, and the modularization is thus independent of the theoretical model.

made. This point is less important for systems that are small enough to be run on single processors than more computationally intensive ones, but as the power of processors increases the benefits even there are becoming increasingly obvious.

- *Modular approaches facilitate collaborations between scientists, labs and universities*. This is extremely important, as HeLDs are difficult to build; sharing models and runnable code between scientists and even institutions may speed up the progress of cognitive research by orders of magnitude [49].

Even though modularization with abstract modules can be kept independent of the underlying model, this does not mean that modularization in a runnable model should never mirror modularization (whether hypothetical or real) in the system being modeled – quite the contrary: It may in fact sometimes be beneficial to make modules in a model directly mirror modules in the modeled system. To take the example of Fodor again, we could build a model where our abstract modules directly implement the way in which he intended his mind modularity to work.[8] It is important to keep in mind, however, that when such assumptions about actual modularity are made they must be made explicitly and clearly and not implicitly, as is often the case. This allows the validity of such a local hypothetical modularization to be further investigated – and eventually decided – in the course of the model verification procedure.

[8] This would presumably require a significant amount of detail to be added "between the lines" in his theory, as it is a relatively high-level and coarse-grained.

The idea of abstract modules as presented here continues along the line of Marr's [50] three levels of analysis, theory, representation and implementation[9] – but abstract modules go further, dealing with complex system architectures, relationships between semi-independent entities, and abstractions at multiple levels of detail. Although to some extent compatible with Minsky's Society of Mind [51] (multiple interacting subcomponents), the idea of abstract modules differs significantly from it in its emphasis (a) the importance of gross architecture in complex systems, (b) hierarchical models, and (b) the practical benefits of modularity for building runnable models. As Simon [3, 46, 47] points out about ND systems, they can be described as a hierarchy at multiple levels of abstraction (detail) where mechanisms at each level interact more between each other than any other part of the system as a whole. This concept is illustrated in Figure 6. The decomposition into levels, and subsystems at each level, can be structural and/or functional. A functionally decomposable system will have functions that can be isolated and implemented computationally as abstract modules – independently of how or whether its functional decomposition mirrors its actual physical/structural instantiation.[10] As one descends down this scale, detail, i.e. physical and temporal granularity, increases – the model involves smaller objects operating at higher frequencies.

Figure 7 shows three canonical abstracted examples of systems resulting from applying this methodology. In Example I a target system is decomposed at two main levels, the highest and the lowest. An example is e.g. a goal-stack for topics to be discussed (the high level), and neural mechanisms coding for the speaker's representation of spatial relations so that he is able to look at the listener (low level). Alternatively, the lowest level could be a neural model of goal representation, with spatial relations simply modeled at the top level as Cartesian points in space. In Example II the system is decomposed into three levels; take our first model and add a middle level describing how neurally-encoded spatial information (lowest level) informs the control of neck and eye-muscle tension (mid-level) to bring head and eyes to the desired positions, relative to the speaker's and listener's bodies. Example III is an example of "the modeler's nightmare": Here a system has multiple valid decompositions at any level (there are no discernable levels), potentially all equally good (or bad).

Furthermore, some abstract modules at each level have causal connections to abstract modules at different levels of description. Encountering this situation may in fact point to a possibility that (a) the phenomena one is trying to model are in fact not causally connected or that (b) they are in essence atomic. Note though that this does not mean one cannot model the system in a modular fashion, only that the modules and their connections will have a high level of arbitrariness.

---

[9] It is important to note that Marr's usage of "implementation" referred to the substrate – the hardware – that a model runs on, that is, how a system can be realized physically e.g. the brain or a CPU, while my use of the term "implementation" is used throughout in the sense of "software implementation", i.e. how a system is implemented as a software program. A key point here is that software implementations can approximate the hardware implementation to various degrees, along almost a continuous scale of fidelity.

[10] Of course any functional feature of an abstract runnable system must be implemented as causal chains at some physical level (not necessarily in a one-to-one relationship), lest we assume some sort of metaphysical causation – see Scheutz [52].

A key feature of ND heterogeneous systems is that the causal chains between their elements are a tangled web of different types of interactions, or couplings (Figure 8). We can classify these along at least two main dimensions, *density* and *tightness*. A dense coupling between two components makes them highly dependent on each other on many variables; a sparse coupling means only one or a few variables on either side affect the other. A variable in component *A* is tightly coupled to a variable in component *B* if changes in it affect changes in the other in a (close to) 1:1 relationship. A loose coupling implies a statistical relationship or that e.g. only part of the range of one variable affects the other. Of course, any given HeLD may be composed of a combination of components that vary along both dimensions; what makes a system a HeLD is its large number of components and the existence of a significant number of dense couplings. But what does it mean to implement a tightly coupled causal relationship (theoretical model) as loosely coupled modules (implementation model)? It means that the implementation model will probably represent the theoretical model incompletely (and the theoretical model may in turn implement the actual phenomenon incompletely), resulting in lower fidelity of simulation, and less predictive power. Depending on the questions asked of such a system, the answer may be wrong, or only correct to certain approximation.

