UNIVERSITEIT
AMSTERDAM

VU Research Portal

Formal Analysis of Dynamics Within Philosophy of Mind by Computer Simulation

Bosse, T.; Schut, M.C.; Treur, J.

Minds and Machines
2009

DOI (link to publisher)
10.1007/s11023-009-9177-z

document version
Publisher's PDF, also known as Version of record

Link to publication in VU Research Portal

citation for published version (APA)
Bosse, T., Schut, M. C., & Treur, J. (2009). Formal Analysis of Dynamics Within Philosophy of Mind by
Computer Simulation. *Minds and Machines*, *19*(4), 543-555. https://doi.org/10.1007/s11023-009-9177-z

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners
and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

  • Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
  • You may not further distribute the material or use it for any profit-making activity or commercial gain
  • You may freely distribute the URL identifying the publication in the public portal ?

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately
and investigate your claim.

E-mail address:
vuresearchportal.ub@vu.nl

Download date: 23. May. 2021

Formal Analysis of Dynamics within Philosophy of Mind by Computer Simulation

Tibor Bosse · Martijn C. Schut · Jan Treur

Abstract Computer simulations can be useful tools to support philosophers in validating their theories, especially when these theories concern phenomena showing nontrivial dynamics. Such theories are usually informal, whilst for computer simulation a formally described model is needed. In this paper, a methodology is proposed to gradually formalise philosophical theories in terms of logically formalised dynamic properties. One outcome of this process is an executable logic-based temporal specification, which within a dedicated software environment can be used as a simulation model to perform simulations. This specification provides a logical formalisation at the lowest aggregation level of the basic mechanisms underlying a process. In addition, dynamic properties at a higher aggregation level that may emerge from the mechanisms specified by the lower level properties, can be specified. Software tools are available to support specification, and to automatically check such higher level properties against the lower level properties and against generated simulation traces. As an illustration, three case studies are discussed showing successful applications of the approach to formalise and analyse, among others, Clark's theory on extended mind, Damasio's theory on core consciousness, and Dennett's perspective on intertemporal decision making and altruism.

Keywords Computer simulation · Dynamics · Philosophy of mind

T. Bosse (✉) · M. C. Schut · J. Treur
Department of Artificial Intelligence, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
e-mail: tbosse@cs.vu.nl
URL: http://www.few.vu.nl/∼tbosse

M. C. Schut
e-mail: schut@cs.vu.nl
URL: http://www.few.vu.nl/∼schut

J. Treur
e-mail: treur@cs.vu.nl
URL: http://www.few.vu.nl/∼treur

Introduction

This paper introduces ideas on a research methodology that aims to bridge the gap between philosophical thought experiments and computer simulation. Computer simulations can be used as *intuition pumps*, to scale up and analyse such non-empirical experiments, as described by Dennett (2003):

> Computer simulations … add further discipline: a way of discovering hidden assumptions of one's models, and a way of exploring the dynamic effects, by "turning the knobs" to see the effect of different settings of the variables. It is important to recognize that these computer simulations are actually philo-sophical thought experiments, intuition pumps, not empirical experiments. … Philosophers used to have to conduct their thought experiments by hand, one at a time. Now they can conduct thousands of variations in an hour … (Dennett 2003, Chap. 7, p. 218).

Even more so, the computer can be considered a tool to support philosophers in their thinking process on particular consequences of considered models:

> … the evolutionary perspective … permits us to explore the interactions over time between agents that philosophers typically just handwave about. For instance, philosophers often ask "What if everybody did it?" as a rhetorical question, and don't stop to consider the answer, which they typically think is obvious. They never even address the more interesting question: What if *some* people did it? (What percentage, over what time period, under what conditions?) (Dennett 2003, Chap. 7, pp. 217–218).

However, setting up a simulation model often requires quite some technical work on programming in some programming language, which makes it a not very attractive activity for the average philosopher. Sometimes cooperation with a computer scientist or AI researcher who is interested in philosophical themes may help, but this type of researcher is also a bit rare.

