Tracking agents – the scrub-jay testbed for Minimal Theory of Mind 17.07.10

1. Motivation
2. Against strong bias in explanations for Theory of Mind-abilities
3. Developing Minimal Theory of Mind
   1. Applications in avian research
   2. Principles of Minimal Theory of Mind
   3. Inference-Preference-Goal-agents
   4. Possible biological underpinnings
4. Technical tracking agents
   1. Technical agency and Minimal Theory of Mind
   2. Theory of Mind in robotics
   3. A sketch of a scrub-jay –like companion
5. Discussion
6. Conclusion
7. Motivation

Agents engage Theory of Mind (ToM) to infer mental states, intentions and goals of other agents. ToM is related to goal-setting-theory and belief-desire-intentionality (BDI)-models of agents, which were originally constructed in philosophy of agency and action. However, a long lasting dogma held that only human agents, especially adults, have a ToM. This was probably due to high cognitive demands and complexity of the underlying neural mechanisms: ToM relies on relations between intentional agents and abductions in complex belief-networks. But recently approaches applying ToM-like explanations to infants, chimps, scrub-jays and others have emerged, despite objections by philosophers holding a top down approach to ToM.

In man-machine-systems as well as multi-agent-modeling inferring intentions also is an actual and crucial topic. Giving the user a suitable image of the machine and vice versa demands an adequate realization of behavior. This holds especially for dialogical situations, for the machine shall be able to “understand” user’s aims and respond accurately. This cannot be done with pure behavior reading: recognizing user’s mimics, gestures and prosody is not enough to “understand” a dialogical scene or intentions. But starting with a full blown ToM for technical agents may be like handling all conceptual and implementation-related problems at once. Indeed, it seems unrealistic to believe in an “intention module” generating intentions, which could be implemented literally in technical agents.

Our approach is Minimal ToM, which we have already applied as an alternative explanation for infants, chimps and scrub-jays. It relies on relations between agents and their encountered objects and is quite far away from common sense. This can be realized without referring to intention-reading. Minimal ToM-demands are situated between behavior reading and full blown ToM. In limited situations explanations of ToM and Minimal ToM may partially overlap, enabling agents with Minimal ToM to behave intentional-like and being able to fake a strong ToM. In these situations “faking” allows to give the user a suitable image of the machine without implementing an intentional architecture literally.

We apply Minimal ToM to technical agents and explore applications and biological and technical underpinnings in companion-systems.

1. Against strong bias in explanations for Theory of Mind-abilities

Comparative psychology and ethology, cognitive biology and other fields employ several types of explanation for behavior, e.g. causal and functional explanations. In the last years, researchers in these branches have used more and more Theory of Mind (ToM)-explanations in both humans and other animals, e.g. in infants, chimps and birds. A Theory of Mind-explanation is a hybrid-type of explanation, thus it contains mechanistic (combination of functional and causal) as well as a sort of design explanations.

ToM-explanations refer to ToM-concepts, e.g. intentions, knowledge, beliefs, desires, mental states and mental concepts. A typical ToM-type-explanation may have the following structure:

**ToM-explanation**

A Functional: Agent x performs behavior b to phi.

B Functional: ToM-abilities t are required for b.

C Mechanistic & Design: t is realized by human ToM-cognition tc in x.

--- ---

D Iff A & B & C, then x has ToM.

Clearly, C is a difficult constraint of the Scopus of ToM-explanations. First, it refers to an empirical statement: ToM-cognition (the underlying mechanisms) may be species-bounded. Second, it refers to a methodology: a top down-approach is used to compare humans and other species, starting with human concepts of agency. This is also a design explanation, because humans designed an ideal of agency: the belief-desire-intentionality (BDI)-agent. Consequently, the designs of underlying mechanisms are also compared to an ideal of human mechanisms and inferred agency. For example, typically ToM is tight to human language.

Both constraints together are comparable to a bias in research: instead of comparing behavior and underlying mechanisms in face of ecology and contexts, researchers typically reject ToM-explanations applied to animals for the following reasons:

(R1) Conceptual cognitive costs and complexity

(R2) Evolutionary distance between human and other species

(R3) Scepticism of non-human-cognition: top down approach to intentionality, commitment to human language as a condition sine qua non for ToM

(R4) Easier explanations may be available

In view of these objections, ethologists have three possibilities to apply ToM-explanations in animals:

E1) Missing evidence: One argues for further research for similar realizers in animals as well as humans, such that tc in humans may be tc in animals.

E2) Evolutionary progress: one argues that realizations found in animals and humans are in a linear sequence of progression; such that for example chimp’s mechanisms would be evolutionary predecessors of man’s mechanisms. For this reason they may have preserved similar functions. Wedded to this theory are the Aristotelian “scala naturae” and modern approaches of evolutionary macro-trends like “progress”.

E3) Counterexamples against strong assumptions: One could argue against conditions held to be necessary with the help of empirical counterexamples, e.g. in the case of ToM and linguistics in humans (Varley, R., & Siegal, M. Evidence for cognition without grammar from causal reasoning and ‘theory of mind’ in an agrammatic aphasic patient. Current Biology 2000 10(12):723-726.).

E4) Easier explanations may be ruled out by sharp experimental designs and control experiments.

For likely any empirical case, there will be a difference in detail between human and animal agents in tc, e.g. it is discussed, whether rats have a prefrontal cortex or to what degree the avian hippocampus is similar to the human hippocampus. Unfortunately, homology is an open field.