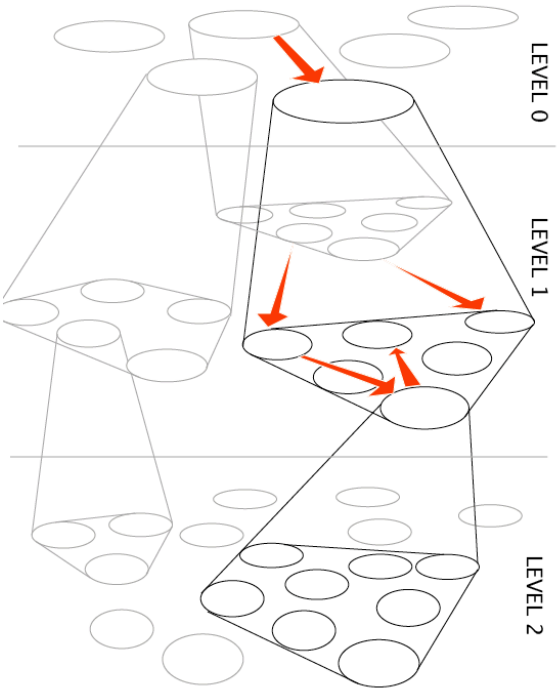


Fig. 6. Abstraction levels: Each level of description can be broken into smaller constituents that interact through rules different from the level above. A causal connection at Level 0 (top arrow) may in fact stand for a complex set of causal interactions at the next level below (small arrows).

3.3 Model Examples

Abstract modules can be used as building blocks for any sort of system; they can be used to turn architectural ideas into *runnable models* that can be put the scrutiny that only dynamic runnable models can (e.g. interaction with the real world). They have for example been used in one form or other for robotics [53, 54], models of market innovation [39], and neurocognitive modeling [55].

To exemplify the use of abstract modules in the context of multimodal communication, we will now take a brief look at two systems that use abstract modules, both implementing turntaking skills. The first system implements a new version of the Ymir Turntaking Model (YTTM [56]), a model based on a broad set of psychological research on human face-to-face communication. The model incorporates multiple modes and has been tested extensively in realtime dialogue with people. A recent implementation of the model is speech-only but has several new perception mechanisms and a new system for managing real-time decision-making and planning. The model is built using around 20 abstract modules that implement various functions such as managing architecture-wide semi-global (internal) states that concern realtime resource (CPU) allocation, decisions and perceptual tasks, as well as speech recognition and speech synthesis (one module each). To take an example, intonation is processed in a special prosody processing module, decisions when to take and give turn are managed by a group of decider modules and the decision to start speaking is managed by a relatively large, modular planner. Most of these modules take input from 2-3 other abstract modules in the system. None of the implemented modules are purported to map directly, or even indirectly, to brain or cognitive "modules". But they *are* assumed to implement functionalities that influence each other in the way that the system architecture implies.

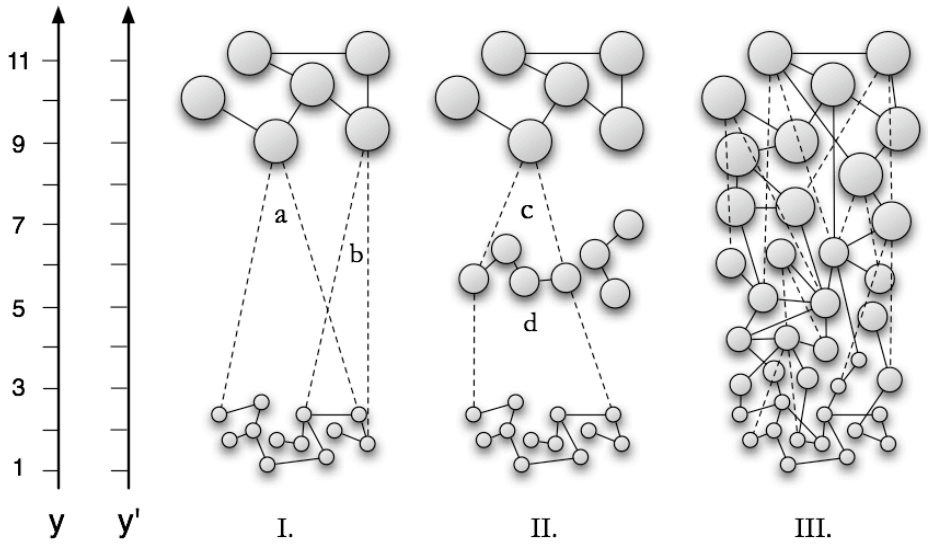


Fig. 7. All systems can be decomposed at multiple levels of abstraction; the more complete a theory is, the better such decompositions are theoretically connected. Here, circles represent decompositions of a system into abstract modules – each module containing an appropriate method for its function, e.g. an artificial neural net, rule set, fuzzy logic, etc.). The lowest level (bottom) represents the finest-grained grouping of functions and/or structure in the system; the highest level (top) represents the most abstract; lines represent coupling. Example I: Modularization at two scales; example II: modularization at three scales; example III: modularization and decomposition at multiple, overlapping scales. For number of neurons in the brain $y' \approx 10^y$ neurons.

