# Part A Chapter 2

# Reactive Behaviour

An agent has *simple reactive behaviour* if it immediately responds to stimuli from its environment. Such agents are also called behaviour-based or situated agents; the pattern of behaviour is called *stimulus-response behaviour*. These agents make their decisions based on information they receive as input, and simple situation-action associations. The stimuli they get as input can either consist of perceived changes in the external world or received communications from other agents. Changes in the external world are perceived by the agent by observation. The response behaviour of the agent affects its environment. Several architectures have been developed for (artificial) reactive software and hardware agents. In (Müller, 1996) an extensive overview of these architectures and the motivations behind them can be found.

Within the first two subsections stimulus-response behaviour is addressed (external view and internal viewpoint, respectively). This illustrates the perspective of *behaviourism*; see, e.g., (Kim, 1996), Ch. 2, pp. 25-46, (Maslin, 2001), Ch 4, pp. 105-129. Another variant of reactive behaviour, delayed response behaviour, is addressed from the external and internal viewpoint in Sections 3 and 4. Here both the perspectives of behaviourism and *functionalism* (e.g., (Kim, 1996), Ch 3-4, pp. 47-124; (Maslin, 2001), Ch 5, pp. 130-161) become manifest.

## 1 External Dynamics Characterizing Stimulus-Response Behaviour

The four simple example traces in Table 1 for situation 1 of Chapter 1 (the situation without cups) illustrate the stimulus-response pattern of behaviour within the experimental setting.

| *time*<br>*trace* | *time point 0* | *time point 1* | *time point 2* | *time point 3* | *time point 4* | *time point 5* |
|---|---|---|---|---|---|---|
| *trace 1* | food at p2<br>screen | no food at p2<br>screen | food at p2<br>screen | food at p2<br>no screen | food at p2<br>no screen<br>goes to p2 | no food at p2<br>no screen |
| *trace 2* | no food at p2<br>screen | no food at p2<br>screen | no food at p2<br>screen | food at p2<br>screen | food at p2<br>no screen | food at p2<br>no screen<br>goes to p2 |
| *trace 3* | no food<br>screen | no food<br>no screen | no food<br>no screen | food<br>no screen | food<br>no screen<br>goes to p2 | no food<br>no screen |
| *trace 4* | food at p2<br>screen | no food at p2<br>no screen | food at p2<br>screen | food at p2<br>screen | food at p2<br>no screen | food at p2<br>no screen<br>goes to p2 |

**Table 1** Example set of observed stimulus-response traces

Here the time points indicated are time points that show a different state, so no uniform time scale is assumed. The observed (animal) agent receives observation input on the availability of food (indicated by food), and of the limitation of its moving around due to the presence or absence of a screen in the experimental setting (indicated by screen). Depending on the circumstances it can decide to eat the food (indicated by the action eat). Assume that the traces depicted in Table 1 are observed. In this table, for example, the fact that at a certain point in time it is observed that there is food at p2 is simply denoted by food at p2, and goes to p2 denotes that the agent performs the action to go to p2.

**Definition (input and output specification, functional input-output association)**
a) An *specification* of *input and output* is an indication of what state properties are considered input and what state properties are considered output of an agent.
b) Stimulus-response behaviour is usually viewed (from an external viewpoint) as a functional association between the agent's input states and output states. This provides for this specific type of behaviour a description of the input-output correlation, as a *functional association between input and output states*, i.e., as a (mathematical) function

$$F : \text{Input\_states} \rightarrow \text{Output\_states}$$

of the set of possible *input states* to the set of possible *output states*.

Note that according to this definition, stimulus-response behaviour is deterministic. Behaviour of this type is transparent and predictable; for the same input always the same behaviour is repeated: it does not depend on earlier processes, nor does it on (not observable) internal states. For example, no information on previous experiences is stored in some form of memory so that it can be remembered and affect behaviour.

If also non-deterministic behaviour is taken into account, the function in the definition above can be replaced by a relation between input states and output states, which relates each input state to a number of alternatives of behaviour. Also probabilistic accounts of stimulus-response behaviour are possible, where for a given input state, for each of the possible behavioural alternatives a probability is specified. For the sake of simplicity, however, in this chapter stimulus-response behaviour is restricted to the functional form defined above.

An external description of a pattern of behaviour as given by an input-output association fits well in the externalist perspective adopted by *behaviourism*; see (Kim, 1996), Ch. 2, pp. 25-46. Within Computer Science, well-known traditional programming methods are based on this paradigm; for example, program specification and refinement based on preconditions and postconditions as developed in, e.g., (Dijkstra, 1976). The following questions are relevant here.