The approach proposed here aims at improving this situation by offering computer-supported methods to obtain specifications that can be used within a computer at a conceptual modelling level, based on a gradual formalisation process of dynamic properties from a temporal linguistic and logical perspective. On the one hand this formalisation process provides a conceptual level specification of a simulation model describing the basic mechanisms underlying a process at the lowest aggregation level. Within a dedicated software environment that has been developed, this specification can be used to obtain a simulation model to perform simulations. On the other hand, conceptual level specifications of dynamic properties at a higher aggregation level of the process that is considered (for example, properties that are expected to emerge from the basic mechanisms) can be expressed, and formalised. Using available software tools, automated verification can be performed of such higher level properties against (1) the lower level properties of the simulation model, (2) generated simulation traces, or (3) empirically available traces.

The methodology has been successfully applied in case studies addressing major themes within Philosophy of Mind, in particular themes that involve phenomena

with nontrivial dynamics. Philosophical themes that have been clarified in this manner concern dynamically emerging properties (such as representation relations; e.g., Jacob 1997; Kim 1996) for an overall process based on given or assumed mechanisms. In this paper, three case studies are described in which such emergent dynamic properties have been analysed. These case studies address the following themes: the (shared) extended mind (Clark 1997, 2001; Clark and Chalmers 1998) within ant colonies, the notion of core consciousness (Damasio 2000), and the idea of altruistic behaviour based on intertemporal decision making (Dennett 2003).

In section "Conceptual Analysis from Informal to Formal" the conceptual analysis method involving specification from informal to formal format is discussed. Section "Extended Mind (Clark)" presents a first case study as an application of this method, on the use of extended mind (Clark 1997, 2001; Clark and Chalmers, 1998) within ant colonies. In section "Core Consciousness (Damasio)" it is shown how the method was used to analyse Damasio (2000)'s theory on core consciousness. Section "Intertemporal Decision Making and Altruism (Dennett)" addresses Dennett (2003)'s perspective on how cognitive capabilities for intertemporal decision making play a role in the evolution of altruistic behaviour. Final section is a "Discussion".

Conceptual Analysis from Informal to Formal

Within our approach a *dynamic property* is considered to be a building block from which we can construct complex dynamic structures (cognitive agents, organisations, societies, complex systems). Typically, a dynamic property identifies a relation between something that happens at some time and something that happens at, possibly but not necessarily, another time. We distinguish a number of different formats for expressing dynamic properties. Depending on the contexts, these formats can be based on *informal natural language*, *semi-formal structured natural language* or a *formal language*.

An example of a simple dynamic property (in informal format) is the following: "If an agent observes that it is raining, then later on it will believe that it is raining". For less formal discussions, for example, with domain experts, informal (and semi-formal) formats are more appropriate than a formal format. On the other hand, if automated checking software is used, dynamic properties have to be in a formal format. A natural process during the analysis of philosophical questions is that first informal specifications (of dynamic properties) are expressed, and later these informal expressions are translated into semi-formal, and possibly into formal formats. A first step in such an analysis involves **acquisition of a (domain) ontology**. In this process, different **state properties** are identified and distinguished from each other: concepts that relate to an agent's **input state**, **output state**, **internal state**, and to **external world states**. An example of an internal state property is belief (agent,_A itsraining). The ontology later facilitates the formalisation of dynamic properties, as the different concepts are already defined. Moreover, a formalisation of a scenario (for example empirical or imagined data over time) can be made by using the formal ontologies for the different states, in order to formalise a sequence of events as a temporal **trace**.

Usually, also a **temporal structure** has to be reflected in the representation of a dynamic property. This entails that terms such as 'at any point in time', 'at an earlier point in time', 'for all time points between t1 and t2', 'after', 'before' are used to clarify the temporal relationships between different fragments in the dynamic property. This temporal structure is a main aspect distinguishing informal properties from semi-formal properties. For example, the semi-formal variant of the informal property shown above is the following: "If at any point in time t1 and agent A observes that it is raining, then there exists a time point t2 after t1 such that at t2 the agent A believes that it is raining".

To obtain formal representations of dynamic properties, the input, output, internal and external ontologies are chosen as formal ontologies for states, specified by sorts, constants, functions and predicates within an order-sorted predicate logic language. In addition, the temporal structure, as present in a semi-formal representation, is expressed in a predicate logic format, using an ordering relation between time points, the usual logical connectives $(\wedge, \vee, \neg, \Rightarrow)$ and universal and existential quantifiers over time $(\forall, \exists)$. In our methodology, we use the temporal language TTL (Bosse et al. 2009) for this purpose. In TTL, the property shown above is formalised as follows:

$$\forall t1 \quad [\ state(\gamma, t1) \models observes(agent\_A, itsraining) \ \Rightarrow$$
$$\exists \ t2 \geq t1 \quad state(\gamma, t2) \models belief(agent\_A, itsraining) \ ]$$

Here, γ is a variable that stands for an arbitrary trace, and t1 and t2 stand for time points. Moreover, `state(`γ`, t)` $|$ `= p` denotes that state property p is true in the state of trace γ at time point t.