In this system the density of the coupling between modules is moderate (averaging 5 connections per module), information transfer between modules is on the order of 6-12k bits per second per module, and the coupling tightness is relatively high, implying that the majority of the modules operate highly predictably based on their input, and thus embody relatively simple internal processes.

Although the YTTM so far seems a reasonable initial step in the direction of modeling complex multimodal realtime turntaking, significant additional research is needed, in particular regarding how the behavior produced by the model compares to real human data and whether the factors (causal chains) it proposes can be implemented by neural mechanisms – a necessity for any model claiming cognitive realism. Both of these are currently work in progress – the latter to be discussed in our next example.

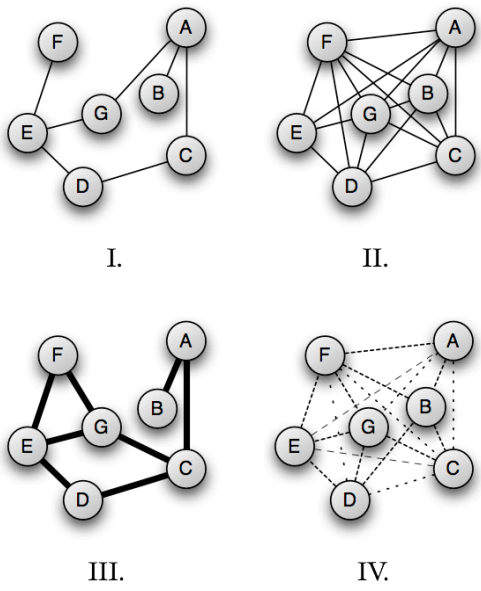


Fig. 8. Types of couplings between a set of heterogeneous abstract modules. I: Sparse coupling. II: Dense coupling. III: Tight coupling. IV: Loose coupling.

The second example of a system we will look at exemplifies how abstract modules can help connect theories at different levels of abstraction. In this system, selected parts of the YTTM, more specifically some of the abstract modules it proposes for handling decision and planning in multimodal turntaking [55], have been implemented as neural mechanisms. The neural planner is a modified version of the Augmented Competitive Queuing model (ACQ [57]), which was built to model Macaque brain mechanisms that control grasping [58, 59]. This new implementation of the YTTM gives it learning capabilities, but the key advance on other implementations of it is the increased level of detail in the implementation and its link to brain research.

At the neural level the model now proposes *motor schemas* that compete for execution in a way that produces emergent action sequencing. Motor schemas are a kind of abstract modules that map directly onto purported neural mechanisms. Albeit somewhat simpler than the original decision and motor control mechanisms in the YTTM, the replaced sections are now anchored to empirical brain modeling: rather than being purely motivated by data from gross-behavior analysis they are theoretically motivated at a much finer level of granularity. Because of the flexibility of abstract modules, some parts of this neural implementation of the YTTM still include abstract modules that have not yet been linked to models from other fields, including modules that control motivational levels and perception of speech and

gesture: these have been implemented to handle particular functions without any claims that their existence is somehow based on existing modularity found in the real world. In accordance with the above account, however, the factors these modules control, e.g. motivational level, are assumed to represent real-world factors. Because abstract modules help isolate logical, as well as structural, parts of a system in its model, modeling can be done without necessarily representing the whole system at a single level of abstraction.

These two models of embodied multimodal turntaking together exemplify how levels of abstraction can be bridged with the help of abstract modules; the synthesis of the original cognitive turntaking model and a neural model of planning and motor control produced a mixed-level, mixed-abstraction approach, bridging cognitive and neural levels in a manageable way, and perhaps more importantly, in a way that can produce performance data that can be directly compared to the systems being studied.

4 Neurocognitive Architecture

One of the questions that arises when modeling cognitive phenomena as varied as those encountered in natural communication is whether we are likely to encounter the need to have our abstract modules represent actual modules. Put in a different way, is the brain/mind modular, and if so, to what extent? (That is, are there systems or mechanisms that clearly, or perhaps not so clearly, form spatial and/or structural groups?) One way to begin to answer this is to look at data produced by recent brain research, since neural structures is ultimately how all animal cognitive functionality must be implemented.

