Patterns in World Dynamics Indicating Agency

Tibor Bosse and Jan Treur

Vrije Universiteit Amsterdam, Department of Artificial Intelligence
De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands
{tbosse, treur}@cs.vu.nl, http://www.cs.vu.nl/~{tbosse, treur}

Abstract

In this paper, the question is addressed which patterns in world dynamics are an indication for a conceptualisation of a world's process as an agent. Six criteria are discussed that provide an indication for the world to show a form agency, and allows for suitable agent-based conceptualisation. The criteria take the form of relationships between the occurrence of certain patterns in the world's dynamics, and are expressed as second-order properties of world dynamics. They are formalised in a reified temporal predicate (meta-)logical language and their use is illustrated in a case study, supported by automated support in the form of simulation and verification.

1 Introduction

To conceptualise processes in the world, often an agent-oriented perspective is a useful conceptual tool. By having distinguished a number of agents and their interaction, the overall process can be analysed from a collective intelligence perspective, as emerging from the individual agent processes and their interactions. However, a fundamental question, usually solved implicitly when agent-based modelling is applied, is which parts of the world's process can reasonably interpreted as agents. Whether or not to choose for an agent-based conceptualisation might be considered just a modelling choice, which is to a certain extent a subjective issue for the modeller. However, not just any process can just be considered an agent in a reasonable manner. The patterns shown by the dynamics of the world should not contradict the possibility of an agent-based conceptualisation. At least certain aspects of agency should show themselves in the world's dynamics. In other words, there are certain criteria for the dynamics of the world that indicate a form of agency. This paper addresses the question which properties patterns occuring in the world's dynamics are indications for agency, and enable a modeller to choose for an agent-based conceptualisation in a justified manner.

Dissatisfaction with agents that are modelled in a way isolated from the physical world, not taking into account adequate criteria for agency, has led to recent attention for the question how to embody agents, and how to embed them in the physical world. The perspective taken in this paper, in a sense, starts at the other end: the world's dynamics and patterns that occur in these dynamics. Using such a perspective, an agent emerges from the world's processes, and thus is fully integrated in them in a natural manner.

In this paper, in Section 2 six agency-indicating criteria are identified and discussed informally: boundary separating internal and external, isolation, modular world dynamics, input-output dynamics relations, internal-interaction dynamics relations, and representation relations. Next, in Section 3 the formal language MetaTTL is introduced. In subsequent Sections 4 to 9 for each of the criteria it is shown how, using this language, it can be formalised as a second-order dynamic properties of the world. After that, in Section 10, a simple case study illustrates the use of the criteria. This case study has been addressed using automated support in the form of simulation and verification. Finally, as Section 11 a discussion is included.

2 Agency Criteria for Patterns in World Dynamics

In this section, both ontological assumptions on the world state ontology and assumptions on the dynamics of the world are explored as indications for agency. Note that these indications are not assumed to be non-overlapping, nor independent. Moreover, different notions of agency can be covered by taking different subsets of them. For example, a world showing a purely reactive deterministic agent with behaviour fully determined by the input states will fulfill a subset of properties different from the subset fulfilled by a world showing an agent with goal-directed behaviour with some degrees of freedom or randomness in its behaviour.

2.1 Boundary Separating Internal and External

A first criterion for agency concerns the often-mentioned issue that there is a *boundary* separating *internal* states and processes for the agent (internal milieu, body) from states and processes *external* to the agent; cf. Bernard (1865), Brewer (1992), Cannon (1932), Damasio (2000), pp. 133-145, Dobbyn and Stuart (2003). The idea is that this boundary can be crossed only by specific processes: from outside to inside by sensor processes (via agent *input states* at the boundary), and from inside to outside by actuator processes (via agent *output states* at the boundary). The rest of the boundary is not affectable (for example, the shell of a sea animal). Abstracting from more precise spatial relations, this is covered here by the assumption that the world state ontology is the union of a collection of sets for areas: internal, external, boundary, input and output. Note that what is external for a given agent, includes the other agents. What is indicated as 'external world' includes both the physical and social environment of the agent.

2.2 Isolation

In addition to the boundary criterion, the fact that the boundary can only be crossed by specific processes via input and output states is formalised by a criterion on patterns in world dynamics called *isolation*. This criterion expresses that (causal) influences between internal and external state properties or processes can only occur in an indirect manner via the input states and output states. As an example, the internal processes for a biological organism are protected against uncontrolled external influences by skin, or bone (protecting the brain), or shell. As another example, a company organised by a 'front office – back office' structure, protects the work going on in the back office against uncontrolled external influences. The front office serves as an interface to the external world, transferring requests for products (input) from external to internal and offers for products (output) from internal to external.

2.3 Modular World Dynamics

Another criterion for agency is that the world's dynamics is composed from dynamics based on two separate but interacting processes, i.e., a purely internal and a purely external process; e.g., Aleksander (1996), Dobbyn and Stuart (2003). Thus, this criterion describes a form of *modularisation of world dynamics*. For a biological organism, the modularisation shows how the internal processes (such as mental processes and digestion) are separated from the external processes. For the company example, the internal back office process is separated from the external processes.

2.4 Input-Output Dynamics Relations

A further criterion is that (by the internal process) in one way or the other the dynamics of the output states relates to the dynamics of the input states. For example, by Kim (1996, pp. 85-91) such a relation is called an *input-output correlation*. For the company example, the output provided by the front office to the external world depends on the input that was received: for example, if a certain type of product was requested, the offer will involve this type of product.

2.5 Internal and Interaction Dynamics Relations

Relations between the dynamics of input states and of output states, (interaction dynamics, for short), depend on the agent-internal processes. By Kim (1996, p. 87) this is expressed as: a formalisation M of internal dynamics (by a Turing machine in his case) 'is a behavioural description of a system S just in case M provides a correct description of S's input-output correlations'. This shows how the system's behaviour as shown by its input and output states depends on its internal mechanisms: *relations between internal dynamics and interaction* dynamics.

2.6 Representation Relations

Finally, *representational content* is a notion that is often related to internal agent states, in particular if a sense of self is at issue; e.g. Kim (1996), Damasio (2000), Dobbyn and Stuart (2003), Stuart (2002), Jacob (1997), Keijzer (2002), Sun (2000). The *relational specification* approach of representational content as introduced by Kim (1996), pp. 200-202, and worked out by Jonker and Treur (2003) and Bosse, Jonker and Treur (2009), is adopted for this criterion. Kim views this as a way to account for a broad or wide content of mental properties:

> ‘The third possibility is to consider beliefs to be wholly internal to the subjects who have them but consider their contents as giving *relational specifications* of the beliefs. On this view, beliefs may be neural states or other types of physical states of organisms and systems to which they are attributed. Contents, then, are viewed as ways of specifying these inner states; wide contents, then, are specifications in terms of, or under the constraints of, factors and conditions external to the subject, both physical and social, both current and historical. (…) These properties are intrinsic, but their specifications or representations are extrinsic and relational, involving relationships to other things and properties in the world. It may well be that the availability of such extrinsic representations are essential to the utility of these properties in the formulation of scientific laws and explanations. (…) … in attributing to persons beliefs with wide content, we use propositions, or content sentences, to represent them, and these propositions (often) involve relations to things outside the persons. When we say that Jones believes that water is wet, we are using the content sentence “Water is wet” to specify this belief, and the appropriateness of this sentence as a specification of the belief depends on Jones’ relationship, past and present, to her environment. (…) The approach we have just sketched has much to recommend itself over the other two. It locates beliefs and other intentional states squarely within the subjects; they are internal states of the persons holding them, not something that somehow extrudes from them. This is a more elegant metaphysical picture than its alternatives. What is “wide” about these states is their specifications or descriptions, not the states themselves.’ (Kim, 1996), pp. 200-202; italics in the original.