For this reason, E2) & E3) seem to offer an ideal way for ethology: while assuming a gradual continuity due to evolution between animals and humans, one wants to apply gradually human-concepts in animals. Indeed, this is for example the case for literally applying instances of Piaget’s cognitive psychology of infant-development in case of object-permanence (Clayton et al. 2009). They stated that birds are likely to reach some stages in these tests, given suitable experiments.

But beside these interesting approaches, by no means B is necessarily connected to C: there should be e.g. homolog realizations of ToM-*like* abilities, as it has been the case for many other abilities (this is an open question in research). This offers an alternative bottom up approach, contrary (and maybe complementary) to the top down approach:

**Bottom up approach**

Comparative multilevel-studies of cognition are required in research, grounded in a bottom level: “The most logical route for comparative cognition, however, is to try to understand the basic processes and common denominators first exploring species-typical specializations.” (c.f. Frans B.M. de Waal1 and Pier Francesco Ferrari. Towards a bottom-up perspective on animal and human cognition. Trends in Cognitive Sciences, Volume 14, Issue 5, May 2010, Pages 201-207).

Prominent examples of comparative cognition are functions of social cognition. Indeed, sociality and social cognition are wide-spread phenomena found in primates, birds, or even in some form in swarms of fishes. In general, abilities refer to functions, which are underdetermined by structure; thus t could be realized by *other* mechanisms than tc at least in principle. Recent evidence suggests that this could be the case: higher forms of cognition *are* realized differently in, e.g., mammalians and avian creatures (c.f. Avian brains and a new understanding of vertebrate brain evolution. Nature Reviews Neuroscience 2005 6:151-159.; Salwiczek, Lucie H.; Watanabe, A.; Clayton N. S. Ten years of research into avian models of episodic-like memory and its implications for developmental and comparative cognition. Behavioral Brain Research 2010 doi: 10.1016/j.bbr2010.06.011;).

In case of ToM-abilities this leads to the call for species-adequate explanations. Due to the possibility of homologs on the mechanistic level and convergent evolution on the conceptual (ability-) level, these different mechanisms do not have to stand necessarily in *a linear sequence* of evolutionary successors, though their functionality may overlap (see claim 1). It will be an issue to find the adequate parameters to map ToM and other explanations in conceptual and possible mechanistic spaces. This lead to terms like “episodic-like” or “mental-like”, thus a typical ToM-like explanation may have this schema:

A Agent x performs behavior b.

B b requires some ToM-abilities t.

C’ ToM-like explanation is conceptually agnostic against tc – t could be realized by tc or others.

--- ---

E1 Weak inference to best explanation: Iff (A & B) and ◊C is known, then x has a ToM-like capacity.

E2 Strong mechanistic explanation: Iff (A & B) and C is known, then x has a ToM-like capacity.

**Minimal Theory of Mind (Minimal ToM)**

With respect to objections to ToM-explanations, a ToM-like explanation that may not rely on costly and disputatious ToM-concepts (like intention-reading) would be preferable for candidates like chimps, infants or jays. Therefore, there can be two issues regarding the focus of a ToM-like explanation:

1. Explaining behavior by ToM-abilities, but conceptually and maybe mechanistically with the help of a ToM-like explanation. This intends that the same abilities can be realized by different mechanisms. For example, some ability of social cognition may be conceptually the same in human and chimp, but is differently realized.
2. Explaining behavior by ToM-like-abilities with the help of a ToM-like-explanation. This intends that these abilities are just similar or maybe somehow compatible (trackable) to ToM-abilities. For example, some ability of social cognition in human and chimp is just similar and differently realized, but may allow the chimp to track the human in some limited situations (and vice versa).

We intend to care for both options, though it may be difficult and unintended to rule out the possibility of the other option in a given case. Generally, and as it was discussed earlier, the commitment to the possibility of convergent evolution, nonlinear evolution and progression requires a viable notion of homology (of e.g. brain areas) and similarity (of e.g. abilities). This is a hot topic in research and solving these problems is beyond this article. However, clarification of possible explanations and alternatives will help.

As we will discuss deeper later on, one must decide between necessary and sufficient conditions for attribution of ToM or alternatives. The fulfilment of all necessary conditions gives no guarantee for sufficiency; however, there might be the hope that many necessary conditions contain automatically the (possibly unspoken) sufficient conditions. (Silverman, P.S. Attributing mind to animals: The role of intuition. *J. Social Biol. Struct.* 1983 6: 231-247.) This is the caveat of necessity.

Giving such an explanation in the sense of option 2) is possible in *limited situations*, as Butterfill and Apperly (2009) showed: Minimal ToM. In view of E, there are two complementary research agendas:

A1) Weak agenda: Giving conceptual and mostly descriptive explanation.

A2) Strong agenda: Giving generative and mechanistic explanations according to biological or technical underpinnings with testable hypotheses.

Our aim is to offer future directions for a strong agenda in both biological and technical areas. We begin with biological applications in the next chapter.

1. **Developing Minimal Theory of Mind**
   1. Applications in avian research

Minimal ToM can be an alternative ToM-like explanation in avian research. For further investigations, we sketch two exemplary scenes with western scrub-jays.

