

Formalising Agency-Inducing Patterns in World Dynamics

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 should occur to enable a conceptualisation of a world's process as an agent. Six criteria on patterns in world dynamics are discussed that indicate when the world shows agency, and allows a faithful agent-based conceptualisation. The criteria, that naturally cover aspects such as embodiment and embeddedness, are formalised and their use is illustrated in a case study.

Introduction

To conceptualise processes in the world, sometimes an agent-oriented perspective may be a possibility. 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 every process can just be considered an agent in a faithful 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 occur in the world's dynamics. In other words, there are certain criteria that need to be satisfied by dynamics of the world, in order to induce a form of agency. This paper addresses the question which patterns in the world's dynamics should occur as criteria 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 can 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. Therefore, the issues of embodiment and embeddedness are automatically fulfilled.

In this paper, first, six agency-inducing criteria are identified and discussed informally: partitioned world state ontology, causal isolation, modular world dynamics, input-output dynamics relations, internal-interaction dynamics relations, and representation relations. Next, the formal language meta-TTL is introduced, and it is shown how, using this language, the criteria can be formalised as second-order dynamic properties of the world. After that, a simple case study illustrates the use of the criteria. Finally a discussion is included.

Agency-Inducing 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. Note that these assumptions and patterns 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.

Partitioned World State Ontology 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). This is covered by the assumption that the world state ontology is the union of a collection of sets for areas: internal, external, boundary, input and output.

Causal Isolation In addition to the partitioned world state ontology, the fact that the boundary can only be crossed by specific processes via input and output states is formalised by a pattern in world dynamics called *causal isolation*. This pattern 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.

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.

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.

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.

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). 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 (2005), is adopted for this criterion. 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 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, 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.

Formalising Patterns in World Dynamics

To formalise the patterns in world dynamics given by the above criteria, as a basis the *Temporal Trace Language* (TTL) to express dynamic properties is used; cf. (Jonker and Treur, 2002).

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 and constitute the set $STATPROP(Ont)$.

In order to express dynamics in TTL, 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 $S: BSTATPROP(Ont) \rightarrow \{true, false\}$. The set of all possible states for ontology Ont is denoted by $STATES(Ont)$. Moreover, a fixed *time frame* τ is assumed which is linearly ordered. Then, a *trace* γ over a state ontology Ont and time frame τ is a mapping $\gamma: \tau \rightarrow STATES(Ont)$, i.e., a sequence of states γ_t ($t \in \tau$) in $STATES(Ont)$. The set of all traces over ontology Ont is denoted by $TRACES(Ont)$, i.e., $TRACES(Ont) = STATES(Ont)^\tau$. Finally, a *temporal domain description* W is a given set of traces over the ontology (usually in a given application domain), i.e., $W \subseteq TRACES(Ont)$.

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. Given a trace γ over state ontology Ont , a certain state of the world at time point t is denoted by $state(\gamma, t)$. These states can be related to state properties via the formally defined satisfaction relation \models , comparable to the Holds-predicate in the Situation Calculus. Thus, $state(\gamma, t) \models p$ denotes that state property p holds in trace γ at time t . Likewise, $state(\gamma, t) \not\models p$ 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 \neg , \wedge , \vee , \Rightarrow , and the quantifiers \forall , \exists (e.g., over traces, time and state properties). For example, consider the following dynamic property for a pattern concerning belief creation based on observation:

for trace $\gamma \in W$,
 if at any point in time t_1 the agent A observes that it is wet outside,
 then there exists a time point t_2 after t_1 such that
 at t_2 in the trace the agent A believes that it is wet outside

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

$\forall t: \tau [state(\gamma, t) \models observes(itswet) \Rightarrow \exists t' \geq t \ state(\gamma, t') \models belief(itswet)]$

The set $DYNPROP(Ont, \gamma)$ is the subset of $DYNPROP(Ont)$ consisting of formulae in which γ is either a constant or a variable without being bound by a quantifier.