Modularity and hierarchy are known to exist in the human brain [60]. The smallest scale thought to matter to its computation is the neuron and the set of chemical compounds known to be able to alter the computational characteristics of these. There are about 500 major groupings of neurons in the human brain where the groups are composed of a selected set of neurons; each of the groups consists of 5 types of neurons, on average, each of which sometimes creates sub-groupings [61]. The human brain thus consists of a total 2500 different types of neurons. Each of the 500 groupings uses specific ways to compute, and each connects to other groups in specific ways. These groups then make up larger interconnected groups, many of which are dedicated to a particular part of mental processing such as vision, hearing, sight, motion control, speech generation, speech understanding, balance, emotions, etc. [61]. Of course these facts are not a proof that cognitive and brain architecture is isomorphic, but it hints strongly at modular organization on *some* levels of analysis. Further signs of modularity can be seen in the numerous gross anatomical areas identified in the last 30-40 years that have particular functional characteristics, including the several layers of visual processing in the back of the brain, neural nuclei for spatial hearing, learning (cf. [62, 63]), cognitive control and planning [64. 65], and more recently, strong evidence of *causal connections* between frontal lobes and decision making, obtained using transcranial magnetic stimulation [66]. These correspondences between mental-level phenomena and neural tissue, which is composed of heterogeneous types of neurons, give us some hope that a modeling methodology based on abstract modules will be relevant and successful in modeling

mental phenomena computationally, including conversation skills. There exist, however, reasons to believe the opposite: Results from the complete mapping of all 302 brain cells of the nematode C. elegans, and their 7000 synaptic connections [67], have not resulted in any significant deepening of our understanding of how a brain operates, and even simple tasks such as the nematode's crawling seem as unexplained as ever, from the perspective of its brain and neural architecture (cf. [68]).

Elsewhere I have argued that AI researchers need to study intelligence in a larger context than they typically do, and that to do so they need new tools, methodologies and collaboration strategies to build larger models of the phenomena they are studying [49]. The same can be said about cognitive scientists and psychologists working on understanding human communication and cognition in general. Again, it is imperative that we build *runnable models* of these phenomena – it is the only (known) way to address the complexity explosion that happens when we try to understand larger parts of the human and animal minds. It is equally imperative that we aim at modeling these phenomena *in toto* – as completely and comprehensively as possible. [11] It is not enough for psychology to limit itself to (observable) behavior, or to a purely cognitive level – data, theories and methods from neurology, medicine, artificial intelligence and other fields must be used to help constrain the vastness of the possible explanations for the observable surface phenomena (even those produced in controlled experiments with large and repeatable results).

So to understand complex systems such as the human mind/brain system we are going to need models at various levels of detail, for various purposes. In employing this method, one can choose the level of abstraction in accordance with the desired model resolution, available data, and available measuring techniques – in essence employing numerous instances of the model in Figure 4 at various levels in Figure 6, and then expand as more information comes to light. Scientific theory thus becomes built up over time, incrementally covering and explaining more phenomena and excluding alternative explanations. Evidence from research based on these assumptions, including the use of various versions of hierarchies of abstract modules, points to significant benefits of the approach (cf. [33, 40, 53, 55, 64, 71]). Eventually we will want them to be interconnected enough to present the "mind atlas", with details of fine and gross anatomy and function equally represented in a runnable simulation. What that model will look like is an empirical question; however, the evidence so far indicates that it would be equally absurd to expect the human mind/brain to be exclusively built out of a huge number of uniform, specialized modules (what Bryson [69] calls "vertical modules"), as advocated – to different degrees – by e.g. some "massive modularity" hypothesis enthusiasts (see Carruthers [70] for a review), as it is to contend that the human mind/brain is a massive collection of undifferentiated neurons with no discernable low or high-level structures – neural spaghetti. Barrett and Kurzban [72] provide a thorough overview of arguments on both sides of the modularity debate.

---

[11] In particular, a direct result of the requirement for a detailed, comprehensive, runnable model is that our models will – for the most part – *never* be complete; we will be doomed to build models for various purposes, for answering various sets of questions. This can already be seen in modeling efforts for many natural phenomena such as ocean currents, weather systems, ecosystems, etc.

Clearly the human brain, and thus by extension the mind, is organized in many ways and on many levels, but both extremes on the organizational spectrum seem implausible. Assuming then that the substrate of the human mind lies somewhere along the spectrum from being perfectly hierarchically organized to being complete spaghetti, we need powerful tools to analyze and model it.

5 Conclusion

We have painted a picture of multimodal dialogue as a complex system composed of a large set of functions, produced by a multitude of systems and subsystems that interact in complex ways, producing emergent properties. With examples from cellular automata, cognitive modeling and brain research, I have argued that because of this, if we wish to obtain a comprehensive model of multimodal embodied communication, we cannot apply a divide-and-conquer approach exclusively, or only study the various (surface) dialogue phenomena or brain function in isolation – we have to take an integrative approach. Because of its complex-system nature, multimodal realtime dialogue calls for an approach that can model it at a high level granularity – a necessity for achieving correctness in models of complex systems.