**Scene 1: Pilfering and object permanence**

There is a lively debate concerning whether scrub-jays are able to adjust their behavior using complex cognitive strategies. In one scrub-jay study it was found that jays changed their behavior after observing other jays pilfering, which was used to argue for complex observational understanding (Emery NJ, Dally JM, Clayton NS. Western scrub-jays (Aphelocoma californica) use cognitive strategies to protect their caches from thieving conspecifics. Anim Cogn. 2004 Jan;7(1):37-43. Epub 2003 Jun 26.) By contrast, chickadees do not change their behavior if they are victims of pilfering just once (Baker, M. C. & Anderson, P. Once-pilfered cache sites not avoided by black-capped chickadees. Anim. Behav. 49, 1599-1602 (1995).).

This and other experiments indicate that jays have a complex understanding of other jays’ behavior and use it to adapt their own behavior. Beside phylogeny, a further interesting appeal in Theory of Mind is its development in ontogeny in humans.

In a recent article, Salwiczek et al. deal with the development of Piagetian object permanence in western scrub-jays’ ontogeny: being able to deal with visible and invisible objects is a necessary condition for caching and pilfering (Salwiczek, Lucie H.; Schlinger, Barney; Emery, Nathan J.; Clayton, Nicola S. The development of acting and object permanence in western scrub-jays (Aphelocoma californica: Which emerges first? J Comp Psychol. 2009 123(3):295-303. Salwiczek et al. argue for application of Piaget’s schema of six developmental stages)

|  |
| --- |
| ToM-explanation terms in scene 1 |
| Observational understanding in pilfering includes maybe a kind of perspective-taking  Possessing episodic(-like) memory to remember when and where to pilfer what |
| Possessing concepts of visibility and invisibility for objects (pilfering and caching) |
| Ontogeny of object-related concepts is tight to higher forms of learning |
| Physical cognition of object-pushing (causal cognition) |

**Scene 2: Catching a scene by use of modality-related information**

Stulp et al. 2009 did a behavioral experiment to investigate in western scrub-jay’s capacity to use scene-related modal information. In this experiment, western scrub-jays had to cache either 1) alone in the cage, 2) in the presence of another one in a cage split by a transparent plate (seen + heard), 3) in the presence of another one in a cage split by an opaque plate (heard but not seen). One tray was prepared in a noisy condition such that caching produced noise, the other tray was prepared in a quiet condition. Briefly, in 1) and 2) jays preferred the noisy tray, whereas in 3) they used the quiet tray. Because jays’ preference was not determined by the presence of another jay alone (they also used the noise tray in 1)), Stulp et al. concluded “that food-caching western scrub-jays conceal auditory information if-and only if-the competitors can hear, but cannot see the cachers”

|  |
| --- |
| ToM-explanation terms in scene 2 |
| Assessing the other’s possible perspective requires perspective-taking |
| Possessing concepts of possible scopes of different modalities of a scrub-jay related to the scene  Possible possessing a self-concept and related body-schema |
| Physical cognition of scene-related features in terms of causal cognition |

Both scenes are typically explained with the help of ToM in context with the development of complex cognitive strategies. A very similar process can be found in primatology (e.g. Call, J., & Tomasello, M. Does the chimpanzee have a theory of mind? 30 years later. Trends in Cognitive Sciences 2008 12(5):187-192.). In the next chapter, we introduce alternatives of explanation for ToM: Minimal ToM.

**3.b Principles of Minimal Theory of Mind**

The following approach is thought to be only one of possibly many alternative approaches. It rests on the following four claims and was introduced in Butterfield and Apperly 2010. Then we introduce a first set of five possible concepts forming a corpus of an alternative explanation.

M1) Conceptual claim: Properties

Some Minimal ToM-states share properties of ToM-states. This may not be related to the level of cognition but abilities. In limited situations it is possible to use proxies for reasoning about beliefs and fake a full blown ToM.

M2) Non-conceptual claim: Fewer resources needed

Minimal ToM has states demanding less cognitive and conceptual resources than ToM-states.

M3) Empirical claim: Explanatory relevance

Some of the ToM-abilities found in infants, scrub-jays or chimps are consequences of having a Minimal ToM. A lower level of explanation might be also true in some situations (e.g. Penn&Povinelli) for scrub-jays or humans.

M4) Empirical speculation: Minimal ToM and full blown ToM may be realized in parallel (but maybe not in synchronization) in humans.

These claims follow the aim to establish a strong research agenda that does not only offer alternative concepts. Thus, we state empirical claims of explanatory relevance and speculate: If Minimal ToM and ToM are realized by different mechanisms, the case that humans can possess both ones would be an interesting empirical challenge. Are there conflicts between the systems? Are their outputs integrated, synchronized or follow completely other pathways?

In the next chapter we introduce an exemplary conceptual framework of five design principles. Then we bring them together with the two scenes above to show explanatory relevance.

**A subsumption-architecture for Minimal Theory of Mind**

Butterfield and Apperly 2009 suggested five concepts, which may not be 1) the only ones, 2) a closed set, 3) literally true. These concepts are:

C1) encountering: an object or another agent is in the agent’s perceptual field.

C2) registration: an object or another agent is actively registered by the agent.

C3) goal-directed actions exist and can be represented: an agent is able to recognize goals and engage in goal-directed actions (instead of just object-direct actions).

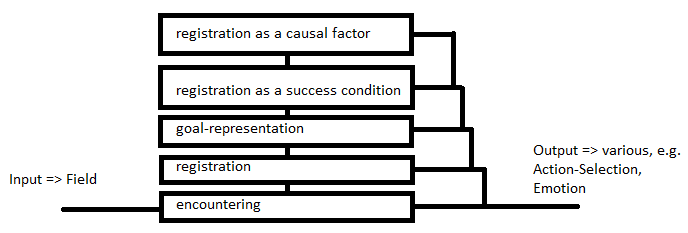
C4) registration as a success-factor: correct registration is needed for a successful goal-directed action involving the object.