Past, Future and Interval Patterns

Let two traces γ_1, γ_2 coincide on ontology Ont , and interval $[t1, t2]$, denoted by $\text{coincide_on}(\gamma_1, \gamma_2, \text{Ont}, t1, t2)$ or $\gamma_1 =_{\text{Ont}, [t1, t2]} \gamma_2$ iff

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

When no interval is mentioned it is meant that it holds for the whole time frame. Notice that for $\varphi(\gamma)$ in $\text{DYNPROP}(\text{Ont})$ it holds that $\gamma =_{\text{Ont}} \gamma' \Rightarrow [\varphi(\gamma) \Leftrightarrow \varphi(\gamma')]$. An *interval pattern* for the time interval $[t1, t2]$ is a statement that does not depend on time points before $t1$ or after $t2$. The subset $\text{IPROP}(\text{Ont}, \eta, u1, u2)$ of $\text{DYNPROP}(\text{Ont}, \eta)$ (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

$$\text{interval_statement}(\varphi(\eta, u1, u2), \text{Ont}, \eta, u1, u2) \equiv \forall \gamma_1, \gamma_2, t1, t2 [\gamma_1 =_{\text{Ont}, [t1, t2]} \gamma_2 \Rightarrow [\varphi(\gamma_1, t1, t2) \Leftrightarrow \varphi(\gamma_2, t1, t2)]]$$

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

$$\begin{aligned} \text{past_statement}(\varphi(\eta, u2), \text{Ont}, \eta, u2) &\equiv \forall \gamma_1, \gamma_2, t2 [\gamma_1 =_{\text{Ont}, < t2} \gamma_2 \Rightarrow [\varphi(\gamma_1, t2) \Leftrightarrow \varphi(\gamma_2, t2)]] \\ \text{future_statement}(\varphi(\eta, u1), \text{Ont}, \eta, u1) &\equiv \forall \gamma_1, \gamma_2, t2 [\gamma_1 =_{\text{Ont}, \geq t1} \gamma_2 \Rightarrow [\varphi(\gamma_1, t1) \Leftrightarrow \varphi(\gamma_2, t1)]] \end{aligned}$$

Second-Order Dynamic Properties

The formalisations 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 meta-TTL: the meta-language of TTL. The language meta-TTL includes sorts for $\text{DYNPROP}(\text{Ont})$ and its subsets as indicated above, which contain TTL-statements (for dynamic properties) as 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 meta-TTL statement, quantifiers over TTL statements can be used. In the next sections, the six criteria for agency will be formalised in meta-TTL.

Partitioned World State Ontology

To start with the first criterion (i.e., a partitioned world state ontology), 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 $\text{WorldOnt} = \bigcup \text{LOC} = \bigcup_{L \in \text{LOC}} L$. Based on this, the set of *local basic world state properties* for location L is $\text{BSTATPROP}(L)$, and the set of *local world state properties* is $\text{STATPROP}(L)$. Finally,

$$\begin{aligned} \text{WBSTATPROP} &= \bigcup_{L \in \text{LOC}} \text{BSTATPROP}(L) \\ \text{WSTATPROP} &= \bigcup_{L \in \text{LOC}} \text{STATPROP}(L) \end{aligned}$$

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:

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

According to this, the following ontologies are defined:

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

On this basis also the other sets can be partitioned; e.g., $\text{BSTATPROP}(\text{IntOnt})$, $\text{STATPROP}(\text{IntOnt})$, and $\text{DYNPROP}(\text{IntOnt})$.

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

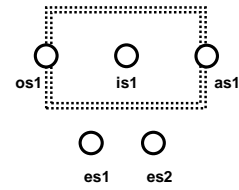


Figure 1. Example world

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:

$$\text{IntOnt} = \{\text{is1}\} \quad \text{ExtOnt} = \{\text{es1}, \text{es2}\} \quad \text{InOnt} = \{\text{os1}\} \quad \text{OutOnt} = \{\text{as1}\}$$

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 $\text{BoundOnt} = \text{InteractionOnt}$ 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.

Causal Isolation

The causal isolation principle expresses that causal influences between internal and external state properties can only occur via the input states and output states. In meta-TTL, this principle can be formalised for causal influences from outside to inside as follows:

$$\text{causal_isolation}(\text{ExtOnt}, \text{InputOnt}, \text{IntOnt}) \equiv$$

$\forall \varphi_1: \text{IPROP}(\text{ExtOnt}, \eta, u_1, u_2)$
 $\forall \gamma: W \forall t_1, t_2: T [\varphi_1(\gamma, t_1, t_2) \wedge t_1 \leq t_2 \Rightarrow \exists t_3, t_4: T [t_2 \leq t_3 \leq t_4 \ \& \ \varphi_3(\gamma, t_3, t_4)]] \Rightarrow$
 $\exists \varphi_2: \text{IPROP}(\text{InputOnt}, \eta, u_1, u_2)$