According to this approach, an internal state property has representational content in the sense that a *representation relation* exists that relates the occurrence of this state property to occurrences of certain patterns in the external part of the world. Such patterns may occur in the past and the future. Similarly, the internal state property may be related to interaction states (for *interactivist* representation; cf. Bickhard, 1993), or to other internal states (*second-order* representation; e.g., Damasio, 2000, pp. 168-182). For the company example, for example, a choice made within the back office relates to the (past) input from a certain customer and also to the (future) output to be provided to this customer.

3 Formalising Patterns in World Dynamics and their Relationships

To formalise the patterns in world dynamics that play a role in the above criteria, as a basis the *Temporal Trace Language* (TTL) to express dynamic properties is used; cf. (Jonker and Treur, 2002; Bosse, Jonker, Meij, Sharpanskykh and Treur, 2009; Sharpanskykh and Treur, 2010). This language can be classified as a sorted reified temporal predicate logic language (see, e.g., Galton, 2003, 2006), in contrast to, for example, modal-logic-based temporal logics as the ones discussed in, e.g., (Fisher, 2005). The language is briefly introduced here. For more details, including its semantics, see (Bosse, Jonker, Meij, Sharpanskykh and Treur, 2009; Sharpanskykh and Treur, 2010) [1]. Moreover, to express properties (second-order patterns), that have the form of relationships between the occurrence of patterns, by second-order dynamic properties, the language MetaTTL is introduced.

[1] Note that in the current paper a slightly different notation is used.

3.1 States and Traces

In TTL, ontologies for world states are formalised as sets of symbols in sorted predicate logic. For any ontology Ont, the ground atoms form the set of *basic state properties* BSTATPROP(Ont). Basic state properties can be defined by nullary predicates (or proposition symbols) such as hungry, or by using n-ary predicates (with n>0) like has_temperature(environment, 7). The *state properties* based on a certain ontology Ont are formalised by the propositions (using conjunction, negation, disjunction, implication) made from the basic state properties; they constitute the set STATPROP(Ont).

In order to express dynamics in TTL, in addition to state properties, important concepts are *states*, *time points*, and *traces*. A *state* S is an indication of which basic state properties are true and which are false, i.e., a mapping S: BSTATPROP(Ont) → {true, false}. The set of all possible states for ontology Ont is denoted by STATES(Ont). Moreover, a fixed *time frame* T is assumed which is linearly ordered. Then, a *trace* γ over a state ontology Ont and time frame T is a mapping γ : T → STATES(Ont), i.e., a sequence of states γ_t ($t \in T$) in STATES(Ont). The set of all traces over ontology Ont is denoted by TRACES(Ont), i.e., TRACES(Ont) = STATES(Ont)^T. Finally, a *temporal domain description* W is a given set of traces over the state ontology (usually in a given application domain), i.e., W ⊆ TRACES(Ont). This set W represents the world that is considered.

3.2 Patterns in World Dynamics as Dynamic Properties

Patterns in world dynamics are described by dynamic properties. The set of *dynamic properties* DYNPROP(Ont) is the set of temporal statements that can be formulated with respect to traces based on the state ontology Ont in the following manner. Traces and time points can be related to state properties via the relation at, comparable to the Holds-predicate in event calculus (Kowalski and Sergot, 1986) or situation calculus (Reiter, 2001). Thus, at(γ, t, p) denotes that state property p holds in trace γ at time t. Here state propererties are considered objects and denoted by term expressions in the TTL language. Likewise, at(γ, t, ¬p) denotes that state property p does not hold in trace γ at time t. Based on these statements, dynamic properties can be formulated in a formal manner in a sorted predicate logic, using the usual logical connectives such as negation, conjunction, disjunction, implication (denoted by ¬, &, ∨, ⇒ respectively), and universal and existential quantifiers (denoted by ∀, ∃), for example, over traces, time and state properties. An example is the following dynamic property for a pattern concerning belief creation based on observation:

for trace $\gamma \in$ W,
if at any point in time t1 the agent observes that it is wet outside,
then there exists a time point t2 after t1 such that at t2 in the trace the agent believes that it is wet outside

This property can be expressed as a dynamic property in TTL form (with free variable γ) as follows:

$$\forall t{:}T\ [\ at(\gamma, t, observes(itswet)) \ \Rightarrow\ \exists t' \geq t\ \ at(\gamma, t', belief(itswet))\ ]$$

The set DYNPROP(Ont, γ) is the subset of DYNPROP(Ont) consisting of formulae in which γ is either a constant or a free variable.

3.3 Past, Future and Interval Patterns

Let two traces γ_1, γ_2 *coincide* on ontology Ont, and interval [t1, t2), denoted by

$$coincide\_on(\gamma_1, \gamma_2, Ont, t1, t2) \quad \text{or} \quad \gamma_1 =_{Ont,\ [t1,\ t2)} \gamma_2$$

if and only if

$$\forall t{:}T\ \forall a{:}BSTATPROP(Ont)\ \ [\ t1 \leq t < t2 \ \Rightarrow\ \ [at(\gamma_1, t, a) \ \Leftrightarrow\ at(\gamma_2, t, a)]\ ]$$

When no interval is mentioned it is meant that it holds for the whole time frame. Notice that for $\varphi(\gamma)$ in DYNPROP(Ont) it holds that

$$\gamma =_{Ont} \gamma' \Rightarrow\ [\ \varphi(\gamma) \Leftrightarrow \varphi(\gamma')\ ].$$

An *interval pattern* for the time interval [t1, t2) is formalised as a statement that does not depend on time points before t1 or after t2. The subset IPROP(Ont, η, u1, u2) of DYNPROP(Ont, η) (where u1 and u2 are constant parameters for time points and η for traces) is the set of *interval statements* over state ontology Ont with respect to trace η and interval from time point u1 to time point u2. This set is defined by the predicate

$$\begin{array}{l} interval\_statement(\varphi(\eta, u1, u2), Ont, \eta, u1, u2) \quad \equiv \\ \quad \forall \gamma_1, \gamma_2, t1, t2\ \ [\ \ \gamma_1 =_{Ont,\ [t1,\ t2)} \gamma_2 \ \Rightarrow\ [\ \varphi(\gamma_1, t1, t2) \Leftrightarrow\ \varphi(\gamma_2, t1, t2)\ ]\ ] \end{array}$$

In principle, instances of this set can be defined by including for every time quantifier for a time variable s restrictions of the form u1 $\leq$ s, or u1 < s and s $\leq$ u2, or s < u2.
Similarly the sets of past statements and future statements are defined by the predicates

$$\begin{array}{lcl} past\_statement(\varphi(\eta, u2), Ont, \eta, u2) & \equiv & \forall \gamma_1, \gamma_2, t2\ \ [\ \ \gamma_1 =_{Ont,\ <\ t2} \gamma_2 \ \Rightarrow\ [\ \varphi(\gamma_1, t2) \Leftrightarrow\ \varphi(\gamma_2, t2)\ ]\ ] \\ future\_statement(\varphi(\eta, u1), Ont, \eta, u1) & \equiv & \forall \gamma_1, \gamma_2, t2\ \ [\ \ \gamma_1 =_{Ont,\ \geq\ t1} \gamma_2 \ \Rightarrow\ [\ \varphi(\gamma_1, t1) \Leftrightarrow\ \varphi(\gamma_2, t1)\ ]\ ] \end{array}$$