- How can we express basic dynamic properties characterizing stimulus-response behaviour, viewed from an external perspective ?

- In other words, how can we actually specify an input-output correlation for this simple pattern of behaviour ?

To this end, a first property can be expressed informally as follows:

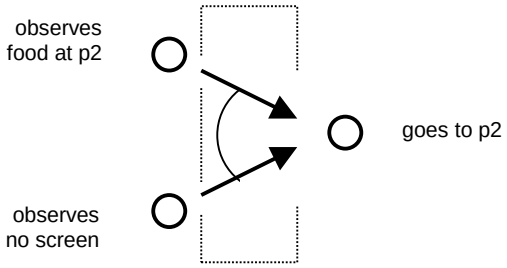
> Every time that the agent observes that there is food at p2, and no screen is present, it will go to p2.

To express such dynamic properties in a more standardized manner, a structured semi-formal temporal language will be used. This language will also turn out useful for more complex patterns of behaviour, as will be seen later in this and subsequent chapters. Using this language, the fact that a given trace shows stimulus-response behaviour, the above informally stated property can be formulated in the form of the following more structured dynamic property, which is expressed in textual form by an immediate temporal relationship (which, in turn, might be based, e.g., on a direct computational or causal relationship) as follows:

**ESR1**

at any point in time,
if the agent observes that food is present at position p2
and it observes that no screen is present,
then it will go to position p2

The naming convention of properties like these is as follows: the first letter refers to whether it is an external or internal property, the rest is a label characterizing the type of behaviour. Notice that this property indeed relates an input states (in which the conditions expressed in the antecedent hold) to output states (in which the consequent holds). In a graphical form such properties are depicted as follows. Here the arrows informally mean 'will lead to', and the arc between two arrows indicates that in combination this is guaranteed to happen. Note that this (graphical or textual) specification by itself does not exclude the action to happen under other circumstances, for example the mouse being pressed to p2 by the experimentator.




Notice that phrases such as 'observes no screen' means that the absence of the screen is observed, so it does not mean that no observation takes place. For example, the (epistemic) difference between 'not observing whether it rains' and 'observing that it does not rain' is

similar to the difference between ‘not knowing whether it rains’ and ‘knowing that it does not rain’.

**Explanation of Stimulus-Response Behaviour from a Behaviourist Perspective**

Given the behaviourist description specified above, the following question concerning explanation can be put forward.

- How can observed behaviour be explained using a behaviourist description ?

To make it more specific, assume it is observed that the animal goes to p2. The type of explanation we consider has the following form:

Why does the animal go to p2 ?
The animal goes to p2 because it observed that food is present at p2, and that the screen is absent.

This raises the following question.

- Which *law* is behind this explanation, or, on the basis of what *general property* can this be concluded ?

The answer on this question is that this explanation is based on the property, given in the description **ESR1**, stating that ‘always if the animal observes food at p2 and that the screen is absent, it will go to p2’. This property may be a special case of a still more general formulation of a law, but we will not discuss this further; the idea remains the same.

The generic format behind the dynamic property **ESR1** is as follows:

at any point in time,
if for the input state X holds
then for the output state Z holds
where X may cover a conjunction of observation results, i.e., of the form
the agent observes that P holds
and it observes that Q does not hold
[ and it observes ……. ]

The dynamic property **ESR1** as defined leaves completely open the response time. An animal going to p2 next day, or next year will also count as an instance of stimulus-response behaviour. A question is what would be a reasonable maximal response time for a response to count as a response to a given stimulus. Probably this maximal response time depends on the case at hand. For an animal in our experimental setting, a maximal response time of 1 to 4 seconds might be reasonable. But in other cases this may be too long or too short (imagine the response time of a snail).

Within the dynamic property above, it is not difficult to include a parameter for the maximal response time taken into account. Thus more refined variants of the dynamic property can be

made, for example, taking into account a maximal response time r; in structured semiformal form this obtains the following parameterised (by r) property:

**ESR2(r)**

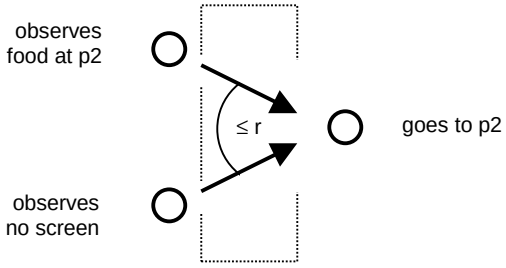
at any point in time,

if the agent observes that food is present at position p2

and it observes that no screen is present,

then within at most r seconds it goes to position p2

In graphical form, this dynamic property can be depicted as follows.