Using the formal ontologies, and the formalisation of the temporal structure, a formalisation is obtained of dynamic properties. This formalisation can be used to perform automated analysis and simulation within a software environment that has been developed for these purposes.

Dynamic properties can be specified for different **aggregation levels**, from the lowest level of the direct causal relationships between state properties within a process (modelling the basic mechanisms assumed) to higher aggregation levels for properties of the process as a whole (modelling properties that emerge from the basic mechanisms). Within analysis, the different aggregation levels provide automated verification possibilities to check whether the higher level properties are consistent with, or even are entailed by the lower level properties. The properties at the lowest aggregation level are often specified in **executable** format, close to the format of a transition system or a finite automaton. This format is suitable as a basis for a simulation model, to obtain simulated traces of the process. In our methodology, the executable language LEADSTO (Bosse et al. 2007), which is a sub-language of TTL, is used to specify such executable properties. The basic building blocks of this language are expression of the format $\alpha \ \beta$ (pronounced α *leads to* β), which informally means the following: if state property α holds for a certain time interval, then after some delay, state property β will hold for a certain time interval. For a precise definition, see (Bosse et al. 2007).

Extended Mind (Clark)

As a first case study, using the approach described above, it is discussed how Clark's theory on extended mind was analysed and formalised (Bosse et al. 2005, 2006). This theory expresses that behaviour is often not only supported by internal mental structures and cognitive processes, but also by processes based on patterns created in the external environment that serve as external mental structures; cf. (Clark 1997, 2001; Clark and Chalmers 1998; Dennett 1996). In particular, in the context of an ant society, where pheromone levels in the environment play a role as external mental state properties, the focus mainly was on (1) logical specification of a simulation model for the lower level mechanisms, and (2) at a higher aggregation level on the representational content (e.g., Bickhard 1993; Jacob 1997; Kim 1996) of external mental state properties (i.e., the pheromone levels in the environment). The latter properties describe representational relations in a formalised form, for which it is to be verified whether they emerge in the process shown in simulated traces. Notice that in the case of an ant colony, the external pheromone states are used in a collective manner, they are shared by multiple agents.

Through modelling the following challenging issues on cognitive modelling and representational content were encountered in this case study: (1) how to define representational content for an external mental state property; (2) how to handle decay of a mental state property; (3) how can joint creation (by multiple agents) of a shared mental state property be modelled; (4) what is an appropriate notion of collective representational content of a shared external mental state property; and (5) how can representational content be defined in a case where a behavioural choice depends on a number of mental state properties.

To model the ant society, the following ontology was used:

	body positions in world:
pheromone level at edge e is i	pheromones_at(e, i)
ant a is at location l coming from e	is_at_location_from(a, l, e)
ant a is at edge e to l2 coming from location l1	is_at_edge_from_to(a, e, l1, l2)
	world state properties:
edge e connects location l1 and l2	connected_to_via(l1, l2, e)
location 1 is the nest location	nest_location(l)
location 1 is the food location	food_location(l)
	input state properties:
ant a observes that it is at location l coming from edge e	observes(a, is_at_location_from(l, e))
ant a observes that it is at edge e to l2 coming from location l1	observes(a, is_at_edge_from_to(e, l1, l2))
ant a observes that edge e has pheromone level i	observes(a, pheromones_at(e, i))
	output state properties:
ant a initiates the action to go to edge e to l2 coming from location l1	to_be_performed(a, go_to_edge_from_to(e, l1, l2))
ant a initiates the action to go to location l coming from edge e	to_be_performed(a, go_to_location_from(l, e))
ant a initiates the action to drop pheromones at edge e coming from location l	to_be_performed(a, drop_pheromones_at_edge_from(e, l))
ant a initiates the action to pick up food	to_be_performed(a, pick_up_food)
ant a initiates the action to drop food	to_be_performed(a, drop_food)

An example of a semiformal representation and a formalisation of a dynamic property in the executable LEADSTO format (Bosse et al. 2007) is the following (note that LP stands for 'Local Property', to be able to distinguish between Local (or executable, at lower aggregation level) and Global Properties (GPs, at a higher aggregation level)).