$\forall \gamma: W \forall t_1, t_2: T [\varphi_1(\gamma, t_1, t_2) \wedge t_1 \leq t_2 \Rightarrow \exists t_3, t_4: T [t_2 \leq t_3 \leq t_4 \ \& \ \varphi_2(\gamma, t_3, t_4)]] \ \&$
 $\forall \gamma: W \forall t_1, t_2: T [\varphi_2(\gamma, t_1, t_2) \wedge t_1 \leq t_2 \Rightarrow \exists t_3, t_4: T [t_2 \leq t_3 \leq t_4 \ \& \ \varphi_3(\gamma, t_3, t_4)]]$

Informally, this criterion states the following:

for all dynamic properties φ_1 referring to only external states,
 and for all dynamic properties φ_3 referring to only internal states,
 if for all traces γ , φ_1 implies later φ_3 ,
 then there is also a dynamic property φ_2 referring to only input states, such that φ_1
 implies later φ_2 and φ_2 implies later φ_3 in all traces.

This definition can be illustrated by considering Figure 2. This picture shows how possible instances of φ_1 , φ_2 and φ_3 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 causal 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: $\text{causal_isolation}(\text{IntOnt}, \text{OutputOnt}, \text{ExtOnt})$.

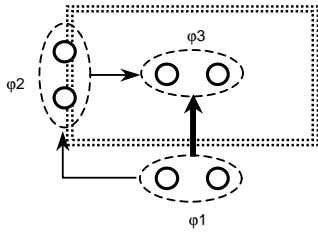


Figure 2. Causal Isolation Principle

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. In meta-TTL, this criterion is formalised as follows:

$\text{modular_world_dynamics} =$
 $\forall \psi: \text{IPROP}(\text{WorldOnt}, \eta, u_1, u_2)$
 $\forall \gamma: W \forall t_1, t_2: T [\psi(\gamma, t_1, t_2) \wedge t_1 \leq t_2 \Rightarrow$
 $\exists \varphi_1: \text{IPROP}(\text{ExtOnt} \cup \text{InteractionOnt}, \eta, u_1, u_2)$
 $\exists \varphi_2: \text{IPROP}(\text{IntOnt} \cup \text{InteractionOnt}, \eta, u_1, u_2)$
 $\varphi_1(\gamma, t_1, t_2) \ \& \ \varphi_2(\gamma, t_1, t_2) \ \&$
 $[\forall \gamma': W [\varphi_1(\gamma', t_1, t_2) \ \& \ \varphi_2(\gamma', t_1, t_2)] \Rightarrow \psi(\gamma, t_1, t_2)]]$

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 φ_1 , referring to only external and interaction states,
 and there is a dynamic property φ_2 , referring to only internal and interaction states,
 such that φ_1 and φ_2 hold for γ , and for all traces γ' , φ_1 and φ_2 together imply ψ .

Also see the (two-dimensional) Figure 3. Again, the three dotted shapes (named ψ , φ_1 , and φ_2) indicate dynamic properties which are built up from the state properties they contain, and arrows denote (temporal) implications between dynamic properties.

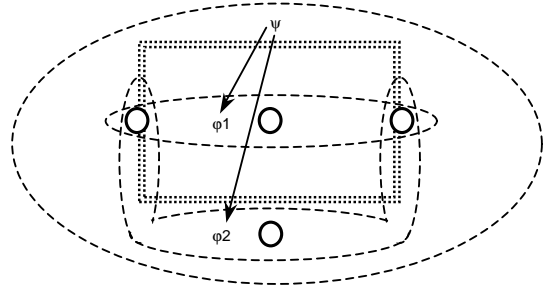


Figure 3. Modular World Dynamics Principle

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 pattern 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:

$\text{IOR} : \text{TRACES}(\text{InOnt}) \times \text{TRACES}(\text{OutOnt})$.

If this relation is functional, i.e., if $\text{IOR}(\gamma_1, \gamma_2)$ and $\text{IOR}(\gamma_1, \gamma_3)$ implies $\gamma_2 = \gamma_3$, then a function IOF exists:

$\text{IOF} : \text{TRACES}(\text{InOnt}) \rightarrow \text{TRACES}(\text{OutOnt})$.