Notice that **ESR2(r)** implies **ESR1**, i.e., **ESR2(r)** is a stronger property. Actually, **ESR1** can be viewed as a form of **ESR2(r)** with maximal response time infinity, and if r1 ≤ r2, then **ESR2(r1)** implies **ESR2(r2)**.

Are these all relevant external properties to be taken into account? Let's consider another set of possible traces; see Table 2.

| *time*<br>*trace* | *time point 0* | *time point 1* | *time point 2* | *time point 3* | *time point 4* | *time point 5* |
|---|---|---|---|---|---|---|
| *trace 1* | food at p2<br>screen | no food at p2<br>screen | food at p2<br>screen | no food at p2<br>no screen | no food at p2<br>no screen<br>goes to p2 | no food at p2<br>no screen |
| *trace 2* | no food at p2<br>screen | no food at p2<br>screen | food at p2<br>screen | no food at p2<br>screen | no food at p2<br>no screen | no food at p2<br>no screen<br>goes to p2 |
| *trace 3* | no food<br>screen | no food<br>no screen | no food<br>no screen | no food<br>no screen | food<br>no screen<br>goes to p2 | no food<br>no screen |
| *trace 4* | food at p2<br>screen | no food at p2<br>no screen | food at p2<br>screen | no food at p2<br>screen | no food at p2<br>no screen | no food at p2<br>no screen<br>goes to p2 |

**Table 2** Another example set of observed traces

The following question can be posed.

- Are these traces instances of stimulus-response behaviour?

Notice that the dynamic properties defined above are satisfied by these behavioural traces. These traces, however, do not satisfy the following dynamic property, defined by:

**ESR3**

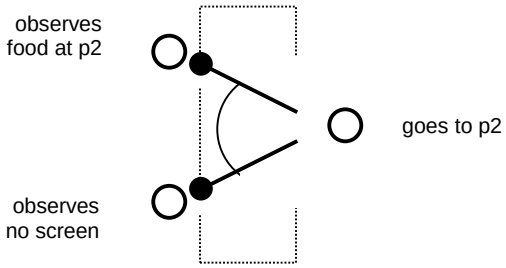
at any point in time,

if the agent goes to position p2

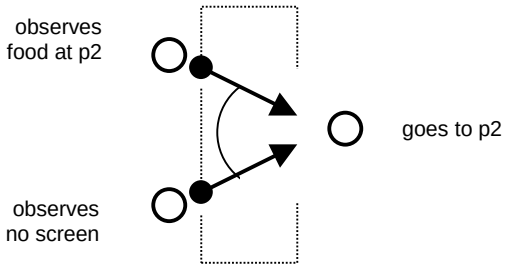
then it was observed that no screen was present,

and that food was present at position p2

In graphical form this can be depicted as follows. Note that, in contrast to the graphical specification of, for example, **ESR1** above, in this specification an arrow has no head but a tail.




The combination of this property **ESR3** with property **ESR1** can be depicted by giving the arrows both a head and a tail, as follows:




It is a matter of choice whether or not this property is assumed to hold for stimulus-response behaviour. As a general property it may be too strong. In particular applications, however, it may be relevant. The following variant of this property takes into account a maximal response time r:

**ESR4**

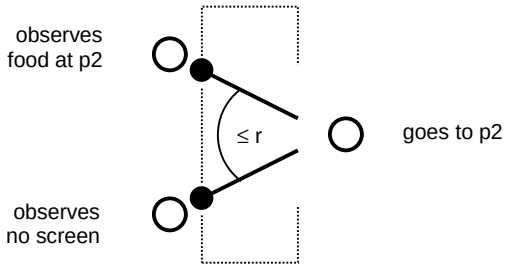
at any point in time,

if the agent goes to position p2

then at most r seconds ago it was observed that no screen was present,

and that food was present at position p2

In graphical form:




## 2 Internal Dynamics Generating Stimulus-Response Behaviour

Although an adequate external description of a pattern of stimulus-response behaviour is possible, this does not exclude to consider the possibility of a description from the internal perspective. For this example, internal mental concepts b1 and b2 (relating to food present and screen absent, respectively) can be assumed with the following dynamic properties:

**ISR1**

at any point in time,
if the agent observes that food is present at position p2
then internal state property b1 will hold