LP5b (Selection of Edge)

If an ant observes that it is at location A, and edge e1 connected to location A has the highest number of pheromones, compared to edge e2 connected to location A, then the ant goes to edge e1.

> Formal representation:
>   observes(a, is_at_location_from(A, e0)) and connected_to_via(A, l1, e1) and
>   observes(a, pheromones_at(e1, i1)) and connected_to_via(A, l2, e2) and
>   observes(a, pheromones_at(e2, i2)) and i1 > i2
>   ⟶≫ to_be_performed(a, go_to_edge_from_to(e1, A, l1))

This is one of the executable dynamic properties that make up the logical specification of the simulation model that was used to perform simulations. For the complete specification of this simulation model, see Bosse et al. (2005).

The executable dynamic properties discussed above address the process at the lowest aggregation level (the local dynamic properties). The remainder of this section discusses dynamic properties of a higher aggregation level (in the TTL format by Bosse et al. 2006) and their verification against lower level properties. Within these properties, γ is a variable that stands for an arbitrary trace. First a language abstraction is given:

> food_delivered_by(γ, t, a) ≡
>   ∃ l, e [state(γ,t) |= is_at_location_from(a, l, e)) &
>   state(γ,t) |= nest_location(l) & state(γ,t) |= to_be_performed(a, drop_food)]

One of the properties considered at the highest aggregation level is:

GP1 Food Delivery Succesfulness

There is at least one ant that brings food back to the nest.

> ∃ t ∃ a: food_delivered_by(γ, t, a).

Another type of dynamic property at a higher aggregation level is a representation relation (e.g., Bickhard 1993; Jacob 1997, Kim 1996, pp. 184–210) for the pheromone states of the environment. The backward case of a representation relation for the pheromone states in the environment involves a summation over multiple agents at different time points, and decay rate r with $0 < r < 1$ is used to indicate that after each time unit only a fraction r is left; see Bosse et al. (2006).

Backward Representation Relation for Pheromone States

There is an amount v of pheromone at edge e, if and only if there is a history such that at time point 0 there was ph(0, e) pheromone at e, and for each time point k from 0 to t a number dr(k, e) of ants were present at e, and

$$v = ph(0, e) * r^t + \Sigma_{k=0}^{t} dr(t\text{-}k, e) * r^k$$

A formalisation of this property in the logical language TTL is as follows:

$\forall t \ \forall e \ \forall v \ \ state(\gamma, t) \models pheromones\_at(e, v) \Leftrightarrow$

$\Sigma_{k=0}^{t} \Sigma_{a=ant1}^{ants} case([\ \exists l,l1 \ state(\gamma, k) \models is\_at\_edge\_from\_to(a, e, l, l1)], 1, 0) * r^{t\text{-}k} = v$

Here for any formula `f`, the expression `case(f, v1, v2)` indicates the value `v1` if `f` is true, and `v2` otherwise. Likewise, according to the relational specification approach the following forward representation relation was specified.

Forward Representation Relation for Pheromone States

If at time t1 the amount of pheromone at edge e1 (connected to location 1) is maximal with respect to the amount of pheromone at all other edges connected to that location l, except the edge that brought the ant to the location,
    then, if an ant is at that location l at time t1,
    then the next edge the ant will be at some time t2 > t1 is e1.
    If at time t1 an ant is at location 1 and for every ant arriving at that location 1 at time t1, the next edge it will be at some time t2 > t1 is e1, then the amount of pheromone at edge e1 is maximal with respect to the amount of pheromone at all other edges connected to that location l, except the edge that brought the ant to the location.
    A formalisation of this property is as follows.

```
∀t1,l,l1,e1,e2,i1
  [ e1≠e2 &
  state(γ, t1) |= connected_to_via(l, l1, e1) &
  state(γ, t1) |= pheromones_at(e1, i1) &
  [∀l2≠l1, e3≠e2 [ state(γ, t1) |= connected_to_via(l, l2, e3) ⇒
      ∃ i2 [0≤i2<i1 & state(γ, t1) |= pheromones_at(e3, i2) ] ]
  ⇒ ∀a [ state(γ, t1) |= is_at_location_from(a, l, e2) ⇒
      ∃ t2>t1 state(γ, t2) |= is_at_edge_from_to(a, e1, l, l1) &
      [∀t3 t1<t3<t2 ⇒ state(γ, t3) |= is_at_location_from(a, l, e2) ]]]]

∀t1, l,l1,e1,e2
  [e1≠e2 &
  state(γ, t1) |= connected_to_via(l, l1, e1) &
  ∃ a  state(γ, t1) |= is_at_location_from(a, l, e2) &
  ∀a [ state(γ, t1) |= is_at_location_from(a, l, e2)  ⇒
      ∃ t2>t1 state(γ, t2) |= is_at_edge_from_to(a, e1, l, l1) &
      [∀t3 t1<t3<t2 ⇒ state(γ, t3) |= is_at_location_from(a, l, e2) ]]]
  ⇒ ∃ i1 [ state(γ, t1) |= pheromones_at(e1, i1) &
      [∀l2≠l1, e3≠e2 [ state(γ, t1) |= connected_to_via(l, l2, e3)
      ⇒∃i2 [0≤i2≤i1 & state(γ, t1) |= pheromones_at(e3, i2) ]]]]
```