C5) registration as a causal factor: registration is behaviorally relevant.

Conceptually, each set of the first n of these concepts of Minimal Theory of Mind can be arranged in a subsumption-architecture (after Brooks), which has the following properties:

1. n-consistency: Every set of the first n concepts of this architecture is consistent for itself. For example, encountering & registration ends in behavior reading. Higher arranged concepts depend on lower arranged ones, but feedback down to the output of next lower one again.
2. Compositionality: Two consistent sets may be composed to generate a higher-order-architecture. These are not necessarily the five concepts shown here; one could add further concepts (see above).
3. Tracking-Capacity: The bigger the n, the bigger the creature’s capacity to track higher-order-architectures (the more ways to handle cues by composition).

We arrange these concepts in a subsumption-architecture.



For example, packing the set {encountering, registration} is enough to characterize pure behavior reading. Encountering is a necessary condition for registration, because what is not in the field cannot be registered. The output of the registration-process is fed back to the encountering level, where it may be signaled to an action-selection module, e.g. to increase attention or enable a behavior-adequate motor-action.

We return to this schema again in the second part of this article, where we discuss technical agency. At first, we introduce an agent-schema complementary to BDI-agents and suitable for ToM-like explanations.

**3.c Inference-Preference-Goal-agents**

Theory of Mind and BDI-agents do fit. For alternative ToM-like-explanations, the application of BDI-schema may not fit, e.g., if no beliefs are held by the agent. For these creatures we suggest another schema according to our set of concepts for Minimal ToM given above.

Furthermore, because Minimal ToM claims to be explanatory relevant in some cases of ToM-abilities, the space of explanations for these abilities cannot be drawn one-dimensionally. For, there are at least two theories of ToM-abilities or ToM-like-abilities and their possible realizers.

It is important to keep in mind that Minimal ToM cannot be only thought as a predecessor of ToM, but also as a complementary or different approach. It will be an interesting empirical business to draw ToM-phylogenies (Tecumseh Fitch, W.; & Huber, L.; & Bugnyar, T. Social Cognition and the Evolution of Language: Constructing Cognitive Phylogenies. Neuron 65(6):795-814.).

Let us question, what *parameters and scales* are adequate to classify ToM and alternatives? We suggest distinguishing 1) the conceptual level of abilities and the mechanistic level of realizations, and 2) the notions of goal, preference and inference for abilities of agents, differed by their complexities.

The levels of concepts and mechanisms follow from the possibility of different mechanisms for the same or behaviorally similar ToM-abilities. Furthermore, the notions of goal, preference and inference follow from the set of five conceptions given above. They can be inferred from the greater framework of tracking agents, too (see the following paragraph and chapter 4).

Theory of Mind and alternatives describe abilities to track other agents. Conceptually, this means to track the components of agent-schemas, for example to track intentions in BDI-agents. Because chimps, infants and scrub-jays may not be able to track beliefs, desires and intentions, they may be able to track *proxies* for them instead. These proxies may be adequate conceptions for non-BDI-agents themselves and in tracking each other, too. We describe them now and then turn to possible biological underpinnings and technical agency.

**Goal**

Let us start with goal, because it is the entry-point for interesting biological cases like scrub-jays. A goal can be defined depending on agent’s demands and facilities:

Definition 1: Goal is an intentionally set element of a plan related to the agent’s beliefs, desires or needs.

This definition would be suitable for a BDI-agent; however, it engages notions of intentionality and beliefs. This may be far too much or would be conceptually different.

Definition 2: Goal is a property of units of behavior (a function in teleological sense), whereas these goals are related to the agent’s needs.

This second notion of goal does not imply human intentionality or complex networks of beliefs, and it does not imply representation of goals as (theoretical) functions. If it is the function of caching to prevent other scrub jays from stealing, this function only occurs with the whole behavioral unit of caching and is not included in a subtask (e.g. digging). It may be reasonable to assume that a scrub-jay is not an expert in the theoretical account of evolutionary biology and has an explicit representation of goal in this manner. However, planning behavior or individual goals should not be excluded; rather we state the behavioral goal can be a proxy for the intentional goal.

**Preference, and Type-representations and Token-representations**

Preferences can be type-like (“Birds (generally) like worms”) or token-like (“Tom likes (this) Mary”). They shape attention and processing in perception like a bias, and do explain a part of the goal-driven activities of creatures. ~~Related to the definition of goal~~, preferences can be defined in these two forms:

1. Type: Preference of Y is a *species-wide* preference to do/like/... rather Y than X.
2. Token: Preference of Y is an *individual* preference to do/like/... rather Y than X.

Which kind of preference an agent can possess depends on its other abilities. For example, an episodic-like memory with a binding function for What-Where-When enables the identification of individual events. This seems to be the case for scrub-jays, but not for pigeons (Salwiczek, Lucie H.; Watanabe, A.; Clayton N. S. Ten years of research into avian models of episodic-like memory and its implications for developmental and comparative cognition. Behavioral Brain Research 2010 doi: 10.1016/j.bbr2010.06.011). However, both jays and pigeons may represent food in a type-manner. Preferences may be proxies for complex desires.