3.4 Formalising Second-Order Dynamic Properties in MetaTTL

The criteria for agency have the form of (second-order) properties of patterns in world dynamics. As patterns in world dynamics are formalised by TTL formulae, formalisation of the criteria for agency take the form of second-order dynamic properties, i.e., properties that refer to dynamic properties expressed within TTL. Such second-order dynamic properties are expressed in MetaTTL: the meta-language of TTL. For more information on how to formalise such a meta-language, see, for example, Attardi and Simi (1984), Bowen and Kowalski (1982), Bowen (1985), Weyhrauch (1980), Vila and Reichgelt (1996). The language MetaTTL includes sorts for DYNPROP(Ont) and its subsets as indicated above, which contain TTL-statements (for dynamic properties) as objects denoted by term expressions. Moreover, a predicate holds on these sorts can be used to express that such a TTL formula is true. When no confusion is expected, this predicate can be left out. To express second-order dynamic properties, in a MetaTTL statement, quantifiers over TTL statements can be used. As TTL-statements are used to formalise patterns in the world's dynamics, quantifiers over TTL statements can be used to express properties about all patterns or about the existence of patterns with certain properties. For example, for $\varphi1$ of sort IPROP(ExtOnt, η, u1, u2) and $\varphi2$ of sort IPROP(IntOnt, η, u1, u2) the MetaTTL formula

$$\forall\gamma{:}W\ \forall t1,t2{:}T\ [\ [\ holds(\varphi1(\gamma,\ t1,\ t2))\ \ \&\ \ t1\leq t2\ ]\ \ \Rightarrow\ \ \exists t3,t4{:}T\ [\ t2\leq t3\leq t4\ \ \&\ holds(\varphi2(\gamma,\ t3,\ t4))\ ]\ ]$$

expresses that

> for any trace γ and time points t1 and t2, when the pattern $\varphi1(\gamma, t1, t2)$ occurs in γ between t1 and t2, then after t2 the pattern $\varphi2(\gamma, t3, t4)$ occurs in γ between some t3 and t4.

For this MetaTTL formula the (definable) abbreviation predicate has_effect is used:

$$\begin{array}{l} has\_effect(\varphi1{:}IPROP(Ont,\ \eta,\ u1,\ u2),\ \varphi2{:}IPROP(Ont',\ \eta,\ u1,\ u2))\ \equiv \\ \qquad \forall\gamma{:}W\ \forall t1,t2{:}T\ [\ [\ holds(\varphi1(\gamma,\ t1,\ t2))\ \ \&\ \ t1\leq t2\ ]\ \ \Rightarrow\ \ \exists t3,t4{:}T\ [\ t2\leq t3\leq t4\ \ \&\ \ holds(\varphi2(\gamma,\ t3,\ t4))\ ]\ ] \end{array}$$

In the next sections, the six criteria for agency will be formalised in MetaTTL. This will show various examples where quantifiers over patterns such as $\varphi1$ and $\varphi2$ are used.

4 Boundary Separating Internal and External

To start with the first boundary criterion suppose WorldOnt is the world state ontology used. It is assumed that this set is the union of a collection of subsets, each of which collects the ontology elements within WorldOnt related to a certain location (local ontology). This collection of local ontologies can be considered a set of locations; it is called LOC, so WorldOnt = $\bigcup$ LOC = $\bigcup_{L\in LOC}$ L. Based on this, the set of *local basic world state properties* for location L is BSTATPROP(L), and the set of *local world state properties* is STATPROP(L). Finally,

$$WBSTATPROP\ \ =\ \bigcup_{L\in\ LOC}\ BSTATPROP(L)$$

$$WSTATPROP\ \ =\ \bigcup_{L\ \in LOC}\ STATPROP(L)$$

denote the overall sets of (basic) world state properties.

An ontological assumption for agency is that in the world a distinction can be made between sets of locations: *internal* and *external* locations, and a *boundary* that has two specific parts: the part affectable from outside (*input*), and the part affectable from inside (*output*). The rest of the boundary (if any) is not affectable (e.g, a shell). To formalise this, the collection LOC is partitioned into three disjoint subsets INTLOC, EXTLOC, BOUNDLOC. Within BOUNDLOC two disjoint subsets INLOC and OUTLOC are distinguished that may not exhaust BOUNDLOC. The union of INLOC and OUTLOC is INTERACTIONLOC. So, the following relationships between these sets exist:

$$\begin{array}{ll} \mathrm{INTLOC} \cup \mathrm{EXTLOC} \cup \mathrm{BOUNDLOC} = \mathrm{LOC} & \text{(disjoint union)} \\ \mathrm{INLOC},\ \mathrm{OUTLOC} \subseteq \mathrm{BOUNDLOC} & \text{(disjoint subsets)} \\ \mathrm{INTERACTIONLOC} = \mathrm{INLOC} \cup \mathrm{OUTLOC} & \end{array}$$

According to this, the following ontologies are defined:

$$\begin{array}{llll} \mathrm{IntOnt} = \bigcup \mathrm{INTLOC} & & \mathrm{ExtOnt} = \bigcup \mathrm{EXTLOC} \\ \mathrm{BoundOnt} = \bigcup \mathrm{BOUNDLOC} & & \mathrm{InteractionOnt} = \bigcup \mathrm{INTERACTIONLOC} \\ \mathrm{InOnt} = \bigcup \mathrm{INLOC} & & \mathrm{OutOnt} = \bigcup \mathrm{OUTLOC} \end{array}$$

On this basis also the other sets can be grouped; e.g., BSTATPROP(IntOnt), STATPROP(IntOnt), and DYNPROP(IntOnt).

To make the above more concrete, consider the example (static) world description depicted in Figure 1.

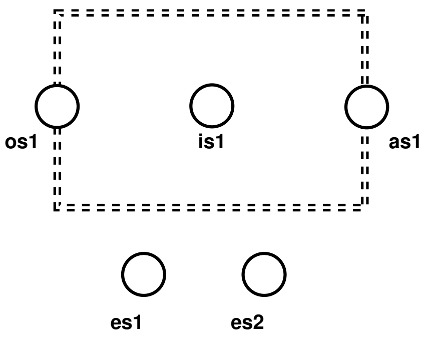

Figure 1. Example world

This figure describes a process in the world to be considered as an agent and its environment. The box indicates the boundaries of the agent, small circles denote basic state properties. Those within the box are internal state properties, those outside are external, those on the left of the box are input state properties, those on the right output state properties. The following ontologies are used for this example:

IntOnt = {is1} ExtOnt = {es1, es2} InOnt = {os1} OutOnt = {as1}

Here, is1 stands for 'internal state 1', es1 stands for 'external state 1', es2 stands for 'external state 2', os1 stands for 'observation state 1', and as1 stands for 'action state 1'. The union is WorldOnt. Note that BoundOnt = InteractionOnt in this description.

Now that the assumptions about the (static) world state ontology have been defined, the next five sections will address criteria concerning the world dynamics.

5 Isolation

The isolation principle expresses that influences between internal and external state properties can only occur via the input states and output states. Informally, this criterion for influences from outside to inside can be stated as follows:

For all dynamic properties φ1 referring to only external states,
and for all dynamic properties φ3 referring to only internal states,
if for all traces γ, φ1 implies later φ3,
then there is also a dynamic property φ2 referring to only input states, such that
φ1 implies later φ2 and φ2 implies later φ3 in all traces.

In MetaTTL, this principle can be formalised as follows, using the abbreviation based on the predicate has_effect:

isolation(ExtOnt, InputOnt, IntOnt) ≡
∀φ1:IPROP(ExtOnt, η, u1, u2) ∀φ3:IPROP(IntOnt, η, u1, u2)
has_effect(φ1, φ3) ⇒
∃φ2:IPROP(InputOnt, η, u1, u2) [has_effect(φ1, φ2) & has_effect(φ2, φ3)]

This definition can be illustrated by considering Figure 2. This picture shows how possible instances of φ1, φ2 and φ3 are located with respect to an agent. Dotted ovals indicate dynamic properties which are built up from the state properties they contain. Arrows denote (temporal) implications between dynamic properties. The idea of the picture is that, if an instance of the thick arrow exists, then also instances of the thin arrows can be found. The isolation criterion for influences from inside to outside via output states can be defined by interchanging ExtOnt and IntOnt and replacing InOnt by OutOnt in the above formalisation: isolation(IntOnt, OutputOnt, ExtOnt).