The properties at a higher aggregation level discussed above and a number of other properties have been formalised and using a checking software environment have been (automatically) verified in simulation traces. This is a first manner for verification. A second way of verification is to establish logical relationships between properties (by mathematical proof). This also has been performed in a number of cases, under a number of assumptions. For more details, see (Bosse et al. 2005, 2006). The results of these verifications show that indeed in the process of the

ant colony, for which the mechanisms are modelled at the lower aggregation level of the simulation model, the assumed representation relations for the external pheromone states emerge, which shows that these external pheromone states play the role of (collective) external mental states in the expected manner.

Core Consciousness (Damasio)

As another case study, Damasio's theory on core consciousness was analysed and formalised (Bosse et al. 2008a). According to this theory, a state of core consciousness (or conscious feeling) for a certain object occurs when an agent monitors a change of its representation of its body state (the protoself) after the occurrence of this object. In other words, the state of core consciousness represents the process of change of the agent's body state representations co-occurring with the occurrence of the object, i.e., it represents transitions between the following states: *protoself at the inaugural instant—object comes into sensory representation— protoself as modified by the object* (Damasio 2000, pp. 177–178).

Based on Damasio's theory, first a formal model was provided of the states and basic processes leading to core consciousness. The building blocks of this model are state properties and their functional roles expressed by executable properties. The following ontology of state properties is used (describing a specific case study about an agent that listens to some very special music, and eventually becomes conscious about this music):

music	a beautiful piece of music is played
sensor_state(music)	the agent is perceiving the music
sr(music)	an internal sensory representation for the music is present
p	the agent's body is preparing to respond to the music
S	the agent is in a body state in responding to the music (e.g., by shivers)
sensor_state(S)	the agent is perceiving its bodily response S
sr(S)	an internal sensory representation for S is present
s0	the agent is in (initial) mental state 0
s1	the agent is in mental state 1
s2	the agent is in mental state 2
speak_about(music)	the agent speaks about the music

In addition, the following executable properties were identified to describe the basic mechanisms of the process at the lowest aggregation level considered.

LP0 music $\twoheadrightarrow$ sensor_state(music)

LP1 sensor_state(music) $\twoheadrightarrow$ sr(music)

LP2 sr(music) $\twoheadrightarrow$ p

LP3 p $\twoheadrightarrow$ S

LP4 S $\twoheadrightarrow$ sensor_state(S)

LP5 sensor_state(S) $\twoheadrightarrow$ sr(S)

LP6 not sr(music) and not sr(S) $\twoheadrightarrow$ s0

LP7 sr(music) and not sr(S) and s0 $\twoheadrightarrow$ s1

LP8 sr(music) and sr(S) and s1 $\twoheadrightarrow$ s2

LP9 s2 $\twoheadrightarrow$ speak_about(music)

Based on these executable properties simulations were performed.