**Inference**

Different species have different capacities to form inferences, which is related to cognitive capacities (e.g. working memory) and standards of rationality (e.g. grades of consistency). At least one species seems to be able to reason in complex causal networks of beliefs (humans), some more species may have a kind of physical or social cognition and possess conceptual representations. Whereas the question for capacities may be genuine empirical, the question for standards of rationality is normative. It is an empirical job to test which creatures are able to maintain some of these rationality-classes. E.g., we can exclude the existence of creatures able to be perfectly rational in our world. Another interesting topic concerns the heuristics that can be used by different types of agents.

Inferences over belief-networks involve relations between propositions and complex concepts, e.g. the self. They are necessary for inferring other BDI-agents’ mental states, having metarepresentational capacities etc. But in limited situations inferences over relations between agents and objects may be enough to act as if one would be able to infer mental states. These agent-object-relations can be proxies for complex belief-networks over concrete agents and objects in a concrete scene.

Summed up, the concept of different agencies concerns a *hypothesis* *about tracking-capacities*:

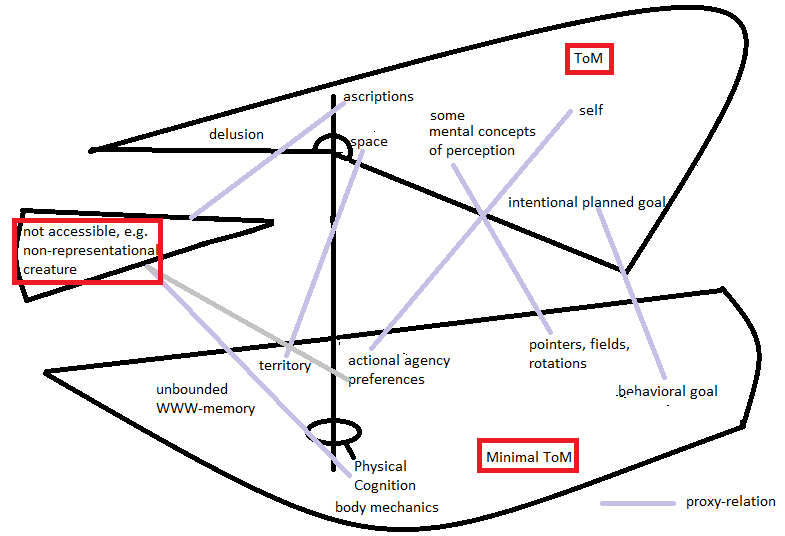
1) Proxies are used by ToM-like-agents (e.g. Minimal ToM) to track ToM-agents.

2) Proxies are used generally to track an agent belonging to another class of agency.

One can easily contrast some schematic differences in conceptions between ToM and Minimal ToM.

|  |  |  |  |
| --- | --- | --- | --- |
| Notion | Theory of Mind | BATCH  TRACK  HANDLE | Minimal Theory of Mind |
| Types of relations | <belief, belief>  <agent, <agent, agent>>  <agent, behavior> <agent, goal>  <agent, object> | <agent, context, behavior, object>  <agent, field>  <agent, behavior> <agent, goal>  <agent, object> |
| Representational capacities | Mental states  Belief, desire, Intention  Episodic memory  Self/Other  Language  Proxemics  Behavioral properties  Goals  Objects  Abstracta  Concreta | Inference, Preference, Goal  Episodic-like memory  Zoosemiotics  Proxetics  Behavioral properties  Goals  Objects  Types  Tokens |
| Examples for problem-solving operations | Mental perspective-taking  Mental rotation  Induction over tokens | (Mental) rotation  Type-learning |

The following map visualizes roughly the concepts so far. Many properties of ToM and Minimal ToM can be tracked; but some may be inaccessible on both sides: e.g. delusion in ToM and unbounded memory in Minimal ToM. There may be other creatures, which are not tractable (due to shared conceptual and biological fundament), but they can be object to ascription or physical cognition.



**Developmental issues**

Human-related ToM-research describes several developmental phases in infants, children and teens. Consequently, standard models of human ToM like Leslie or Baron-Cohen both include the successive emergence of several modules. For example, both models include a succession of physical cognition, actional agency and intentional agency (Leslie: theory of body mechanics, system-1, system-2) and needed modules (e.g. joint attention).

In the case of object permanence also a development of object concept and related activities (e.g. pushing, picking up) is employed. However, this was modelled and explained by a Piagetian theory developed for human infants. This aspect is disputatious, of course, but the general hypothesis seems to be straight:

Developmental hypothesis: Ontogeny plays a role in agents suited with Minimal ToM.

Can creatures evolve during their life? Do learning and development enable or facilitate the emergence of other processing-modules track-capacities? For example, are new types of relations acquirable during lifetime? This issue should be explored in the case of Piagetian development, e.g.

1. **Possible biological underpinnings**

Getting empirical and going from post hoc conceptions to a priori hypotheses is a hard job. Especially, this is the part where we leave the pure conceptual level to find underlying mechanisms. We encourage other researchers to comment or suggest possible mechanisms. According to our earlier stated assumption, Minimal ToM is not necessarily

* a predecessor of ToM
* a part of the ToM-network (e.g. in humans)

Though, missing necessity does not exclude these options, but it also could be:

* a parallel working system (in humans)
* a complementary system to ToM (in humans)
* a case of convergent evolution due to similar ecology (social networks, evolutionary benefit to infer others’ goals or orientation map, ...) and homology (in case of different brains in different species)
* a completely other “operating system” that has evolved and simply uses tractable features to orientate (as if we would use features of alienate species to ascribe, though we are hopelessly unable to know if we really track what we think to track)

The last option opens the theoretical idea of minds that do not have tractable features, or of things that have tractable features though are completely different and challenge our biological bias (just think about Terry Bisson’s story “They’re made out of meat”).