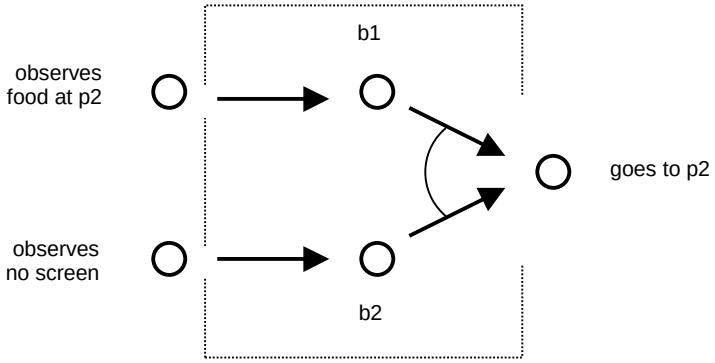
**ISR2**

at any point in time,
if the agent observes that no screen is present,
then internal state property b2 will hold

**ISR3**

at any point in time,
if internal state property b1 holds
and internal state property b2 holds,
then the agent will go to position p2

In graphical form these properties are depicted as follows. Notice that the box indicates the borderline between the internal and externally visible parts.




Notice that for cases that the observations are of a very short duration, for this description of internal dynamics to work, the phrases 'b1 will hold' and 'b2 will hold' either need to be defined by an exact time delay, or they need to persist sufficiently long, so that b1 and b2 occur at a common point in time. Notice that these properties together entail the external property **ESR1**. For completeness' sake this shows what internal dynamics would be possible to generate the external pattern of stimulus-response behaviour; however, this description from an internal perspective does not bring much added value. In fact it is more complex (uses more concepts) than the external description, so it mainly has disadvantages. Actually, Morgan expressed the *principle of parsimony*, stating that:

> 'in no case may we interpret an action as the outcome of the exercise of a higher psychical faculty, if it can be interpreted as the outcome of the exercise of one which stands lower in the psychological scale' (Morgan, 1894).

According to this principle (which can be viewed as a variant of Occam's razor), the external description can be given priority over the internal description: a description of the behaviour without taking into account some internal representation faculty is more parsimonious. However, in this case it may also count that the internal description is not simpler than the external one. If the internal description would be much simpler and more transparent than the external description, it is debatable whether this principle of parsimony should be applied in the absolute sense as expressed by Morgan (actually, another variant of Occam's razor would suggest in such a case not to follow Morgan's principle).

## 3 External Dynamics Characterizing Delayed Response Behaviour

As opposed to behaviour for which an input-output correlation can be defined in the form of a purely functional dependency (association) between input states and output states, as described in Section 1, in less simple cases an agent's behaviour often takes into account previous processes and interactions in which it was involved. Instead of a description as a function of or relation on the set of possible input states to the set of possible output states, in

such more general cases a more appropriate description of an input-output correlation is given in the following definition.

**Definition (input-output correlation and its specification)**

a) A *trace* is mathematically defined as a time-indexed sequence of states, where time points can be expressed, for example, by real or integer values. If these states are input states, such a trace is called an *input trace*. Similarly for an *output trace*. An *interaction trace* is a trace of (combined) states consisting of an input part and an output part.

b) An *input-output correlation* is defined as a binary relation

C : Input_traces x Output_traces

between the set of possible *input traces* and the set of possible *output traces*.

c) A *behavioural specification* S is a *set of dynamic properties* in the form of temporal statements on interaction traces.

A given interaction trace $\mathcal{T}$ *fulfils* or *satisfies* a behavioural specification S if all dynamic properties in S are true for the interaction trace $\mathcal{T}$.

A behavioural specification S is a *specification of an input-output correlation* C if and only if for all interaction traces $\mathcal{T}$ input-output correlation C holds for $\mathcal{T}$ if and only if $\mathcal{T}$ fulfils S.

Notice that this definition takes non-deterministic behaviour into account as well.

In delayed response behaviour, previous observations may have led to internal maintenance of some form of memory of the world state, a model or representation of the (current) world state (for short, *world state model*). This form of memory can be used at any point in time as an additional source (in addition to the direct observations). In that case the same input pattern of stimuli can lead to different behaviour, since the world state models based on observations in the past are different. This makes that agents do not fit strictly in the setting of an input-output correlation based on a direct functional association between (current) input states and output states. This type of behaviour, which just like stimulus-response behaviour occurs quite often in nature, is a bit more complex than stimulus response behaviour. This leads to the question what kind of complexity in the environment is coped with this kind of behaviour that is not coped with by stimulus-response behaviour. An answer on this question can be found in a type of environment with aspects which are important for the animal (e.g., food or predators), and which cannot be completely observed all the time; e.g., food is sometimes covered by other objects.