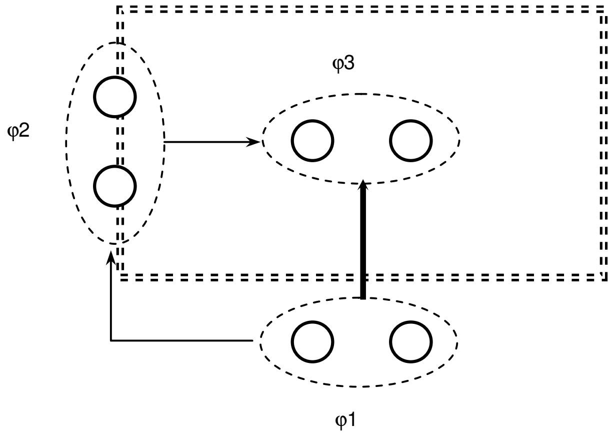

Figure 2. Isolation Principle

6 Modular World Dynamics

According to the modular world dynamics principle, the dynamics of the world is structured in a modular form, based on dynamic relationships that are purely internal and dynamic relationships that are purely external. Informally, this criterion states the following:

> For all traces γ,
> if a certain dynamic property ψ over the world ontology holds for γ,
> then there is a dynamic property $\varphi1$, referring to only external and interaction states,
> and there is a dynamic property $\varphi2$, referring to only internal and interaction states,
> such that $\varphi1$ and $\varphi2$ hold for γ, and for all traces γ', $\varphi1$ and $\varphi2$ together imply ψ.

In MetaTTL, this criterion is formalised as follows:

$$
\begin{aligned}
&\text{modular\_world\_dynamics} \equiv \\
&\quad \forall\psi\text{:IPROP(WorldOnt, }\eta\text{, u1, u2)} \\
&\quad \forall\gamma\text{:W } \forall\text{t1,t2:T [holds(}\psi(\gamma\text{, t1, t2)) \& t1}\leq\text{t2} \;\Rightarrow \\
&\qquad \exists\varphi1\text{:IPROP(ExtOnt } \cup \text{ InteractionOnt, }\eta\text{, u1, u2) } \exists\varphi2\text{:IPROP(IntOnt } \cup \text{ InteractionOnt, }\eta\text{, u1, u2)} \\
&\qquad\qquad \text{holds(}\varphi1(\gamma\text{, t1, t2)) \& holds(}\varphi2(\gamma\text{, t1, t2)) \&} \\
&\qquad\qquad [\forall\gamma'\text{:W [holds(}\varphi1(\gamma'\text{, t1, t2)) \& holds(}\varphi2(\gamma'\text{, t1, t2))]} \;\Rightarrow\; \text{holds(}\psi(\gamma'\text{, t1, t2))]]}
\end{aligned}
$$

Also see the (two-dimensional) Figure 3. Again, the three dotted shapes (named ψ, $\varphi1$, and $\varphi2$) indicate dynamic properties which are built up from the state properties they contain, and arrows denote (temporal) implications between dynamic properties.

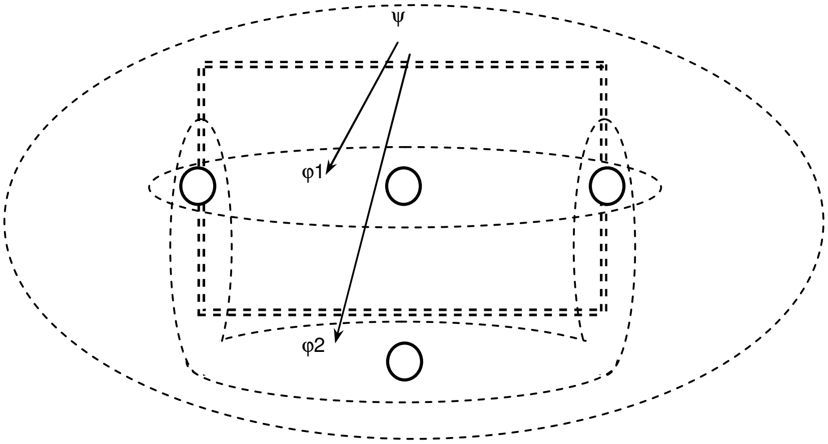

Figure 3. Modular World Dynamics Principle

7 Input-Output Dynamics Relations

In Kim (1996, pp. 85-91) a relation between input and output is called an input-output correlation. In this paper this is considered a relation between series of input states over time (input traces) and series of output states over time (output traces). This relation may or may not be functional. In case the relation is functional, there is a function mapping input state dynamics (traces) onto output state dynamics (traces). In case the relation is not functional, it has a non-deterministic nature (e.g., a probabilistic relation). This criterion on patterns in world dynamics can be formalised as follows. A first step is as a relation or function between input and output traces, generalising the functionality descriptions in Treur (2002), a relation IOR on the cartesian product of input traces and output traces:

IOR : TRACES(InOnt) x TRACES(OutOnt).

If this relation is functional, i.e., if IOR(γ_1, γ_2) and IOR(γ_1, γ_3) implies $\gamma_2 = \gamma_3$, then a function IOF exists:

IOF : TRACES(InOnt) → TRACES(OutOnt).

A further formalisation is by implicit and explicit definability of output traces in terms of input traces, generalising these concepts from Chang and Keisler (1973), and Leemans, Treur, and Willems (2002). For the deterministic, functional case, implicit definability means:

If for two traces the dynamics of input states is the same,
then also the dynamics of the output states is the same.

For this functional case, implicit definability is formally expressed by

$$\forall\gamma,\gamma'{:}W\ \ [\ \gamma =_{InOnt} \gamma' \Rightarrow\ \gamma =_{OutOnt} \gamma'\ ]$$

For the nonfunctional case it can be expressed as:

If for two traces the dynamics of input states is the same,
then there is a trace with
the same external and input dynamics of one of these traces
and the same internal and output dynamics as the other trace.

Formally:

$$\forall\gamma,\gamma'{:}W\ \ \gamma =_{InOnt} \gamma'\ \ \Rightarrow\ \ \exists\gamma''{:}W\ \ \ \gamma'' =_{ExtOnt\cup InOnt} \gamma'\ \ \&\ \ \gamma'' =_{IntOnt\cup OutOnt} \gamma.$$

Explicit definability means:

There is a dynamic property expressed in the specification language used
that relates the input states over time to output states over time.