The evolution of science rests on the power of the tools we have at our disposal; the need for more powerful tools and methodologies for extending – and especially for interconnecting – research in psychology, brain science and cognitive science is as calling today as the need for the microscope was in the early days of biology. Whether or not one takes to either of the extremes in the mind-brain debate (extreme modularity or "neurocogntive spaghetti"), hierarchies of abstract modules can be a powerful approach to modeling large systems with complex causal structure.

Abstract modules are an embodiment of abstracted functionality. They have been used in many systems to date, from humanoid robotics [40, 53, 54] to integrative cognitive models [55, 57, 58] to economic simulations [39], and shown to be a flexible and malleable methodology. Benefits of the approach are numerous, the main ones being (a) easier integration of models based on different theoretical foundations and originating in different disciplines, and (b) the ability to manage a greater amount of overall complexity. A third major benefit is (c) the ability to mix different levels of abstraction when building a model, to get increased levels of detail where needed. A fourth major benefit is that (d) isolated models built with the method can more easily be extended, can more easily be related to other models. Last but not least, (e) they can be run as simulations whose performance can be compared directly to the systems being modeled.

We have looked at some evidence from brain research in support of the idea that functional validity coincides with structural validity, i.e. evidence of brain modularity may in some cases result in certain levels of isomorphism between cognitive models and the brain. As we have seen, however, not only can abstract modules be used for understanding behavior and cognition independently of whether the phenomena modeled are modular or not – that theory and implementation at different levels of abstraction can coincide but need not do so – but that the approach is also independent of physical and functional modularity (or monolithicity); functional and structural validity can be completely disjoint or can overlap to differing degrees in

models based on abstract modules. As long as causal relationships are correctly identified and represented by the theory, abstract module hierarchies can be used to implement any theory as a runnable computer model, with the necessary and appropriate abstractness, fidelity and predictive power. For complex systems such as multimodal dialogue skills, which calls for modeling a wide range of realtime cognitive skills ranging from the control of saccades to social interaction, abstract module hierarchies represent a powerful and proven approach to scientific research.

Acknowledgments. For insightful and illuminating discussions on numerous topics related to this paper I would like to thank Deepa Iyengar, Hannes Högni Vilhjálmsson, Eric Nivel and Hrafn Th. Thórisson, as well as my many colleagues at the ZiF in the summer of 2006, whose thoughts and ideas have influenced several parts of this chapter. Special thanks to Hrafn for the cellular automata simulation and to Eric, Hrafn, Hannes, Deepa and Ipke for numerous insightful and brilliant comments on the manuscript. Big thanks to Eric for important last-minute edits and suggestions. Thanks also to the anonymous reviewers for challenging questions. Big thanks to Ipke Wachsmuth and Günther Knoblich for conceiving and arranging the research year on Embodied Communication in Humans and Machines at ZiF in Bielefeld. Thanks to Jimmy Bonaiuto for our very fruitful and fun collaboration on the neurally-based turntaking model and to Gudny R. Jonsdottir for her work on the next generation of the YTTM. This work was supported in part by a Fellowship grant from Zentrum für interdisziplinäre Forschung, a research grant from RANNÍS, Iceland, and by a Marie Curie European Reintegration Grant within the 6th European Community Framework Programme.

References

[1] Gerver, D.: Simultaneous listening and speaking and retention of prose. Quart. J. Exp. Psych. 26(3), 337–341 (1974)
[2] Lee, T.: Simultaneous Listening and Speaking in English into Korean Simultaneous Interpretation. Meta, 44(4) (1999), http://www.erudit.org/revue/meta/1999/v44/n4
[3] Simon, H.A.: Can there be a science of complex systems? In: Y. Bar-Yam (Ed.), Unifying themes in complex systems: Proceedings from the International Conference on Complex Systems, pp. 4-14. Perseus Press, Cambridge (1999)
[4] Magnusson, M.S.: Understanding Social Interaction: Discovering Hidden Structure with Model and Algorithms. In L. Anolli, S. Duncan Jr., M. S. Magnusson and G. Riva (Eds.), The Hidden Structure of Interaction: From Neurons to Culture Patterns. IOS Press, Amsterdam (2005)
[5] Card, S.K., Moran, T.P., Newell, A.: The Model Human Processor: An Engineering Model of Human Performance. In Boff, K.R., Kaufman, L., Thomas J.P. (eds.) Handbook of Human Perception Vol. II. John Wiley and Sons, New York (1986)
[6] Chandrasekaran, B., Josephson, S.G.: Architecture of Intelligence: The Problems and Current Approaches to Solutions. In: Honavar, V., Uhr, L. (eds.) Artificial Intelligence and Neural Networks: Steps Toward Principled Integration. Academic Press, San Diego (1994)