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 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. For the deterministic, functional case, implicit definability is expressed by $\forall \gamma, \gamma': W \gamma =_{\text{InOnt}} \gamma' \Rightarrow \gamma =_{\text{OutOnt}} \gamma'$. For the nonfunctional case it can be expressed as:

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

Explicit definability means: there is a dynamic property expressed in a certain language that relates the input states over time to output states over time. For the functional case this is as follows. For $\varphi(\eta)$ in $\text{DYNPROP}(\text{InteractionOnt})$, let $\text{input_output_correlation}(\varphi(\eta))$ denote

$\forall \gamma: \text{TRACES} \ \varphi(\gamma) \Leftrightarrow \exists \gamma': W [\gamma =_{\text{InOnt}} \gamma' \ \& \ \gamma =_{\text{OutOnt}} \gamma'] \ \&$

$\forall \gamma: W \exists \gamma': \text{TRACES} [\gamma =_{\text{InOnt}} \gamma' \ \& \ \varphi(\gamma')] \ \&$

$\forall \gamma, \gamma': \text{TRACES} [\varphi(\gamma) \ \& \ \varphi(\gamma') \ \& \ \gamma =_{\text{InOnt}} \gamma' \Rightarrow \gamma =_{\text{OutOnt}} \gamma']$.

Then, for the functional case, explicit definability is: $\exists \varphi(\eta): \text{DYNPROP}(\text{InteractionOnt}) \ \text{input_output_correlation}(\varphi(\eta))$, see Figure 4. For the nonfunctional case the third conjunct can be left out.

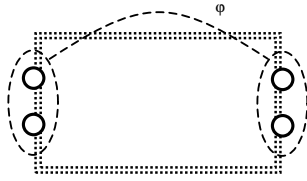


Figure 4. Input-Output Dynamics Relations

Internal and Interaction Dynamics Relations

Given an input-output dynamics relation $\varphi(\eta)$ in $\text{DYNPROP}(\text{InteractionOnt})$, this can be related to the internal dynamics described by $\pi(\eta)$ in $\text{DYNPROP}(\text{IntOnt} \cup \text{InteractionOnt})$ as follows:

$\text{internal_interaction_relation}(\pi(\eta), \varphi(\eta)) \equiv$
 $\forall \gamma:W [\pi(\gamma) \Rightarrow \varphi(\gamma)] \ \& \ \forall \gamma:W [\varphi(\gamma) \Rightarrow \exists \gamma':W [\pi(\gamma') \ \& \ \gamma' =_{\text{InteractionOnt}} \gamma]]$
 Then the criterion is (see also Figure 5):

$\forall \varphi(\eta): \text{DYNPROP}(\text{InteractionOnt}) [\text{input_output_correlation}(\varphi(\eta)) \Rightarrow$
 $\exists \pi(\eta): \text{DYNPROP}(\text{IntOnt} \cup \text{InteractionOnt})$
 $\text{internal_interaction_relation}(\pi(\eta), \varphi(\eta))]$

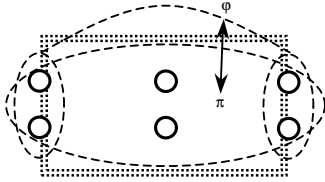


Figure 5. Internal-Interaction Dynamics Relations

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 $\text{DYNPROP}(\text{ExtOnt} \cup \{\text{Ont}(p)\})$ (with $\text{Ont}(p)$ the ontology elements occurring in p) that specifies how a 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 $\text{DYNPROP}(\text{ExtOnt})$ (OR $\text{DYNPROP}(\text{InteractionOnt})$), where the former is a past formula and the latter a future formula. Based on this, the criterion $\text{representation_relation}$ expresses that for all p in $\text{STATPROP}(\text{IntOnt})$ there exist formulae $\varphi_p(\eta, u)$ and $\varphi_F(\eta, u)$ that can be related to p by biconditionals (see also Figure 6):

$\text{representation_relation} \equiv$
 $\forall p: \text{STATPROP}(\text{IntOnt}) \exists \varphi_p(\eta, u), \varphi_F(\eta, u) : \text{DYNPROP}(\text{ExtOnt})$

$\text{past_statement}(\varphi_p(\eta, u), \text{ExtOnt}, \eta, u) \ \&$
 $\text{future_statement}(\varphi_F(\eta, u), \text{ExtOnt}, \eta, u) \ \&$
 $\forall \gamma:W \forall t: T [[\varphi_p(\gamma, t) \Leftrightarrow \text{state}(\gamma, t) \mid = p] \ \& \ [\text{state}(\gamma, t) \mid = p \Leftrightarrow \varphi_F(\gamma, t)]]$

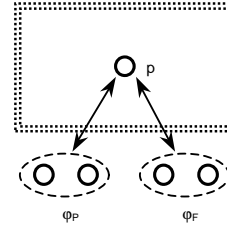


Figure 6. Representation Relation

Case Study

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

Partitioned World State Ontology

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

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

This partitioning satisfies the first criterion. The basic dynamical relationships of the case study are represented graphically in Figure 7.