The (backward) representation relation for the mental state for core consciousness s2 was specified as follows: 'if no body state S and no music occur, and later music occurs and still no body state S occurs, and later music occurs and S occurs, then still later s2 will occur,' and conversely. Formally:

Backward Representation Relation for Core Consciousness States

∀t1, t2, t3 [t1≤ t2 ≤ t3 &
    state(γ, t1, EW) |= ¬S ∧ ¬music &
    state(γ, t2, EW) |= ¬S ∧ music &
    state(γ, t3, EW) |= S ∧ music ⇒
        ∃ t4 ≥ t3 state(γ, t4, internal) |= s2]

∀t4 [state(γ, t4, internal) |= s2 ⇒
    ∃ t1, t2, t3 t1≤ t2 ≤ t3 ≤ t4 &
    state(γ, t1, EW) |= ¬S ∧ ¬music &
    state(γ, t2, EW) |= ¬S ∧ music &
    state(γ, t3, EW) |= S ∧ music]

This corresponds to the transitions indicated by Damasio (2000): *the proto-self exists at the inaugural instant—an object comes into sensory representation—the proto-self has become modified by the object*. For an alternative formalisation, based on the notion of second-order representation, see (Bosse et al., 2008a).

Similarly, when looking forward, the representational content of a mental state can be described by relating it to future world states. The future representational content of state property s2 can be informally described as follows: 'if s2 occurs, then later the agent will speak about the music', and conversely. In the logical language TTL, the expression is formalized as follows.

Forward Representation Relation for Core Consciousness States

∀t1 [state(γ, t1, internal) |= s2 ⇒∃t2 ≥ t1 state(γ, t2, EW) |= speak_about(music)]
∀t2 [state(γ, t2, EW) |= speak_about(music) ⇒∃t1 ≤ t2 state(γ, t1, internal) |== s2]

The backward and forward representation relations are dynamic properties at a higher aggregation level. Part of the analysis has been to automatically verify (using the SMV environment; cf. McMillan 1993) that the lower level properties LP0 through LP9 together entail the representational content specifications. This confirms part of the claims made by Damasio (2000) in the sense that the suggested mechanisms as described at a lower aggregation level indeed entail the emergence of higher level dynamic properties that represent the process of monitoring how the agent's body state is affected by a given object.

Intertemporal Decision Making and Altruism (Dennett)

The third case study is inspired by (Dennett 2003—Chap 7)'s discussion of altruistic behaviour from an evolutionary perspective; see also Sober and Wilson (1998), Trivers (1971). This case study concerns an analysis of how the occurrence of forms of altruistic behaviour within agent communities depends on cognitive capabilities of agents with respect to intertemporal decision making; see also

(Darwin [1871]).[1] For more details, see (Bosse et al. [2008b]). The set up focuses on a population with x members which have some regular (weekly, monthly) interactions with each other. These interactions have the typical form that one agent provides something (a service) to another agent without immediate return. Examples of such interactions are lending money, or assisting the other agent with removal events. Each individual interacts with a subset of the population, which may not be the same set all the time. Each interaction has some future consequence. For example, I may be lending you money today (a 'giving' part of the interaction at time t), and after some time you will return me the money (a 'receiving' part of interaction at time $t' > t$). Based on the 'receiving' parts, individuals assign some *credit value* to individuals that they have had interactions with at different points in time.

Inter-temporal choice is a decision in which the realisation of outcomes may lie in the imminent or remote future. Recently, inter-temporal choice has caught the attention in the literature on behavioural decision making (Loewenstein and Elster [1992]). Before this, results on the subject were mainly due to the research contributions in related fields, like economics and animal psychology. The standard agent model for decision making over time is a framework called time discounting (Loewenstein and Elster [1992]), which works according to the same principles as interest that one receives on a bank account: I calculate a delayed reward back to its current value based on the interest that I would receive for it.

We use a similar agent model for inter-temporal decision making here, extended to our particular decision situation (involving reciprocity for cooperation) in two main ways. Firstly, the decisions involve an explicit model the agent has of (regularities in) the environment, in this case incorporating the other agents. This results in parameters for *trust* of the agent in other agents. As explained below, the value of this parameter evolves over time as a consequence of monitoring (regularities in) the environment over time, i.e., the experienced (non)cooperations. Secondly, the individual decisions are concerned with choosing between (1) a possible reward in the remote future and (2) having no immediate cost, rather than choosing between an immediate and delayed reward (as investigated traditionally in time discounting). In the model, the discounted value (see also Ainslie [2001]; Dennett [2003]) $f_{discounted}$ of a future reward is calculated by: $f_{discounted} = f * 2^{-(1-\alpha)(t/n * (1 - (tr + 1)/2))}$, where

f : REAL	= future reward,
$\alpha \in [0,1]$	= discount factor,
t : INTEGER	= duration after which the future reward is received,
n : INTEGER	= duration of cooperation, and
$tr \in [-1,1]$	= trust in the agent who asks you to cooperate.