[7] Lieberman, M.D.: Reflective and Reflexive Judgment Processes: A Social Cognitive Neuroscience Approach. In: Forgas, J.P., Williams, K.R., von Hippel, W. (eds.) Social judgments: Implicit and explicit processes, pp. 44-67. Cambridge University Press, New York (2003)
[8] Nanayakkara, T., Shadmehr, R.: Saccade Adaptation in Response to Altered Arm Dynamics. J. Neurophysiol., 90, 4016-4021 (2003)
[9] Spivey, M.J., Tannenhaus, M.K., Eberhard, M.K., Sedivy, J.K.: Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. Cogn. Psych., 45, 447-481 (2002)
[10] Thórisson, K.R.: Computational Characteristics of Multimodal Dialogue. AAAI Fall Symposium on Embodied Language and Action, Massachusetts Institute of Technology, Cambridge, MA, November 10-12, pp. 102-108 (1995)
[11] O'Connell, D.C., Kowal, S., Kaltenbacher, E.: Turn-Taking: A Critical Analysis of the Research Tradition. Journal of Psycholinguistic Research 19(6), 345-373 (1990)
[12] Sacks, H., Schegloff, E.A. Jefferson, G.A.: A Simplest Systematics for the Organization of Turn-Taking in Conversation. Language 50, 696-735 (1974)
[13] Wison, M., Wilson, T.P.: An oscillator model of the timing of turn-taking. Psychonomic Bulletin and Review 12(6), 957-968 (2005)
[14] Goodwin, C.: Conversational Organization: Interaction Between Speakers and Hearers. Academic Press, New York (1981)
[15] Duncan, S.Jr.: Some Signals and Rules for Taking Speaking Turns in Conversations. J. of Personality and Soc. Psych. 23(2), 283-292 (1972)
[16] Wolfram, S.: A New Kind of Science. Wolfram Media (2002)
[17] Levi-Strauss, C.: The family. In: Shapiro, H. (ed.) Man, culture and society. Oxford University Press, New York (1956)
[18] Popper, C.: Conjectures and Refutations: The Growth of Scientific Knowledge, Routledge, London (1963)
[19] Newell, A.: You can't play 20 questions with nature and win. In W. G. Chase (ed.) Visual information processing. Academic Press, New York (1973)
[20] Kosslyn, S.M.: You can play 20 questions with nature and win: Categorical versus coordinate spatial relations as a case study. Neuropsychologia 44(9), 1519-23 (2006)
[21] Turing, A.M.: Computing machinery and intelligence. Mind 59, 433-460 (1950)
[22] Turing, A.M.: On Computable Numbers, With an Application to the Entscheidungsproblem. Proceedings of the London Mathematical Society, Series 2 42 (1936)
[23] Laughlin, S.B.: The Implications of Metabolic Energy Requirements for the Representation of Information in Neurons. In: Gazzaniga, M.S. (ed.) The Cognitive Neurosciences III, pp. 187-196. M.I.T. Press, Cambridge (2004)
[24] Thagard, P.: Mind: Introduction to Cognitive Science, 2nd edn. MIT Press, Cambridge (1996)
[25] Chalmers, D.: Does a Rock Implement a Finite State Automaton? Synthese 108, 310-333 (1996)
[26] Chalmers, D.J.: A Computational Foundation for the Study of Cognition. Philosophy-Neuroscience-Psychology Technical Report 94–03, Washington University (1994)
[27] Calabretta, R., Parisi, D.: Evolutionary Connectionism and Mind/Brain Modularity. In: Callebaut, W., Rasskin-Gutman, D. (eds.) Modularity: Understanding the Development and Evolution of Natural Complex Systems, pp. 309-330. MIT Press, Cambridge (2005)
[28] Takahashi, K.: Development of Holistic Climate Simulation Codes for a non-Hydrostatic Atmosphere-Ocean Coupled Systems. In: Annual Report of the Earth Simulator Center, Japan, April 2004 – March 2005, pp. 52-67 (2005)