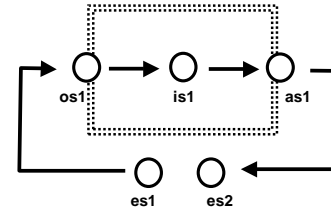


Figure 7. Causal relationships within the case study

Circles denote state properties; the arrows denote causal 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 exist, is1 also ceases to exist. Based on these causal relationships, a number of simulation traces have been produced, using the LEADSTO simulation software (Bosse et al., 2005a). An example of such a trace is shown in Figure 8. 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.

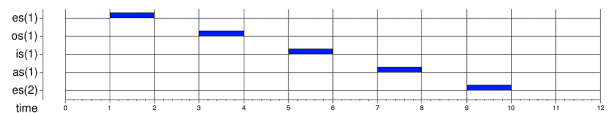


Figure 8. Example world trace

In the following sections, it is explained in detail why the other five criteria for agency hold for these traces.

Causal Isolation

To start, it is illustrated why the property `causal_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 φ_1 and φ_3 in principle is large. Therefore, we restrict ourselves to explaining why the criterion holds for some given instances of φ_1 and φ_3 . Suppose, for example, that the following dynamic properties correspond to φ_1 and φ_3 :

$$\varphi_1(\eta, u_1, u_2) \equiv \text{state}(\eta, u_1) \models \text{es1} \ \& \ \text{state}(\eta, u_2) \not\models \text{es1}$$

$$\varphi_3(\eta, u_1, u_2) \equiv \text{state}(\eta, u_1) \models \text{is1} \ \& \ \text{state}(\eta, u_2) \not\models \text{is1}$$

Then, as $\varphi_1(\gamma:W, t_1:T, t_2:T) \Rightarrow \exists t_3, t_4 \ t_2 \leq t_3 \leq t_4 \ \& \ \varphi_3(\gamma:W, t_3:T, t_4:T)$, according to the causal isolation principle, there is a φ_2 to be found such that, in all traces, φ_1 implies (later) φ_2 and φ_2 implies (later) φ_3 . Such a φ_2 can indeed be found:

$$\varphi_2(\eta, u_1, u_2) \equiv \text{state}(\eta, u_1) \models \text{os1} \ \& \ \text{state}(\eta, u_2) \not\models \text{os1}$$

Given this instance of φ_2 , the property `causal_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 that may be generated on the basis of this model (such as Figure 8), it is clear that if they satisfy φ_1 (i.e., first `es1` holds and later `es1` does not hold) then later φ_3 will hold (i.e., first `is1` holds and later `is1` does not hold), and that they also will satisfy φ_2 in between (i.e., first `os1` holds and later `os1` does not hold). For the set of traces that were generated, this has even been checked automatically, using the TTL checking software described in (Bosse et al, 2005b).

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, suppose

$$\psi(\eta, u_1, u_2) \equiv$$

$$\forall t:T [u_1 \leq t < u_2 \ \& \ \text{state}(\eta, t) \models \text{es1} \Rightarrow \exists t' t < t' \leq u_2 \ \& \ \text{state}(\eta, t') \models \text{is1}] \ \&$$

$$\forall t':T [u_1 < t' \leq u_2 \ \& \ \text{state}(\eta, t') \models \text{is1} \Rightarrow \exists t u_1 \leq t < t' \ \& \ \text{state}(\eta, t) \models \text{es1}]$$

Then, according to the modular world dynamics principle, there are also φ_1 and φ_2 to be found that hold for γ and such that, in all traces, φ_1 and φ_2 together imply ψ . These φ_1 and φ_2 can indeed be found:

$$\varphi_1(\eta, u_1, u_2) \equiv$$

$$\forall t:T [u_1 \leq t < u_2 \ \& \ \text{state}(\eta, t) \models \text{es1} \Rightarrow \exists t' t < t' \leq u_2 \ \& \ \text{state}(\eta, t') \models \text{os1}] \ \&$$

$$\forall t':T [u_1 < t' \leq u_2 \ \& \ \text{state}(\eta, t') \models \text{os1} \Rightarrow \exists t u_1 \leq t < t' \ \& \ \text{state}(\eta, t) \models \text{es1}]$$

$$\varphi_2(\eta, u_1, u_2) \equiv$$

$$\forall t:T [u_1 \leq t < u_2 \ \& \ \text{state}(\eta, t) \models \text{os1} \Rightarrow \exists t' t < t' \leq u_2 \ \& \ \text{state}(\eta, t') \models \text{is1}] \ \&$$

$$\forall t':T [u_1 < t' \leq u_2 \ \& \ \text{state}(\eta, t') \models \text{is1} \Rightarrow \exists t u_1 \leq t < t' \ \& \ \text{state}(\eta, t) \models \text{os1}]$$

Given these instances of φ_1 and φ_2 , the property `modular_world_dynamics` indeed holds for this ψ in this case.