If the discounted future reward evaluates higher than (or equal to) the current (immediate) cost, the agent decides to cooperate. In other words, $if\ f_{discounted} \geq c,$ $then\ cooperate,\ else\ do\ not\ cooperate$, where c : REAL is the immediate cost.

---

[1] 'As the reasoning powers and foresight of the members became improved, each man would soon learn that if he aided his fellow-men, he would commonly receive aid in return' (Darwin [1871], p. 163).

It was tested how agents that use this decision function develop in a multi-agent society. The prediction was that these agents will show altruistic behaviour, will establish a larger social network than agents without such a decision function (i.e., agents that are not able to estimate the future reward, and thus never cooperate), and will eventually get a higher fitness.

Agents adjust their trust values in other agents according to the following principle: if I ask you to cooperate and you accept, then I increase my trust in you; if you decline, then I decrease my trust in you. For modeling such adaptation of trust over time, we use a trust function that was presented in (Jonker and Treur 1999). This function, as applied here, takes the response of an asked agent (accept/decline) to determine how to revise the trust value. Such a response $e \in [-1,1]$ evaluates to 1 if the agent accepts or -1 if the agent declines. A scaling factor $\delta \in [0,1]$ (which is constant throughout the experiments) determines how strongly an agent is committed to its trust values: a higher δ means that an agent puts much weight on its current trust value and lets an (non)cooperation experience not weigh so heavily; and vice versa. In the model, when the outcome of a request to cooperate is known, we calculate the trust value $\mathtt{tr_{new}}$ as follows:

$$\mathtt{tr_{new}} = \delta * \mathtt{tr} + (1 - \delta) * e$$

where

$\mathtt{tr} \in [-1,1]$	current trust value,
$\delta \in [0,1]$	discount factor (constant),
$e \in [-1,1]$	the response of the agent who you asked to cooperate.

Thus, each agent maintains a list of trust values for all other agents in the environment. The model also includes a *cooperation threshold* $ct \in [-1,1]$ such that agent x only requests cooperation with agent y if trust of agent x in agent y is above this threshold.

The mathematical model for trust-based intertemporal decision making described above has been incorporated in a small number of dynamic properties that describe at the lowest aggregation level the basic mechanisms of the (societal) process considered here, thus providing an executable conceptual model for the simulation model. An example of a property that was specified is:

LP1 Trust Adaptation

Trust is adapted on the basis of experiences.

$\forall$x,y:agent $\forall$tr:real $\forall$e:real
has_trust_in(x, y, tr) $\wedge$ has_experience_with(x, y, e) $\twoheadrightarrow$ has_trust_in(x, y, delta $\times$ tr + (1 - delta) $\times$ e)

Here, 'delta' is a constant, e.g. 0.9. State property `has_trust_in(x, y, tr)` represents the fact that agent x has trust in agent y with value tr, and state property has_experience_with(x, y, e) represents the fact that agent x has an experience with agent y with response e.

Based on these properties at the lowest aggregation level a number of simulations have been made. Moreover, a number of dynamic properties at a higher aggregation

level have been identified that are relevant for the domain of trust-based inter-temporal decision making. These properties have been formalised. Three of them are shown below (in an informal format).

FM: Fitness Monotonicity

If x has the cognitive system for decision making, then there exists a time t such that for all t1 and t2 after t, with t1 < t2, the fitness of x at t2 is higher than the fitness of x at t1.

DMAF: Decision Making Agents get Fitter

Eventually, all agents with the cognitive system for decision making, will be more healthy than the agents without this system.

NDMA: Network of Decision Making Agents

All agents with the cognitive system for decision making will always cooperate with each other.

The above properties have been automatically checked against generated simulation traces involving up to 25 agents. They all turned out to hold, which validates the above statements, such as 'decision making agents get fitter', for the simulation traces. These results support the claims about the evolutionary survival value of a cognitive system for intertemporal decision making, as discussed by Dennett (2003, Chap. 7).