In this section dynamic properties are identified that characterize the input-output correlation of delayed response behaviour observed from an external viewpoint. Such a dynamic

property has a temporal nature; it can refer to the agent's input and output in the present, the past and/or the future. Within statements expressing dynamic properties from the external viewpoint no commitment is made to internal properties, i.e., no reference is made to internal states.

| *time*<br>*trace* | *time point 0* | *time point 1* | *time point 2* | *time point 3* | *time point 4* | *time point 5* |
|---|---|---|---|---|---|---|
| *trace 1* | food at p2<br>screen | cup at p2<br>screen | cup at p2<br>screen | cup at p2<br>no screen | cup at p2<br>no screen<br>goes to p2 | cup at p2<br>no screen |
| *trace 2* | no food at p2<br>screen | no food at p2<br>screen | cup at p2<br>screen | cup at p2<br>screen | cup at p2<br>no screen | cup at p2<br>no screen |
| *trace 3* | no food<br>screen | no food<br>no screen | no food<br>no screen | no food<br>no screen | food<br>no screen | food<br>no screen<br>goes to p2 |
| *trace 4* | food at p2<br>screen | no food at p2<br>screen | food at p2<br>screen | cup at p2<br>screen | cup at p2<br>no screen | cup at p2<br>no screen<br>goes to p2 |

**Table 3** Example set of observed delayed response traces

In Table 3 some example traces of delayed response behaviour are depicted. Here, only food is depicted when it is observable. The following questions are addressed:

- What is the pattern behind this observed behaviour?
- Which external dynamic properties can be expressed that characterize the pattern behind these traces?
- Which assumed internal state properties generate this externally observed behaviour?
- What is the pattern of dynamics of these internal state properties?
- How can these internal dynamics be characterized by dynamic properties?

The first two of these questions are addressed (from the perspective of behaviourism; cf. (Kim, 1996), pp. 25-46) in this section. The other three questions are addressed in Section 4 in relation to a functionalist perspective; cf. (Kim, 1996), pp. 73-103.

In an informal manner, a dynamic property characterizing the input-output correlation of delayed response behaviour can be expressed as follows.

> Every time that the agent observes that no screen is present,
> and it observed in the past that there was food at p2,
> it will go to p2.

Using slightly more structured semi-formal temporal language, for the example experimental setting, the input-output correlation for delayed response behaviour can be characterised as:

**EDR1**

at any point in time,
if it observes that no screen is present
and at some earlier point in time the agent observed that food was present at position p2,
then the agent will go to position p2

As in this case the temporal relationships are a bit more complex than the previous types of dynamic properties, we will not introduce a graphical format for this property; the same holds for properties **EDR2** to **EDR5** below. In Section 2.4 it will be defined in more detail for which type of dynamic properties the graphical format (the so-called executable properties) can be used. Notice that such a dynamic property is a property of a behavioural trace relating the input part of the trace (characterized by the two conditions in the antecedent) to the output part of the trace (characterized by the consequent), taking into account the temporal relationships expressed.

**Explanation of Delayed Response Behaviour from a Behaviourist Perspective**

Also for this case the question about explanation can be posed.

- How does an explanation look like if this description is available ?

Consider the following example explanation.

Why does the animal go to p2 ?
The animal goes to p2 because it just observed that no screen is present, and in the past it observed that food was present at p2.

This explanation refers (possible very far) back to the past. This past event cannot be seen as a direct cause of the current behaviour. The question then becomes the following.

- What is, besides the current observation of the absence of the screen, a more direct cause than the reference to observation of food in the past ?

This question cannot be answered easily from the perspective of behaviourism. In Section 2.4 we will come back to this question from the perspective of functionalism.

There is at least one course of affairs in which dynamic property **EDR1** provides debatable results as a specification of the input-output correlation of delayed response behaviour. Suppose the animal observes food. After this episode another animal enters the scene and eats the food, which makes the food disappear. All this is observed by our agent. After this observation a cup is put at p2. Then the screen is taken away, and the agent goes to p2. This example trace, and some similar ones are depicted in Table 4.

| *time* / *trace* | *time point 0* | *time point 1* | *time point 2* | *time point 3* | *time point 4* | *time point 5* |
|---|---|---|---|---|---|---|
| *trace 1* | food at p2<br>screen | no food at p2<br>screen | no food at p2<br>screen | cup at p2<br>screen | cup at p2<br>no screen | cup at p2<br>no screen<br>goes to p2 |
| *trace 2* | food at p2<br>screen | no food at p2<br>screen | cup at p2<br>screen | no food at p2<br>screen | no food at p2<br>no screen | no food at p2<br>no screen<br>goes to p2 |
| *trace 3* | no food<br>screen | food<br>screen | no food<br>screen | no food<br>no screen | no food<br>no screen | no food<br>no screen<br>goes to p2 |

**Table 4** Example set of observed traces

Although after these events the agent's most recent observation is that no food is present at p2, according to the property expressed above, still the animal will go to the position p2, because both conditions are fulfilled. The second and third trace in Table 4 are even more challenging. For example, in trace 3, no cup is involved at all. The agent observes that no food is present all the time after time point 2 (and it may even have eaten the food itself, or moved the food to a different position!), but still goes to p2. It may do so for whatever reason, but what is debatable is that the dynamic property expressed above forces the agent to go to p2 in this case. This happens because this formulation of delayed response does not take into account whether or not since the food was observed, the food has disappeared and this absence of food has been observed. It may well be the case that such a pattern of behaviour occurs in nature. Let's call this a pattern of *rigid delayed response behaviour*.