**Humans and primates**

To investigate for Minimal ToM in humans and primates two options are offered:

1. Lower level cognition working first/only/parallel
2. Higher and lower cognition working synchronized

The first option suggests looking for brain areas which do a job without the need for higher cognitive systems. If these areas generate tracking abilities, they may be interesting candidates for a second parallel mechanism of Minimal Theory of Mind. This can be advantageous in case of primates and monkeys or other partly homolog brains and used to argue for parallel systems. For example, category learning is performed by auditory cortex in primates and gerbils.

The second option assumes that ToM is partly involved in processing of higher cognition. The search for parallel working lower systems doing similar jobs in synchrony with higher systems is sometimes offered by lesions or disorders. In these cases lower or higher systems take control exclusively, e.g. in blindsight, frontotemporal dementia, or dissociations of higher and lower systems emerge because of lesions. Then lower systems alone can do parts of the job without consciousness, language or higher cognition, e.g. registration of objects or causal cognition. The direct or indirect interaction between higher and lower systems can be interesting and can possibly be used to argue for ToM being successive.

Arbib recites two patients having lesions in higher or lower pathways. Patient D.F. had a lesion from V1 towards posterior parietal cortex but not from V1 to inferotemporal cortex. D.F. could grasp correctly, though was not able to declare what she does or play pantomime; Patient A.T. had a bilateral posterior parietal lesion and could play pantomime or declare verbally, but could not grasp correctly. Taken from: Arbib, M.A. From monkey-like action recognition to human language: an evolutionary framework for neurolinguistics. Behavioral and brain sciences 2005 28:105-167.

These case studies show how early systems (of visual processing) can be integrated in higher systems (of language faculty and declarative memory). Maybe there is a parallel case in ToM and Minimal ToM. Let us review some candidates of lower systems that may be involved in generating abilities of Minimal ToM.

Entorhinal cortex: microstructure of space-related cells: grid cells, head cells, view cells, …; allocentric perspective-construction; direct biconnection between hippocampus and entorhinal cortex; other like neocortex, subiculum as well; mathematical notion: continous attractor neural networks (mixed discrete-continous allow to represent discrete objects per continous sets of attractors);

Hippocampus: hippocampal place cells

Cerebellum: low level association learning: motor (x) stimulus

Auditory cortex: category-learning; connected with higher cortical areas and other modal cortices

**Non-human primates**

**Birds**

**Others**

1. **Technical tracking agents**
   1. **Technical agency and Minimal Theory of Mind**

In this section we will give a sketch of a scrub-jay-like companion system. Because mind-reading and implementation of intention is a challenging but desired ability, it may be satisfying to use the Minimal ToM-approach in limited situations. How can biological and technical concepts inform each other to help to model and build technical companions?

David Marr suggested three levels of analysis, which are classically used to classify approaches.

|  |  |  |
| --- | --- | --- |
| Computational level | What does the system do and why does it so? | Goal  Representational account |
| Algorithmic level | How can the system’s activity be algorithmically specified? | Method  Connectionist account |
| Implementational level | How is it realized? | Means  Dynamic systems account |

Furthermore, related to the degree of using biological principles or solutions, one can distinguish between biologically informed (copy solutions), inspired (extract principles) and agnostic (do not care) approaches. A classic set up came from Brian Scassellati (Scassellati, B. Theory of Mind for a Humanoid Robot. Autonomous Robots 2002 12(1):13-24), who suggested implementations that can mimic the computational level of biology, for example in case of vision.

According to the distinction of trivial bionics and bionics made by Werner Nachtigall (Nachtigall, W. Bau-Bionik. Springer 2003.), trivial bionics just aims to produce simple copies of nature. Bionics aims to extract the principles to apply them usefully in other fields. This means that we do not intend to copy the biological underpinnings of ToM or Minimal ToM literally, rather the approach should be principled.

**b. Theory of Mind in robotics**

A recent approach dealing with ToM in cognitive robotics situate itself informed on the computational level, whilst being agnostic against the other levels (Butterfield, J., & Jenkins, O. C., & Sobel, D. M., & Schwertfeger, J. Modeling aspects of theory of mind with markov random fields. International Journal of Social Robotics 2009 1(1): 41-51.). Thus they do not *literally* or *necessarily* implement a biologically informed ToM at all levels, but implement a generator being able to reproduce some behavioral outcomes of ToM to explain the *role* of ToM. We describe the approach in Butterfield et al. 2009 and show some constraints and confusions. Finally, we suggest that one should interpret their implementation conceptually as a form of a Minimal Theory of Mind.

Butterfield et al. 2009 use Markov Random Fields to model a two-level-approach to Theory of Mind, whereas each level can be described by one variable (in their terms):

Variable x: Latent mental states, knowledge, plans or intentions (laws of psychology) – (a vector consisting of these elements)

Variable y: Observable behavior (laws of physics) – Perception of physical reality

A markov random field describes the conditional dependencies (influences) between different random variables indexed on an undirected graph with Markov property.