For the functional case this is as follows. For $\varphi(\eta)$ in $\mathrm{DYNPROP(InteractionOnt)}$, let

$$\mathrm{input\_output\_correlation}(\varphi(\eta))$$

denote

$$\begin{aligned}
&\forall\gamma{:}\mathrm{TRACES}\ \ [\ \ \mathrm{holds}(\varphi(\gamma)) \Leftrightarrow\ \ \exists\gamma'{:}W\ \ [\ \gamma =_{InOnt} \gamma'\ \&\ \gamma =_{OutOnt} \gamma'\ ]\ \&\\
&\forall\gamma{:}W\ \ \exists\gamma'{:}\mathrm{TRACES}\ [\ \gamma =_{InOnt} \gamma'\ \ \&\ \ \mathrm{holds}(\varphi(\gamma'))\ ]\ \ \&\\
&\forall\gamma,\gamma'{:}\mathrm{TRACES}\ [\mathrm{holds}(\varphi(\gamma))\ \&\ \mathrm{holds}(\varphi(\gamma'))\ \&\ \gamma =_{InOnt} \gamma'\ \ \Rightarrow\ \ \gamma =_{OutOnt} \gamma'\ \ ]\ ]
\end{aligned}$$

Then, for the functional case, explicit definability is:

$$\exists\varphi(\eta){:}\mathrm{DYNPROP(InteractionOnt)}\quad \mathrm{input\_output\_correlation}(\varphi(\eta)),$$

see Figure 4. For the nonfunctional case the third conjunct can be left out:

$$\begin{aligned}
&\forall\gamma{:}\mathrm{TRACES}\ \ [\ \ \mathrm{holds}(\varphi(\gamma)) \Leftrightarrow\ \ \exists\gamma'{:}W\ \ [\ \gamma =_{InOnt} \gamma'\ \&\ \gamma =_{OutOnt} \gamma'\ ]\ \&\\
&\forall\gamma{:}W\ \ \exists\gamma'{:}\mathrm{TRACES}\ [\ \gamma =_{InOnt} \gamma'\ \ \&\ \ \mathrm{holds}(\varphi(\gamma'))\ ]\ ]
\end{aligned}$$

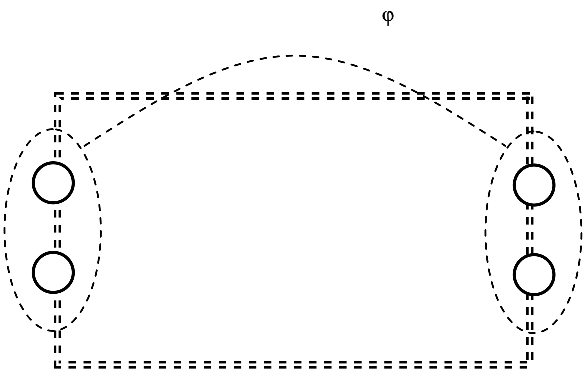

Figure 4. Input-Output Dynamics Relations

8 Internal and Interaction Dynamics Relations

Given an input-output dynamics relation φ(η) in DYNPROP(InteractionOnt), this can be related to the internal dynamics described by π(η) in DYNPROP(IntOnt∪InteractionOnt) as follows:

internal_interaction_relation(π(η),φ(η)) ≡
$\forall\gamma$:W [holds(π(γ)) ⇒ holds(φ(γ))] & $\forall\gamma$:W [holds(φ(γ)) ⇒ $\exists\gamma'$:W [holds(π(γ')) & $\gamma' =_{\text{InteractionOnt}} \gamma$]]

Then the criterion is (see also Figure 5):

$\forall$φ(η):DYNPROP(InteractionOnt) [input_output_correlation(φ(η)) ⇒
$\exists$π(η):DYNPROP(IntOnt∪InteractionOnt) internal_interaction_relation(π(η),φ(η))]

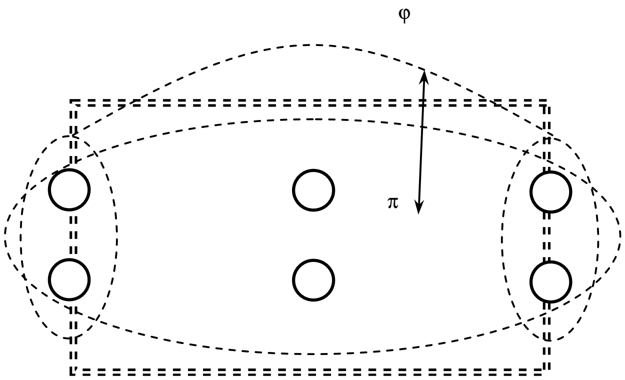

Figure 5. Internal-Interaction Dynamics Relations

9 Representation Relations

In the literature on Philosophy of Mind different types of approaches to representational content of an internal state property have been put forward, for example the causal/correlational, interactivist and relational specification approach; cf. Bickhard (1993); Kim (1996), pp. 191-193, 200-202. For this paper we adopt the relational specification approach; cf. Kim (1996), pp. 200-202. The formalisation of this approach can be done as follows. Suppose p is an internal state property. A relational specification for p is made by a formula $\varphi(\eta, u)$ in DYNPROP(ExtOnt $\cup$ { p}) that specifies how a certain pattern in the dynamics of external world states relates to p. Here ExtOnt can also be replaced by InteractionOnt to relate p to a pattern in the dynamics of the interaction states. A relational specification can also be obtained in a more specific manner by relating p separately to a past pattern and to a future pattern. Then two formulae $\varphi_P(\eta, u)$ and $\varphi_F(\eta, u)$ exist in DYNPROP(ExtOnt) (or DYNPROP(InteractionOnt)), where the former is a past formula and the latter a future formula. Based on this, the criterion representation_relations expresses that for all p in STATPROP(IntOnt) there exist formulae $\varphi_P(\eta, u)$ and $\varphi_F(\eta, u)$ that can be related to p by biconditionals (see also Figure 6):

is_past_representation_relation_for($\varphi_P(\eta, u)$, p) $\equiv$
past_statement($\varphi_P(\eta, u)$, ExtOnt, η, u) &
$\forall\gamma$:W $\forall$t:T [holds($\varphi_P(\gamma, t)$) $\Leftrightarrow$ holds(at(γ, t, p)]
is_future_representation_relation_for($\varphi_F(\eta, u)$, p) $\equiv$
future_statement($\varphi_F(\eta, u)$, ExtOnt, η, u) &
$\forall\gamma$:W $\forall$t:T [holds($\varphi_F(\gamma, t)$) $\Leftrightarrow$ holds(at(γ, t, p)]
has_two_sided_representation_relations(p) $\equiv$
$\exists\varphi_P(\eta, u), \varphi_F(\eta, u)$:DYNPROP(ExtOnt)
past_representation_relation_for($\varphi_P(\eta, u)$, p) & future_representation_relation_for($\varphi_F(\eta, u)$, p)

representation_relations $\equiv$ $\forall$p:STATPROP(IntOnt) has_two_sided_representation_relations(p)

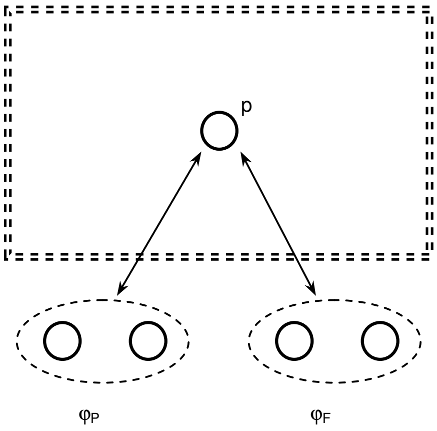

Figure 6. Representation Relation

10 Case Study

To illustrate how the above criteria for agency apply to a specific example, this section describes a simple case study.

10.1 Boundary Separating Internal and External

In the case study, the following five basic state properties are considered (similar to Figure 1):

IntOnt = {is1} ExtOnt = {es1, es2} InOnt = {os1} OutOnt = {as1}

This satisfies the first criterion. The basic dynamical relationships of the case study are represented graphically in Figure 7; this defines the set of traces W for the example.

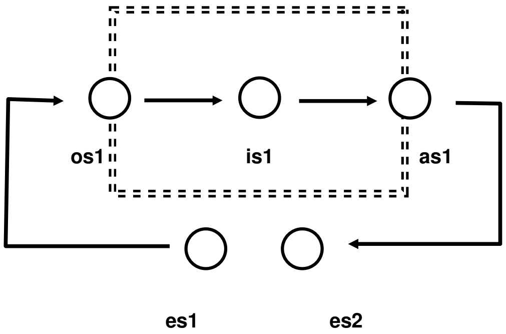

Figure 7. Relationships within the case study

Circles denote state properties; the arrows denote relationships between state properties. For example, the arrow from os1 to is1 indicates that the occurrence of os1 leads to the occurrence of is1. Furthermore, the state properties are assumed to be non-persistent. Thus, whenever os1 ceases to exists, is1 also ceases to exist.