Discussion

This paper describes a computer-supported method to transform philosophical thought experiments into computer simulation, thereby 'pumping up' the intuitions of philosophers. The method involves both informal and formal conceptual analysis including specification of dynamic properties from linguistic, informal, through structured, semiformal, to formal, temporal logical formats. Within the developed software environment a dedicated editor is available to support this process from informal to formal specification. These specifications can be made both at the lower aggregation level of the basic mechanisms underlying the considered process and at the higher aggregation levels of dynamic properties expected to emerge from these basic mechanisms. Within the software environment, the former specifications can be used to perform simulation, whereas the latter type of properties can be checked automatically against simulated (or empirical) traces. Moreover, a model checker environment such as SMV (McMillan 1993) can be used to verify whether these higher level properties are entailed by the lower level properties. The method is particularly relevant for those philosophical thought experiments where dynamics play a crucial role.

The method was illustrated by three case studies that were undertaken. For these case studies, dynamic properties at the lowest aggregation level of basic mechanisms were specified, constituting a simulation model, as well as properties at a higher level

of aggregation, that are expected to emerge from the lower level properties. In each of the cases, an analysis was performed based on computer-supported formalisation, simulation and verification. These analyses supported claims made by Clark (1997, 2001); Clark and Chalmers (1998), Damasio (2000), and Dennett (2003), respectively.

References

Ainslie, G. (2001). *Breakdown of Will*. Cambridge, MA: Cambridge University Press.

Bickhard, M. H. (1993). Representational content in humans and machines. *Journal of Experimental and Theoretical Artificial Intelligence, 5*, 285–333.

Bosse, T., Jonker, C. M., Schut, M. C., & Treur, J. (2005). Simulation and analysis of shared extended mind. *Simulation Journal: Transactions of the Society for Modeling and Simulation International, 81*, 719–732.

Bosse, T., Jonker, C. M., Schut, M. C., & Treur, J. (2006). Collective representational content for shared extended mind. *Cognitive Systems Research Journal, 7*, 151–174.

Bosse, T., Jonker, C. M., & Treur, J. (2008a) Formalisation of Damosio's Theory of emotion, feeling and core consciousness. *Consciousness and Cognition, 17*, 94–113.

Bosse, T., Jonker, C. M., van der Meij, L., Sharpanskykh, A., & Treur, J. (2009). Specification and verification of dynamics in agent models. *International Journal of Cooperative Information Systems, 18*, 167–193.

Bosse, T., Jonker, C. M., van der Meij, L., & Treur, J. (2007). A language and environment for analysis of dynamics by simulation. *International Journal of Artificial Intelligence Tools, 16*, 435–464.

Bosse, T., Schut, M. C., Treur, J., & Wendt, D. (2008b). Trust-based inter-temporal decision making: Emergence of altruism in a simulated society. In: Antunos, L., Paolucci, M., Norling, E. (Eds.) *Proceedings of the 8th International Conference on workshop on multi-agent-based simulation, MABS'07*. Lecture notes in AI, Vol. 5003, 96–111.

Clark, A. (1997). *Being there: Putting brain, body and world together again*. Cambridge, MA: MIT Press.

Clark, A. (2001). Reasons, robots and the extended mind. *Mind & Language, 16*, 121–145.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*, 7–19.

Damasio, A. (2000). *The feeling of what happens: Body*. Cambridge, MA: Emotion and the Making of Consciousness. MIT Press.

Darwin, C. (1871). *The descent of man*. London: John Murray.

Dennett, D. C. (1996). *Kinds of mind: Towards an understanding of consciousness*. New York: Basic Books.

Dennett, D. C. (2003). *Freedom evolves*. New York: Viking Penguin.

Jacob, P. (1997). *What minds can do: Intentionality in a non-intentional world*. Cambridge: Cambridge University Press.

Jonker, C. M., & Treur, J. (1999). Formal Analysis of Models for the Dynamics of Trust based on Experiences. In: F.J. Garijo, M. Boman (eds.), *Multi-agent system engineering, proceedings of the 9th european workshop on modelling autonomous agents in a multi-agent world, MAAMAW'99*. Lecture Notes in AI, vol. 1647, Springer, Berlin, 1999, (pp. 221–232). Extended version as Technical Report.

Kim, J. (1996). *Philosophy of mind*. Boulder: Westview Press.

Loewenstein, G. F., & Elster, J. (1992). *Choice over time*. New York: Russel Sage Foundation.

McMillan, K. L. (1993). Symbolic Model Checking: An approach to the state explosion problem. PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, 1992. Published by Kluwer Academic Publishers, 1993.

Sober, E., & Wilson, D. S. (1998). *Unto others: The evolution and psychology of unselfish behaviour*. Cambridge, MA: Harvard University Press.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology, 46*, 35–57.