Here, two fields define an agent attached with variables. Butterfield et al. 2009 further introduce a *local evidence* relation and a *compatibility* relation between two agents. The local evidence relation gives a measure how well perceptual reality of agent 1 and his attribution of a mental state to agent 2 fit; the compatibility relation results in a measure of influence of the mental states of agent 1 on agent 2. For each pair of agents I and j a compatibility relation exists, and a local evidence relation for each pair of agent i’s perception xi and state yi.

The joint probability distribution Pr of the given random variables is given by



and describes the likelihood that x occurs given the relevant agents’ states and perceptions. x can be defined as the content of an agent’s mental state, so it can be modeled how likely a belief occurs given an observational space.

Butterfield et al. apply this approach to several applications of uncertainty and reliability elegantly and interestingly, but surprisingly not to a traditional false belief test. What are the reasons? First, it is worth to note their following assumptions and constraints:

1. Perceptual reality includes social cues.

* This is a notion of preference for attention.

1. Perceived state of another agent is equal to its actual state, there is no deception.

* This is a notion of inference. Reasoning over actual perceived states is equal to reasoning over potential latent states, or alternatively: the mental world is transparent.

How ToM-perception can be modeled or constraint, if the mental world is transparent? This does not mean that information of physical and mental world would clash; both are modeled by different variables.

1. One can argue for a ToM with direct perception. Then the true mental states can be “seen” literally.
2. Further, one can add scripted preferences for relations between agents. These preferences can be arguably defined in “mental” terms; however, here preference is just a constraint for a perceptual channel: the mental is perceived in a way analog to the physical data, because there is only what is perceived. (This should not be understood to be antirealism that there could not be theoretical terms; nevertheless the robots do not handle with theoretical terms even in the case of the mental – they handle with special perceptual content)
3. Furthermore, in their application to belief propagation in multi-robot-coordination “*advice-messages*” are introduced, which are continuously submitted between robots. Given that agent 1 has to decide what action he has to do next, he can care for suggestions (advice-messages) by other agents what to do. This is a simple mechanism for planned joint action without a central planner. Advice-messages of agent 1 are selectively computed given local evidence, compatibility between agent 1 (sender) and agent 2 (receiver), and the advice-messages of robots located in the neighbourhood (without messages from agent 2). Butterfield et.al describe using these messages as “using others’ intentions” and limit their intentions to one content perceivable by others:
4. An agent’s mental state only contains the action to be done next.

* This is a notion of goal.

There is some confusion regarding the label “Theory of Mind” in view of the serious constraints. Indeed, in ch.4 Butterfield et al. 2009 mention:

*The compatibility functions and even the methods for obtaining the state would be very different for robotic and human collaborators. The simplest application of the MRF model to multi-robot coordination is where the state is the action to be taken and all the agents are robots.*

We suggest that Butterfield et al. should use the label “Minimal Theory of Mind” to describe their approach conceptually, because it is not *full blown adult human* Theory of Mind. There are no unobservable mental states, because mental states are perceived and are truly the way they are perceived. Thus, here are only relations of perceptions and goal-representations. The compatibility relation (mental, mental) is just a relation

(agent1, agent2) := (agent1, perception\_mental\_agent2) \* (agent2, perception\_mental\_agent1)

And because the mental contents here are goals, it is the like in Minimal ToM; not a theory-theory, simulator or something else one would attribute to human ToM.

Here we seem to have the case that preference-inference-goal agents track each other: robots track robots – but not the case that humans track humans, or robots track humans, or humans track robots. These constellations would involve other relations, as Butterfield et al. 2009 argue. Interestingly, one can argue to have a first proof of principle for agents of the same class tracking each other.

**Preference**: Preferences shape processing and attention in perception (you have a bias). Preferences in Butterfield et al. 2009 are prescripted (biased) compatibility relations, which can be thought to have a bias in perception of the mental.

**Inference**: Though the label “intention” is involved, it is not necessary to speak about relations of intentions in the demonstrated applications, if the actual mental state *is* or *is perfectly mapped onto* the perceived state. In first case, mental states would be perceived; in second case there exists a perfect behavioral proxy for intentions. Further, the extension of gaze, reliability and uncertainty could be in limits the same in ToM and Minimal ToM (c.f. Butterfill and Apperly 2009). Inferences here are only

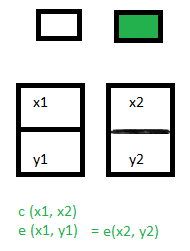
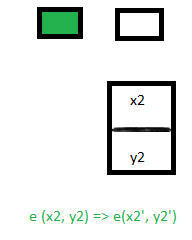
**Goal**: Their notion of goal in this robot-world is compatible with our definition of goal as a property of behavior (see above). Furthermore, robot-relations do not necessarily contain human-relations (see quotation above). Goals as properties of behaviors can be perceived by agents having a Minimal ToM.

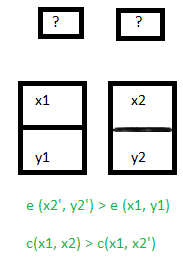
Another coincidence is about illusion: In their robot-world no deception exists. This is also true for Minimal Theory of Mind-agents in case of perceptual illusions and identity (Butterfill and Apperly 2010). But having no illusions or deceptions as well as having only robot-robot-relations does not exclude “false belief” – in a special sense.

**A false belief-test**

Although Butterfield et al. 2009 did not model a false belief-test; this can be done for short within their framework and given limitations. Note that this is a false belief-test in the sense of preference inference goal-agents. It will give hints in view of *tracking* false beliefs.