[29] Vlachos, D.G.: A review of multiscale analysis: Examples from systems biology, materials engineering, and other fluid-surface interacting systems. Adv. Chem. Eng. 30, 1-61 (2005)
[30] Abadi, M.G., Navarro, J.F., Steinmetz, M., Eke, V.R.: Simluations of Galaxy Formation in a Lambda CDM Universe II: The Fine Structure of Simulated Galactic Disks. Astrophys. J. 597, 21-34 (2003)
[31] Wildberger, A.M.: A.I. and Simulation. Simulation 1-2 (March 1999)
[32] Simon, H.A.: Complex systems: The interplay of organizations and markets in contemporary society. Computational & Mathematical Organization Theory 7(2), 79-85 (2001)
[33] Scwabacher, M., Gelsey, A.: Multi-Level Simulation and Numerical Optimization of Complex Engineering Designs. 6th AIAA/NASA/USAF Multidisciplinary Analysis & Optmization Symposium, Bellevue, WA. AIAA-96-4021 (1996)
[34] Schaffner, K.F.: Reduction: the Cheshire cat problem and a return to roots. Synthese 151(3), 377-402 (2006)
[35] Gaud, N., Gechter, F., Galland, S., Koukam, A.: Holonic Multiagent Multilevel Simulation Application to Real-time Pedestrians Simulation in Urban Environment. Proceedings of IJCAI-07, pp. 1275-1280 (2007)
[36] Dayan, P.: Levels of Analysis in Neural Modeling. In Encyclopedia of Cognitive Science. London, England: MacMillan Press (2000)
[37] Arbib, M.A.: Levels of Modeling of Visually Guided Behavior (with peer commentary and author's response), Behavioral and Brain Sciences 10, 407-465 (1987)
[38] Bakker, B., den Dulk, P.: Causal Relationships and Relationships between Levels: The Modes of Description Perspective. In Hahn, M., Stoness, S.C. (eds.) Proceedings of the Twenty-First Annual Conference of the Cognitive Science Society, pp. 43-48 (1999)
[39] Saemundsson, R., K. R. Thórisson, G. R. Jonsdottir, M. Arinbjarnar, H. Finnsson, H. Gudnason, V. Hafsteinsson, G. Hannesson, J. Ísleifsdóttir, Á. Th. Jóhannsson, G. Kristjánsson, S. Sigmundarson: Modular Simulation of Knowledge Development in Industry: A Multi-Level Framework. WEHIA – 1st International Conference on Economic Sciences with Heterogeneous Interacting Agents, University of Bologna, Italy, 15-17 June (2006)
[40] Thórisson, K. R., H. Benko, A. Arnold, D. Abramov, S. Maskey, A. Vaseekaran: Constructionist Design Methodology for Interactive Intelligences. A.I. Magazine 25(4), 77-90. Menlo Park, CA: American Association for Artificial Intelligence (2004)
[41] Fink, G. A., N. Jungclaus, F. Kummer, H. Ritter, G. Sagerer: A Distributed System for Integrated Speech and Image Understanding. International Symposium on Artificial Intelligence, Cancun, Mexico, pp. 117-126 (1996)
[42] Fink, G. A., N. Jungclaus, H. Ritter, G. Saegerer: A Communication Framework for Heterogeneous Distributed Pattern Analysis. International Conference on Algorithms and Architectures for Parallel Processing, Brisbane, Australia, pp. 881-890 (1995)
[43] Martinho, C., A. Paiva, & M. R. Gomes: Emotions for a Motion: Rapid Development of Believable Pathematic Agents in Intelligent Virtual Environments. Applied Artificial Intelligence, 14(1), 33-68 (2000)
[44] Bischoff, R.: Towards the Development of 'Plug-and-Play' Personal Robots. In: 1st IEEE-RAS International Conference on Humanoid Robots, September 7-8, 2000, vol. 8, MIT, Cambridge (2000)
[45] Simmons, R., D. Goldberg, A. Goode, M. Montemerlo, N. Roy, B. Sellner, C. Urmson, A. Schultz, M. Abramson, W. Adams, A. Atrash, M. Bugajska, M. Coblenz, M. MacMahon, D. Perzanowski, I. Horswill, R. Zubek, D. Kortenkamp, B. Wolfe, T. Milam, B. Maxwell: GRACE: An Autonomous Robot for the AAAI Robot Challenge. A.I. Magazine, 24(2), 51–72 (2003)