Based on these relationships, that define the set of traces W, a number of simulation traces have been produced, using the LEADSTO language and its simulation software (Bosse et al., 2005). This language enables to model direct temporal dependencies between two state properties in successive states. This executable format is defined as follows. Let α and β be state properties of the form 'conjunction of atoms or negations of atoms', and e, f, g, h non-negative real numbers. Then the notation $\alpha \twoheadrightarrow_{e, f, g, h} \beta$, means:

If state property α holds for a certain time interval with duration g

then after some delay (between e and f) state property β will hold for a certain time interval of length h.

A trace γ *satisfies* a LEADSTO expression $\alpha \twoheadrightarrow_{e,\, f,\, g,\, h} \beta$ denoted by $\gamma \models \alpha \twoheadrightarrow_{e,\, f,\, g,\, h} \beta$ if

$$\begin{array}{l} \forall t1\ \ [\forall t\ \ [t1{-}g \le t < t1 \ \Rightarrow\ at(\gamma, t, \alpha)\,]\ \Rightarrow \\ \qquad \exists d\ \ [e \le d \le f\ \&\ \forall t'\ \ [t1{+}d \le t' < t1{+}d{+}h\ \Rightarrow\ at(\gamma', t, \beta)\,]\,] \end{array}$$

which also can be used as a definition of the LEADSTO format in terms of the language TTL. A specification of dynamic properties in LEADSTO format has as advantages that it is executable and that (besides in textual or formal format), it can often easily be depicted graphically (as in Figure 7). The LEADSTO format has shown its value especially when temporal relations for basic mechanisms in the (continuous) physical world are modelled and simulated; for example, in cooperation with cell biologists, the bacterium *E. coli* and its intracellular chemistry have been modelled as an agent in LEADSTO (Jonker, Snoep, Treur, Westerhoff, and Wijngaards, 2008). The textual specification in LEADSTO format of the example depicted in Figure 7 is as follows:

LP1 $\quad es1 \twoheadrightarrow_{e,\, f,\, g,\, h} os1$

LP2 $\quad os1 \twoheadrightarrow_{e,\, f,\, g,\, h} is1$

LP3 $\quad is1 \twoheadrightarrow_{e,\, f,\, g,\, h} as1$

LP4 $\quad as1 \twoheadrightarrow_{e,\, f,\, g,\, h} es2$

Here LP2 and LP3 describe the internal process, and LP1 and LP4 describe (part of) the external process. This specification describes the set W characterising the world for the example:

$$W\ =\ \{\ \gamma \in \text{TRACES(Ont)}\ \mid\ \gamma \models \text{LP1}\ \&\ \text{LP2}\ \&\ \text{LP3}\ \&\ \text{LP4}\ \}$$

Simulation is performed by execution of the LEADSTO rules (similar to executable temporal logic; e.g., Fisher, 2005), thus generating a trace that satisfies all of these rules, and therefore in W. An example of a trace in W as generated by the LEADSTO specification described above is shown in Figure 8 (e, f, g, h all have been taken 1). Here, time is on the horizontal axis, and the state properties are on the vertical axis. A mark on top of a line indicates that a state property is true at that time point.

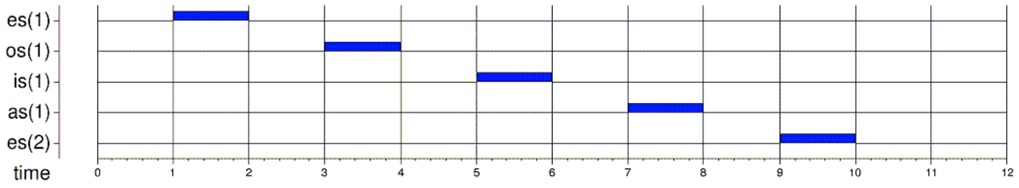

Figure 8. Example world trace

In the following sections, it is explained in more detail why in addition to the boundary criterion, also the other five criteria for agency hold for these generated traces, which form a representative subset of W.

10.2 Isolation

To start, it is illustrated why the property isolation holds in this case study (also see Figure 2 and the informal description above this figure). Obviously, it is difficult to provide a complete proof for this criterion, since the number of dynamic properties that can be filled in for $\varphi 1$ and $\varphi 3$ in principle is large. Therefore, we restrict ourselves to explaining why the criterion holds for some given instances of $\varphi 1$ and $\varphi 3$. Suppose, for example, that the following dynamic properties correspond to $\varphi 1$ in terms of the external state ontology and $\varphi 3$ in terms of the internal state ontology:

$$\varphi 1(\eta, u1, u2) \quad \equiv \quad at(\eta, u1, es1)\ \&\ at(\eta, u2, \neg\, es1)$$
$$\varphi 3(\eta, u1, u2) \quad \equiv \quad at(\eta, u1, is1)\ \&\ at(\eta, u2, \neg\, is1)$$

Then, as

$$holds(\varphi 1(\gamma{:}W, t1{:}T, t2{:}T)) \Rightarrow \exists t3, t4\ [\ t2 \leq t3 \leq t4\ \&\ holds(\varphi 3(\gamma{:}W, t3{:}T, t3{:}T))\ ]$$

it holds

$$has\_effect(\varphi 1, \varphi 3)$$

Therefore, according to the isolation principle, there is a $\varphi 2$ in terms of the input ontology to be found such that, in all traces, $\varphi 1$ implies (later) $\varphi 2$ and $\varphi 2$ implies (later) $\varphi 3$, or:

$$has\_effect(\varphi 1, \varphi 2)\ \&\ has\_effect(\varphi 2, \varphi 3)$$

Such a $\varphi 2$ can indeed be found:

$$\varphi 2(\eta, u1, u2) \quad \equiv \quad at(\eta, u1, os1)\ \&\ at(\eta, u2, \neg os1)$$

Given this instance of $\varphi 2$, the property isolation indeed holds for the case study. This can be made more clear by looking at the model described in Figure 7. Intuitively, for all traces in W, which are the traces that can be generated on the basis of this model (such as Figure 8), it is clear that if they satisfy $\varphi 1$ (i.e., first es1 holds and later es1 does not hold), then later $\varphi 3$ will hold (i.e., first is1 holds and later is1 does not hold), and that they also will satisfy $\varphi 2$ in between (i.e., first os1 holds and later os1 does not hold). For the set of traces that have been generated as representative example traces in W, this has been checked automatically, using the TTL checking software described in (Bosse et al, 2006), and found confirmed.