There may be two agents described by each two variables, Agent1 (x1, y1) and Agent2 (x2, y2), whereas x are physical states and y are Minimal Theory of Mind-states. We assume that their notion of mental state is indeed Minimal Theory of Mind-state and contains the behavioral goal. There may be a relation of local evidence e: e(x1, y1) and e(x2, y2); and a compatibility function c(x1, x2). Then one applies Onishi and Baillargeon’s false belief-task:

1. Agent1, Agent2, and two boxes are in a room. A melon is placed into one box. Both agents have high agreement in terms of high output of compatibility relation and same or very similar local evidence relations.
2. Agent1 goes away and the melon is placed into the other box. Agent2 can perceive this event and consequently his perceptual state notes a change in the registration of the melon. This means higher fit between actual state and perceptual reality in Agent2 after registering the change: e(x2, y2) < e(x2’, y2)

1. Agent1 returns, but now the situation completely changed. Note that agent2 neither does deceive nor agent1 has illusions – he correctly registered in scene1. The following will depend on advice-messages:

If they are allowed, agent1 and agent2 will handle the situation, given a suitable preference (e.g. “Higher priority for another agent’s state if it was already there.”). Then there is no test-situation.

If they are not allowed, one can model agent2’s registrations.

Let us “test” agent2 in prediction and explanation of agent1’s behavior.

**Prediction of behavior (success)**: If agent2 has Minimal ToM, it will predict that agent1 will look after the melon in the false place: correct registration is a success factor. With the help of episodic memory and an allocentric map this could be realized:

<agent1, agent2, melon> => <-agent1, agent2, melon> => <agent1> <agent2, melon>

The third allocentric map cannot contain the same relations as the first one. Whatever agent1 will do, it cannot depend on a successful registration. So it is more likely that agent1 will not do the right thing.

**Explanation of behavior (cause)**: With Minimal ToM, agent2 may be able to represent agent1’s field. Because (correct) registration is a causal factor it is behaviorally relevant. In agent1’s field the false registration was a causal factor for behavior.

The probability for choosing the false place by agent1 in the perspective of agent 1 would be:

Pr1 (x1 = false box) = 1 - 1/Z \* e(x1 = box,y1) \* c(x1,y1)

The probability for choosing the false place by agent1 in the perspective of agent 2 would be:

Pr2 (x1 = false box) = 1/Z \* (Pr1 \* e(x2‘=box, y2‘)) \* (c(x1, y1) + c(x2‘, y2‘))

Both probabilities include different local evidences and compatibility relations; in fact, they diverge (Pr1 tends towards 0, Pr2 tends towards 1). Because there are no illusions for both agents and Pr1 is not equal to Pr2, agent1’s behavior is “explainable” in terms of Minimal ToM.

Without question it would be desirable to extend such a technical case study to integrate intra-class-tracking as well as inter-class tracking. A biologically informed case study may also elucidate the scrub-jay experiments illustrated above.

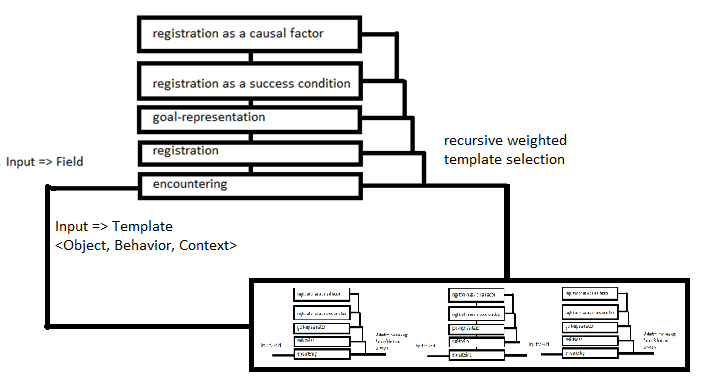
1. **A sketch of a scrub-jay –like companion**

How could a scrub-jay-like companion look like and what capabilities would it have?

1. Control architecture

A possible control architecture for behavior-selection may use the subsumption-architecture outlined above integrated in a parallel-like architecture.

Taken the subsumption-architecture for itself, it can be argued that such a Minimal ToM-module is just a special case of recursive behavior-selection on previous (when) goal-directed behavior (what) with objects in a context (where). Behavior is the interaction between an instantiated template of a reactive module and its ecology including environmental fields as well as its previous template.



A subsumption-template gets input from sensor-data and previous templates: objects, behavior and context.

One will need further modules, for example to model compulsive behavior, and further architectural elements, for example to model ontogeny and (higher) learning. Some details have to be declared, for example tracking relations, and the type of plans.

For example, should plans be explicit (declarative) or implicit by the relations of behaviors to each other? Which other modules should be added in accord with biology?

1. Tracking and faking
2. Simulation

**5. Discussion**

“Kim Sterelny (personal communication) sees the core hypothesis of cognitive ethology to be:

“animals represent their world and exploit those representations in modulating their particular

behaviors to their particular problems. So we can neither predict nor explain animal behavior without an understanding of *what about their world they represent, and how they represent it*”.” (Bekoff, Marc. Cognitive ethology, vigilance, information gathering, representation: Who might know what and why? Behavioral Processes 35 (1996) 225-237; emphasis added)

Some years ago, Kim Sterelny stated the central tenet of what animals represent of the world and how they do it. Full blown ToM sets the What as being intrinsically human: relations in intentional systems or relations between ToM-possessing agents. Minimal Theory of Mind states that in limited situations it would be enough to represent relations between agents and objects. Thus, conceptually both theories operate on different types of relations.