The considerations above make it worth while yet to look for another, more flexible form of delayed response behaviour. The input-output correlation of such a pattern of behaviour can be specified by a dynamic property, which excludes these at least sometimes undesired phenomena. In an informal manner, such a more sophisticated dynamic property of, let's call it *updating delayed response behaviour*, can be expressed as follows.

Every time that the agent observes that no screen is present,
and it observed in the past that there was food at p2,
and it did not observe after this time point that there was no food at p2,
it will go to p2.

Using semi-formal temporal language, this property characterizing the input-output correlation of the pattern of updating delayed response behaviour can be expressed as follows:

**EDR2**

at any point in time t1,
if the agent observes that no screen is present

and at some earlier point in time t the agent observed that food was present at position p2,
and at every point in time t' after t up to t1,
the agent did not observe that no food was present at p2,
then the agent will go to position p2

This property does not force the (debatable) behaviour depicted in Table 4. Notice, however, it still allows this behaviour, it does not exclude the agent to go to p2 in other cases, for whatever reason.

If desired, an additional property to exclude such behaviour can be expressed in a manner similar to the comparable property in Section 1. In semi-formal temporal form, this additional property can be expressed as follows:

**EDR3**

at any point in time t2,
if the agent goes to position p2
then the agent observed at some time point t1 that no screen was present,
and at some point in time t earlier than t1 the agent observed that food was present at position p2,
and at every point in time t' after t up to t1,
the agent did not observe that no food was present at p2.

Just as in the case of stimulus-response behaviour, whether or not this should be considered as a property included in the specification of the input-output correlation of all cases of delayed response behaviour is debatable. But it can be seen as a property that applies in some cases. Other, more refined input-output correlations can be expressed as well (by appropriate properties). First, as in the stimulus-response case in Section 1, it is possible to limit the response time by a bound r after it was observed that no screen is present. Another refinement is to limit the past observation by some bound b. Using these parameters, a more refined property can be expressed as follows in a structured semi-formal manner:

**EDR4(b, r)**

at any point in time t1,
if the agent observes that no screen is present
and at some earlier point in time $t \geq t1-b$ the agent observed that food was present at position p2,

and at every point in time t' after t up to t1,
the agent did not observe that no food was present at p2,
then within at most r seconds it goes to position p2

Using theses parameters, property **EDR3** can be reformulated in parameterised form as follows:

**EDR5(b, r)**

at any point in time t2,
if the agent goes to position p2
then the agent observed at some time point t1 ≥ t2 - r that no screen was present,
and at some point in time t earlier than t1 but not earlier than t1 – b the agent observed
that food was present at position p2,
and at every point in time t' after t up to t1,
the agent did not observe that no food was present at p2.

## 4 Internal Dynamics Generating Delayed Response Behaviour

In Section 3 dynamic properties were established that characterize the input-output correlation of delayed response behaviour from an external, behaviourist viewpoint; these properties refer to the agent's input in the past, in relation to the agent's present output. In this section the internal viewpoint is taken, based on the perspective of *functionalism*; cf. (Kim, 1996), pp.73-103. Assumptions are made about the existence of internal state properties and internal dynamics in such a way that these internal dynamics produce the externally characterized delayed response behaviour. To characterize these internal dynamics, dynamic properties are identified that not only refer to input states and output states over time, but also to *internal states*.

Since the external characterisations of delayed response behaviour refer to the agent's input in the past, it makes sense to assume that the agent makes use of some mechanism to maintain past observations (in the form of a certain kind of *world state model*) by means of persisting internal properties, i.e., some form of memory. These persisting properties are sometimes called *beliefs* on the world state. For the example case, assume that an internal state property b1 is available that defines part of the agent's world state model, with the following dynamics:

**IDR1**

for all time points
if the agent observes that food is present at position p2,
then internal state property b1 will hold