10.3 Modular World Dynamics

Next, the criterion modular_world_dynamics is addressed. Consider the description of this criterion given earlier (also see Figure 3). Again, it is explained why the criterion holds for a given instance of ψ. Thus, as an example for ψ suppose

$$\begin{aligned}&\psi(\eta, u1, u2) \equiv \\ &\quad \forall t{:}T\ [\ [u1 \leq t < u2\ \&\ at(\eta, t, es1)] \Rightarrow \exists t'\ [\ t < t' \leq u2\ \&\ at(\eta, t', is1)]\ ]\ \& \\ &\quad \forall t'{:}T\ [\ [u1 < t' \leq u2\ \&\ at(\eta, t', is1)] \Rightarrow \exists t\ [u1 \leq t < t'\ \&\ at(\eta, t, es1)]\ ]\end{aligned}$$

Then, according to the modular world dynamics principle, there are $\varphi1$ and $\varphi2$ to be found that hold for γ and such that, in all traces, $\varphi1$ and $\varphi2$ together imply ψ. These $\varphi1$ and $\varphi2$ can indeed be found:

$$\begin{aligned}&\varphi1(\eta, u1, u2) \equiv \\ &\quad \forall t{:}T\ [\ [u1 \leq t < u2\ \&\ at(\eta, t, es1)] \Rightarrow \exists t'\ [\ t < t' \leq u2\ \&\ at(\eta, t', os1)]\ ]\ \& \\ &\quad \forall t'{:}T\ [\ [u1 < t' \leq u2\ \&\ at(\eta, t', os1)] \Rightarrow \exists t\ [u1 \leq t < t'\ \&\ at(\eta, t, es1)]\ ]\end{aligned}$$

$$\begin{aligned}&\varphi2(\eta, u1, u2) \equiv \\ &\quad \forall t{:}T\ [\ [u1 \leq t < u2\ \&\ at(\eta, t, os1)] \Rightarrow \exists t'\ [\ t < t' \leq u2\ \&\ at(\eta, t', is1)]\ ]\ \& \\ &\quad \forall t'{:}T\ [\ [u1 < t' \leq u2\ \&\ at(\eta, t', is1)] \Rightarrow \exists t\ [u1 \leq t < t'\ \&\ at(\eta, t, os1)]\ ]\end{aligned}$$

Given these instances of $\varphi1$ and $\varphi2$, the property modular_world_dynamics indeed holds for this ψ in this case.

10.4 Input-Output Dynamics Relation

Next, it is shown that the criterion input_output_correlation (see Figure 4) can be satisfied for the case study. This can be done by choosing the following instance for φ:

$$\begin{aligned}\varphi(\eta) \equiv\ &\forall t{:}T\ [\ at(\eta, t, os1) \Rightarrow \exists t' > t\ \ at(\eta, t', as1)\ ]\ \& \\ &\forall t{:}T\ [\ at(\eta, t, as1) \Rightarrow \exists t' < t\ \ at(\eta, t', os1)\ ]\end{aligned}$$

10.5 Internal-Interaction Dynamics Relation

Next, the case study satisfies the criterion internal_interaction_relation (see Figure 5) with the following instance for π:

$$\begin{aligned}\pi(\eta) \equiv\ &\forall t\ [\ at(\eta, t, os1) &&\Rightarrow \exists t' > t\ \ at(\eta, t, is1)\ ]\ \& \\ &\forall t\ [\ at(\eta, t, is1) &&\Rightarrow \exists t' > t\ \ at(\eta, t, as1)\ ]\ \& \\ &\forall t\ [\ at(\eta, t, is1) &&\Rightarrow \exists t' < t\ \ at(\eta, t, os1)\ ]\ \& \\ &\forall t\ [\ at(\eta, t, as1) &&\Rightarrow \exists t' < t\ \ at(\eta, t, is1)\ ]\end{aligned}$$

10.6 Representation Relations

Finally, it is shown that appropriate representation relations can be defined for the internal state properties in the case study. To this end, consider the criterion representation_relations (see Figure 6). Suppose that p corresponds to the state property is1. Then, for φ_P and φ_F the following dynamic properties yield correct representation relations:

$$\varphi_P(\eta, u) \equiv \exists t':T\ [\ t'<u\ \&\ at(\eta, t', es1)\ ]$$
$$\varphi_F(\eta, u) \equiv \exists t':T\ [\ t'>u\ \&\ at(\eta, t', es2)\ ]$$

11 Discussion

In this paper, the question is addressed which criteria on patterns in world dynamics indicate an adequate conceptualisation of a world's process as an agent. Here the world can be a physical or social world. Moreover, artificial and cultural worlds such as virtual worlds and economical worlds are covered as well. Also hybrid worlds are possible, including both natural and artificial elements (e.g., a robot on Mars, or a human interacting with a virtual environment). Whatever world is considered, a minimal demand is that the world's dynamics can be analysed and formalised. Among the examples that can be addressed are biological organisms, organisations within society such as a company structured according to the 'front office – back office' principle, and robots.

Six criteria in the form of (second-order) properties of patterns in world dynamics were discussed that indicate when the world shows agency, or at least allows a reasonable agent-based conceptualisation. As a naturalist perspective is taken, the criteria can be used to find out whether a given dynamic phenomenon can be considered an agent in a faithful manner. Such a phenomenon can be, for example, an organisation within society that attempts to behave in a coherent manner to its environment. If every member of this organisation has its own direct interaction with the external world and is affected by this, then an analysis based on the conceptual framework introduced here will show that there is no separate internal process, and hence the criteria 'isolation' and 'modular dynamics' will fail. If log files of the processes of such a company are given, then such an analysis can be supported by automated checking software that has been developed.

Notice that it is not claimed that the criteria are independent or non-overlapping. For example, under certain conditions isolation may entail also modular world dynamics. In future work relations between the criteria will be investigated more extensively.

Our claim is not that the list of six criteria is the one and only truth about agency emerging from world dynamics. An aspect for further investigation is how different notions of agency can be defined on the basis of certain subsets or specialisations or extensions of the criteria mentioned (e.g., purely reactive agents, or agents with beliefs, desires and intentions, or self-aware agents).

Also in Stuart (2002) and Dobbyn and Stuart (2003), criteria for agency are (informally) discussed. Five of their six criteria seem in line with our criteria, except that they claim that a certain richness (e.g., of external world, of input, of output) should be demanded. Moreover, their criterion of representation indicates internal representations of not only external but also internal processes (they aim that an agent is aware of itself). This can be added to the sixth criterion. Their second criterion deals with the possession of self-directed goals. For us, this could be added as a criterion for a more specialised self-aware, goal-directed agent notion.

Our first criterion deals with the possibility to distinguish a boundary separating the internal and external area in the world. Although much literature exists that supports this as an important criterion, there is also literature that casts doubt on whether always a boundary can be found; e.g., Clark and Chalmers (1998). Indeed for the phenomenon of extended mind the boundary seems larger than the skin of an organism. One of the issues to be further investigated is whether such an extended boundary can be defined according to the framework presented in this paper.

A question that may arise is to which extent the criteria as discussed and formalised are internal-external symmetric in the sense that replacing ‘internal’ by ‘external’ and ‘input’ by ‘output’, and vice versa, obtains the same criteria. Is the external world also an agent according to these criteria? How is the internal area distinguished (as being an agent) by the criteria from the external world (as not being an agent)? Indeed, the first three criteria are internal-external symmetric: boundary separation internal and external, isolation, modular world are all internal-external symmetric. However, the other three criteria are not internal-external symmetric. The fourth criterion on input-output dynamics relations has a direction from input to output, and not in the other direction from output to input. Moreover, there is the fifth criterion on the relation between internal and interaction dynamics, but no criterion on the relation between external and interaction dynamics. Finally, the sixth criterion claims representational relations for internal state properties but not for external state properties.

In how far is it possible to extend the list of criteria in a reasonable manner to obtain internal-external symmetry? It can be imagined that the internal-external mirror image criterion of the fourth criterion on input dynamics relations can also be postulated, thus assuming that the external world can be described in its effects over time on the input states (given the output states over time) by an output-input dynamics relation. In the same line it may be imagined that also for the fifth criterion on relations between internal dynamics and interaction dynamics, the mirror image can be added, claiming a relationship between the external world’s dynamics and the output-input dynamics. The two mirror images of the fourth and fifth criterion would imply additional assumptions on in how far the external world is describable in terms of temporal specifications (not necessarily in a deterministic manner). Cases may be considered that these indeed are reasonable assumptions, but also cases may be possible that these assumptions are not fulfilled. For the sixth criterion on representation relations, the situation seems more inherently asymmetric. The mirror image of this criterion would state that the external state properties have representation relations to the internal processes. This does not make sense for almost all situations that can be imagined. It would mean that the internal process would play a role as a kind of almighty power, determining all states in the external world. This could only make sense when the external world is very limited.