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 :

$$\varphi(\eta) \equiv \forall t:T [\text{state}(\eta, t) \models \text{os1} \Rightarrow \exists t' > t \ \text{state}(\eta, t') \models \text{as1}] \ \&$$

$$\forall t:T [\text{state}(\eta, t) \models \text{as1} \Rightarrow \exists t' < t \ \text{state}(\eta, t') \models \text{os1}]$$

Internal-Interaction Dynamics Relation

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

$$\pi(\eta) \equiv \forall t [\text{state}(\eta, t) \models \text{os1} \Rightarrow \exists t' > t \ \text{state}(\eta, t') \models \text{is1}] \ \&$$

$$\forall t [\text{state}(\eta, t) \models \text{is1} \Rightarrow \exists t' > t \ \text{state}(\eta, t') \models \text{as1}] \ \&$$

$$\forall t [\text{state}(\eta, t) \models \text{is1} \Rightarrow \exists t' < t \ \text{state}(\eta, t') \models \text{os1}] \ \&$$

$$\forall t [\text{state}(\eta, t) \models \text{as1} \Rightarrow \exists t' < t \ \text{state}(\eta, t') \models \text{is1}]$$

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_relation` (see Figure 6). Suppose that ρ corresponds to the state property `is1`. Then, for φ_P and φ_F the following dynamic properties yield correct representation relations:

$$\varphi_P(\eta, u) \equiv \exists t':T [t' < u \ \& \ \text{state}(\eta, t') \models \text{es1}]$$

$$\varphi_F(\eta, u) \equiv \exists t':T [t' > u \ \& \ \text{state}(\eta, t') \models \text{es2}]$$

Discussion

In this paper, the question is addressed which patterns in world dynamics should be present to enable an adequate conceptualisation of a world's process as an agent. Here the world can be a physical world, but also a biological, physiological 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 on patterns in world dynamics were discussed that indicate when the world shows agency, or at least allows a faithful agent-based conceptualisation. As a naturalist perspective is taken, starting from the world's processes, the criteria cover in a direct manner aspects such as embodiment and embeddedness. 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 with 'one face' 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 causal 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 causal 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 of the criteria mentioned (e.g., purely reactive agents), or by specialising some of the criteria or adding criteria (e.g., agents with beliefs, desires and intentions, or 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 at an agent aware of itself). This could easily be added to our 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, internal, 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 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.

Formalisation of the criteria has been done in the form of second-order dynamic properties expressed in the sorted predicate logic-based language meta-TTL. 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, 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
- 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. (2005a). LEADSTO: a Language and Environment for Analysis of Dynamics by SimulaTiOn. In: Eymann, T. et al. (eds.), *Multiagent System Technologies, Proc. of MATES'05*. Lecture Notes in AI, vol. 3550. Springer Verlag, pp. 165-178.
- Bosse, T., Jonker, C.M., Meij, L. van der, and Treur, J. (2005b). A Temporal Trace Language for the Formal Analysis of Dynamic Properties. Technical Report, Vrije Universiteit Amsterdam, Department of Artificial Intelligence.
- Bosse, T., Jonker, C.M., and Treur, J. (2005). Representational Content and the Reciprocal Interplay of Agent and Environment. In: Leite, J., Omicini, A., Torroni, P., and Yolum, P. (eds.), *Declarative Agent Languages and Technologies II, Proc. of DALT'04*. Lecture Notes in Artificial Intelligence, vol. 3476. Springer Verlag, pp. 270-288.
- 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
- 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.
- 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.
- Kim, J. (1996). *Philosophy of Mind*. Westview Press.
- Kim, J. (2005). *Physicalism, or Something Near Enough*. Princeton University Press, Princeton.
- 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.
- 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.
- Treur, J. (2002). Semantic Formalisation of Interactive Reasoning Functionality. *International Journal of Intelligent Systems*, vol. 17, pp. 645-686.