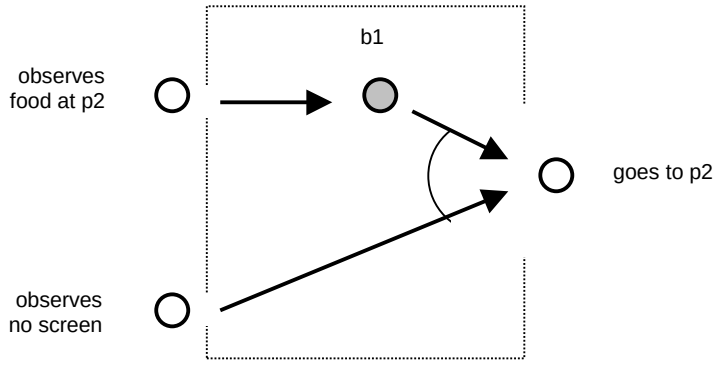
**IDR2**

for all time points
if internal state property b1 holds,
then for every later time point internal state property b1 holds

**IDR3**

at any point in time,
if it observes that no screen is present ,
and internal state property b1 holds,
then the agent will go to position p2

Informally stated, property **IDR1** expresses that any observation of p2 leads to internal property b1. The second property **IDR2** expresses that once b1 holds, it will persist forever. The third property **IDR3** defines that the agent will go to p2 if b1 holds and it observes that no screen is present. In graphical form these dynamic properties can be depicted as follows.




Notice that here a dark node means that this state *persists over time*, i.e., the 'being dark' of node b1 is a graphical way of expressing property **IDR2**. Variants of these properties can be made that take into account a maximal delay time.

These dynamic properties together define how b1 relates to its (immediate) neighbour state properties 'observes food at p2' and 'goes to p2'. Since the relationship of b1 to 'goes to p2' is conditional with condition 'observes no screen', also the latter state property is involved in these relationships. Such relationships of an internal state property b1 to its immediate neighbours defines the *functional role* of b1 as a mediator between the preceeding state properties and succeeding state properties. In the literature on Philosophy of Mind (see, e.g., (Kim, 1996), pp. 74-77), this notion is often illustrated by the internal state property 'having pain'. Having pain is a direct consequence of tissue damage (input) and it leads to output in the form of reactions such as moving away (from the source of the tissue damage). The

internal state property ‘having pain’ mediates between the input and the output: via this internal state property the input ‘tissue damage’ affects the output ‘moving away’. The direct relationships from ‘tissue damage’ to ‘having pain’ and from ‘having pain’ to ‘moving away’ define the functional role of ‘having pain’ as a mediator between this input and output.

A *functionalist* perspective on cognition takes into account internal state properties and characterises them by their functional roles. Using a functionalist perspective, behaviour can be explained by relating it to internal state properties (and their functional roles) preceeding the behaviour. For an example of such a functionalist explanation, see below.

The internal dynamic properties **IDR1**, **IDR2** and **IDR3** together entail the external property **EDR1**. This can be illustrated by the following internal trace.

| *time*<br>*trace 1* | *time point 0* | *time point 1* | *time point 2* | *time point 3* | *time point 4* | *time point 5* |
|---|---|---|---|---|---|---|
| *input* | food at p2<br>screen | food at p2<br>screen | no food at p2<br>screen | cup at p2<br>screen | cup at p2<br>no screen | cup at p2<br>no screen |
| *internal* | | b1 | b1 | b1 | b1 | b1 |
| *output* | | | | | | goes to p2 |

**Explanation of Delayed Response Behaviour from a Functionalist Perspective**

At this point we come back to the explanation of this behaviour. In Section 3 we were left with the situation that an explanation from a behaviourist perspective requires a reference to an observation of food in the past, and thus did not offer a direct cause of the behaviour. So:

- How would an explanation look like from the internal, functionalist perspective ?

The following example explanation illustrates the issue.

Why did the animal go to p2 ?

The animal did go to p2 because it just observed that no screen is present, and it believed that food is present at p2.

Why did it believe that food was present at p2 ?

It believed that food is present at p2 because it already had this belief earlier, and this belief persisted.

When did it start to have this belief ?

It started to have this belief that food is present at p2 when it observed food at p2.

Notice that the properties **IDR1** and **IDR2** could be replaced by the following single property:

**IDR1’**

at any point in time t,

if the agent observes that food was present at position p2,

then a later point in time t’ ≥ t exists such that

the internal state property b1 holds for all time points t" ≥ t’

Also **IDR1'** and **IDR3** together entail **EDR1** (**IDR1'** itself is entailed by **IDR1** and **IDR2**). However, in spite of the lower number of properties (2 instead of 3), the fact that the properties themselves (especially **IDR1'**) have a less simple form, makes that also advantages were lost. For example, it is impossible to visualise this property in the graphical format used thus far; the format would have to be made much more complicated to cover this type of property as well.