[46] Simon, H.A.: Near decomposability and the speed of evolution. Industrial and Corporate Change 11(3), 587-599 (2002)
[47] Simon, H.A., Ando, A.: Aggregation of Variables in Dynamic Systems. Econometrica 29, 111-138 (1961)
[48] Fodor, J.: The Modularity of Mind. Bradford Books / MIT Press Cambridge (1983)
[49] Thórisson, K.R.: Integrated A.I. Systems. Minds & Machines, 17:11-25 (2007); Invited paper at The Dartmouth Artificial Intelligence Conference: The Next 50 Years — Commemorating the 1956 Founding of AI as a Research Discipline, July 13-15, 2006, Dartmouth, New Hampshire (2006)
[50] Marr, D.: Vision. W. H. Freeman, New York (1982)
[51] Minsky, M.: The Society of Mind. Simon & Schuster, New York (1986)
[52] Scheutz, M.: When physical systems realize functions. Minds and Machines 9, 161-196 (1999)
[53] Ng-Thow-Hing, V., List, T., Thórisson, K.R., Lim, J., Wormer, J.: Design and Evaluation of Communication Middleware in a Distributed Humanoid Robot Architecture. In: IROS '07 Workshop Measures and Procedures for the Evaluation of Robot Architectures and Middleware, 29 Oct. - 2 Nov. San Diego, California (2008)
[54] Thórisson, K.R., List, T., Pennock, C., DiPirro, J., Magnusson, F.: Scheduling Blackboards for Interactive Robots. Reykjavik University Department of Computer Science Technical Report, RUTR-CS05002 (2005)
[55] Bonaiuto, J., Thórisson, K.R.: Towards a Neurocogntive Model of Multimodal Turntaking. In: Wachsmuth, I., G. Knoblich, G., Lenzen, M. (eds.) Embodied Communication in Humans and Machines, forthcoming. London: Oxford University Press (2007)
[56] Thórisson, K.R.: Natural Turn-Taking Needs No Manual: A Computational Model, From Perception to Action. In: Granström, B., House, D., Karlsson, I. (eds.) Multimodality in Language and Speech Systems, pp. 173-207. Kluwer Academic Publishers, Dodrecht, the Netherlands (2002)
[57] Bonaiuto, J., Arbib, M.A.: What Did I Just Do? A New Role for Mirror Neurons. (in preparation)
[58] Fagg, A., Arbib M.A.: Modeling parietal-premotor interactions in primate control of grasping. Neural Netw. 7-8, 1277-1303 (1998)
[59] Alstermark, B., Lundberg, A., Norrsell, U., Sybirska, E.: Integration in descending motor pathways controlling the forelimb in the cat: 9. Differential behavioural defects after spinal cord lesions interrupting defined pathways from higher centres to motorneurones. Experimental Brain Research 42(3), 299-318 (1981)
[60] Garel, S., Rubenstein, J.L.R.: Patterning of the Cerebral Cortex. In Gazzaniga, M.S. (ed.) The Cognitive Neurosciences III, pp. 69-84. M.I.T. Press, Massachusetts (2004)
[61] Swanson, L.W.: Interactive Brain Maps and Atlases. In: M. A. Arbib and J. S. Grethe (eds.), Computing the Brain, pp. 167-177. San Diego: Academic Press (2001)
[62] Bryson, J.: Modular Representations of Cognitive Phenomena in AI, Psychology and Neuroscience. In: Davis, D., (ed.), Visions of Mind, pp. 66-89. Idea Group, London (2005)
[63] Bryson, J., Stein, L.A.: Modularity and Specialized Learning: Mapping Between Agent Architectures and Brain Organization. In: Wermter, S., Austin, J., Willshaw, D. (eds.), Emergent Neural Computational Architectures based on Neuroscience. Springer, Heidelberg, Germany (2001)
[64] Koechlin, E., Ody, C., Kouneiher, F.: The Architecture of Cognitive Control in Human Prefrontal Cortex. Science, 302, 1181-1185 (2003)
[65] Miller, E.K., Cohen, J.D.: An Integrative Theory of Prefrontal Cortex Function. Annu. Rev. Neurosci. 24, 167–202 (2001)

[66] van 't Wouta, M., Kahn, R.S., Sanfeyd, A.G., Alemanc, A.: Repetitive transcranial magnetic stimulation over the right dorsolateral prefrontal cortex affects strategic decision-making. Cognitive Neuroscience and Neuropsychology 16(16), 1849-1952 (2005)
[67] Oshio, K., Morita, S., Osana, Y., Oka, K.: C. elegans synaptic connectivity data. Technical Report of CCeP, Keio Future, No.1, Keio University (1998)
[68] Zheng, Y., Brockie, J.P., Mellem, J.E., Madsen, D.M., Maricq, A.V.: Neuronal Control of Locomotion in C. elegans Is Modified by a Dominant Mutation in the GLR-1 Ionotropic Glutamate Receptor. Neuron 24, 347–361 (1999)
[69] Bryson, J.: Evidence of Modularity From Primate Errors During Task Learning. Proceedings of The Ninth Neural Computation and Psychology Workshop (NCPW9), Cangelosi, A., Bugmann, G., Borisyuk, R. (eds.). World Scientific (2005)
[70] Carruthers, P.: The case for massively modular models of mind. In R. Stainton (Ed.), Contemporary Debates in Cognitive Science, pp. 205– 225. Oxford, England: Blackwell (2005)
[71] List, T., Bins, J., Fisher, R.B. , D. Tweed, D., Thórisson, K.R.: Two Approaches to a Plug-and-Play Vision Architecture – CAVIAR and Psyclone. In Thórisson, K.R., Vilhjalmsson, H.H., Marsella, S. (eds.), AAAI-05 Workshop on Modular Construction of Human-Like Intelligence, Pittsburgh, Pennsylvania. American Association for Artificial Intelligence, pp. 16-23 (July 10, 2005)
[72] Barrett, H. C., Kurzban, R.: Modularity in Cognition: Framing the Debate. Psych. Rev. 113(3), 628 – 647 (2006)