Summarising the above deliberations, the background for asymmetry in the criteria is mainly found in these two points:

- The external world may be not fully describable
- Usually not all states in the external world are determined by the internal processes

These seem sufficient reasons to have the criteria asymmetric, which has the positive implication that the criteria do not imply doubt on where the agent is to be found, in the internal area or the external area, or both.

Formalisation of the criteria has been done in the form of second-order dynamic properties expressed in the sorted predicate logic-based language MetaTTL. This approach is comparable to a certain extent to the approach to mental state properties defined as second-order world properties; cf. Kim (2005, pp. 98-102). Here, for example, the mental state ‘being in pain’ is defined as ‘there exists a physical state property p such that tissue damage leads to p and p leads to shouting ouch!’. Mental state properties defined in this manner are called functionalised, as their function is made explicit in this definition, abstracting from their physical realisation. Kim’s second-order properties are limited to *state* properties, which is an important difference with our case, as we deal with second-order *dynamic* properties. On the other hand, the idea of functionalisation seems a common aspect, as also in our case the second-order dynamic world properties indicate how the world functions in the sense of its dynamic pattern(s), abstracting from the specific realisation of such dynamic patterns.

Another area of further research is to combine formalisms for causal or probabilistic networks with the formalisation of agency presented here, to have a way of indicating that a certain subgraph in such a network can be considered an agent.

Acknowledgements

The authors are grateful to Catholijn Jonker, Simon McGregor, Alexei Sharpanskykh, Vera Stebletsova and Allard Tamminga for discussions about parts of this work.

References

Aleksander, I. (1996). *Impossible Minds: My Neurons, My Consciousness*, Imperial College Press, London UK.

Attardi, G., and Simi, M. (1984). Metalanguage and reasoning across viewpoints. In Tim O'Shea, (ed.), *Proc. 6th European Conference on AI, ECAI'84*. North-Holland, 1984, pp. 413-422

Bernard, C. (1865). *Introduction a l'etude de la medecine experimentale*. Paris: J. Baillierre et fils.

Bickhard, M.H. (1993). Representational Content in Humans and Machines. *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 5, pp. 285-333.

Bosse, T., Jonker, C.M., Meij, L. van der, and Treur, J., A Language and Environment for Analysis of Dynamics by Simulation. *International Journal of Artificial Intelligence Tools*, vol. 16, 2007, pp. 435-464.

Bosse, T., Jonker, C.M., Meij, L. van der, Sharpanskykh, A., and Treur, J. (2009). Specification and Verification of Dynamics in Agent Models. *International Journal of Cooperative Information Systems*, vol. 18, 2009, pp. 167 - 193.

Bosse, T., Jonker, C.M., and Treur, J., (2009). Representation for Reciprocal Agent-Environment Interaction. *Cognitive Systems Research Journal*, vol. 10, 2009, pp. 366-376.

Bowen K.A. (1985). Meta-Level Programming and Knowledge Representation, *New Generation Computing*, 3 (1985), 359-383.

Bowen, K. and Kowalski, R., (1982). Amalgamating language and meta-language in logic programming. In: K. Clark, S. Tarnlund (eds.), *Logic programming*. Academic Press, 1982.

Brewer, B. (1992). Self-location and agency, *Mind*, vol. 101, pp 17-34.

Cannon, W.B. (1932). *The Wisdom of the Body*. New York: W.W. Norton and Co.

Chang, C.C., Keisler, H.J. (1973). *Model theory*, North Holland.

Clark, A and Chalmers, D. J. (1998). The Extended Mind. *Analysis* 58(1):7-19

Clarke, E.M., O. Grumberg, D.A. Peled, (1999). *Model Checking*, MIT Press, Cambridge Massachusetts, London England.

Damasio, A. (2000). *The Feeling of What Happens: Body, Emotion and the Making of Consciousness*. MIT Press.

Dobbyn, C., Stuart, S. (2003). The Self as an Embedded Agent. *Minds and Machines,* vol. 13, pp. 187-201.

Fisher, M. (2005). Temporal Development Methods for Agent-Based Systems, Journal of Autonomous Agents and Multi-Agent Systems, vol. 10, pp. 41-66.

Galton, A. (2003). Temporal Logic. *Stanford Encyclopedia of Philosophy*, URL:
http://plato.stanford.edu/entries/logic-temporal/#2.

Galton, A. (2006). Operators vs Arguments: The Ins and Outs of Reification. *Synthese*, vol. 150, 2006, pp. 415-441.

Jacob, P. (1997). *What Minds Can Do: Intentionality in a Non-Intentional World*. Cambridge: Cambridge University Press, 1997

Jonker, C.M., Snoep, J.L., Treur, J., Westerhoff, H.V., Wijngaards, W.C.A., (2008). BDI-Modelling of Complex Intracellular Dynamics. *Journal of Theoretical Biology*, vol. 251, 2008, pp. 1–23.

Jonker, C.M. and Treur, J. (2002). Compositional Verification of Multi-Agent Systems: a Formal Analysis of Pro-activeness and Reactiveness. *International Journal of Cooperative Information Systems*, vol. 11, pp. 51-92.

Jonker, C.M., and Treur, J. (2003). A Temporal-Interactivist Perspective on the Dynamics of Mental States. *Cognitive Systems Research Journal,* vol. 4, pp. 137-155.

Jonker, C.M., Treur, J., and Wijngaards, W.C.A. (2003). A Temporal Modelling Environment for Internally Grounded Beliefs, Desires and Intentions. *Cognitive Systems Research Journal*, vol. 4, 2003, pp. 191-210.

Keijzer, F. (2002). Representation in Dynamical and Embodied Cognition. *Cognitive Systems Research Journal*, vol. 3, 2002, pp. 275-288.

Kim, J. (1996). *Philosophy of Mind*. Westview Press.

Kim, J. (2005). *Physicalism, or Something Near Enough.* Princeton University Press, Princeton.

Kowalski, R., and Sergot, M. (1986). *A Logic-Based Calculus of Events*. New Generation Computing, 4:67--95, 1986.

Leemans, N.E.M., Treur, J., and Willems, M. (2002). A Semantical Perspective on Verification of Knowledge. *Data and Knowledge Engineering,* vol. 40, pp. 33-70.

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.

Reiter, R. (2001). *Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems*. MIT Press, 2001.

Sharpanskykh, A., and Treur, J., (2010). A Temporal Trace Language for Formal Modelling and Analysis of Agent Systems. In: Dastani, M., Hindriks, K.V., and Meyer, J.J.Ch. (eds.), *Specification and Verification of Multi-Agent Systems*. Springer Verlag, 2010, pp. 317-352.

Stuart, S. (2002). A Radical Notion of Embeddedness: A Logically Necessary Precondition for Agency and Self-Awareness, *Journal of Metaphilosophy,* vol. 33, pp. 98-109.

Sun, R. (2000). Symbol grounding: a new look at an old idea. *Philosophical Psychology*, Vol.13, No.2, 2000, pp. 149-172.

Treur, J. (2002). Semantic Formalisation of Interactive Reasoning Functionality. *International Journal of Intelligent Systems,* vol. 17, pp. 645-686.

Vila, L., and Reichgelt, H. (1996). The Token Reificacion Approach to Temporal Reasoning, *Artificial Intelligence*, vol. 83, May 1996.

Weyhrauch, R.W. (1980). Prolegomena to a theory of mechanized formal reasoning, *Artificial Intelligence* 13 (1980), pp. 133-170