The aim is not only to express properties characterizing the internal dynamics in any form, but to express them in such a simple form that, in contrast to the external dynamics, it is immediately clear how these dynamics can be visualised and implemented, for example, as a computational process or as a causal process. In this sense the specification of the internal dynamics as aimed for opens up the possibility of realization of the external dynamics in a formal computational reality, a computer's reality or nature's reality.

In relation to this aim, the advantages of properties **IDR1**, **IDR2**, and **IDR3** derive from the fact that they are in one of the two simple forms defined in the following definition (whereas neither the external property **EDR1** nor **IDR1'** does have this form).

**Definition (executable dynamic properties)**

a) A dynamic property is called *a step property* if and only if it has the form:

if X holds,
then Y will hold

Here X and Y are (conjunctions of) state properties.

b) A dynamic property is called *a (conditional) persistence property* if and only if it has the form:

if X holds
then X will hold at all later time points, as long as not Y holds

Here, as in a), X and Y are (conjunctions of) state properties.

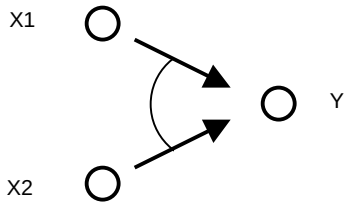
c) A dynamic property is in *executable format* if it is either specified as a step property or as a persistence property.

Note that persistence can also be defined for one step; if such a persistence step is repeated a number of times, the effect is the same as that of a persistence property defined as in b) above. The notion 'executable format' defined in this manner is a rather simple format. Within Computer Science and AI, other, more expressive forms of notions for 'executable' have been defined. Dynamic properties in the simple executable format defined above have three advantages:

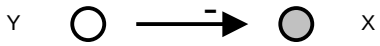
- they can be depicted in a transparent manner using the graphical notation as developed
- they allow for direct simulation or implementation in computational or natural, causal context.
- they allow for explanations based on elementary (e.g., computational or causal) steps

This executable format is particularly useful to specify *functional roles* of internal state properties. In fact, executable dynamic properties define a variant of what by Ashby (1952) is called a *state-determined system*; this is a system for which the properties in subsequent states only depend on properties in the current state, not on the past. Executable dynamic properties provide a representation format to specify such state-determined systems.

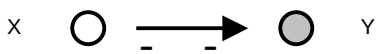
The graphical format used to depict executable properties is as follows. A step property is depicted as:




A (conditional) persistence property is depicted as:




Notice that here the minus sign – above the arrow indicates that Y leads to not X, thereby breaking the persistence of X. This minus sign can also be used to define leads to relations between negations of nodes without having to represent them separately. For example, the step property




expresses that not X leads to not Y.

Within a computational context the step property can be implemented as a simulation step computing the next state out of the current state, or on an abstract (finite state) machine, as a transition step. The persistence property can be implemented by storage in a memory. Within a physical context, a step property can take the form of a causal relation, whereas the persistence property is a basic property of nature (in absence of specific causal effects). That is the reason why dynamic properties in this form are called executable properties. Moreover, the external dynamics is entailed by a set of executable properties, and therefore can be

explained from these executable properties: from the elementary steps and persistencies within the internal dynamics. As these properties can be related to either elementary computation steps or basic causal relations in nature, explanations can be generated for either a computational or natural (physical) context.

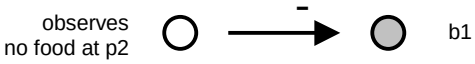
In summary, if the internal dynamics entail the external dynamics, and is described by executable properties, then it provides (a) an implementation possibility, either in a computational context or in a physical context, and (b) an explanation of the externally observable behaviour.

As discussed in Section 1 in relation to the difference between *rigid* delayed response behaviour (external property **EDR1** ) and *updating* delayed response behaviour (external property **EDR2**), internal (persistency) property **IDR2** entails that in case at one point in time the presence of food at p2 is observed, and at a later point in time the absence of food at p2 is observed, then still b1 will persist. For the rigid case, possibly, in addition to b1, another internal property b1' will start to hold after the observation of the absence of food at p2. For the case of udating delayed response behaviour, a slightly more sophisticated option is that, after the absence of food at p2 has been observed, b1 will not be forced to hold anymore. This is expressed by the following dynamic property:

**IDR4**

for all time points t1 and t2 with t1<t2
if internal state property b1 holds at t1,
and between t1 and t2 the agent does not observe that food is not present at position p2,
then internal state property b1 holds at t2

This property can be graphically depicted as follows.




Notice that this property does not exclude b1 to hold at any every future time point, but at least does not entail that b1 holds. If **IDR2** is replaced by **IDR4**, then the properties together entail external property **EDR2**, as is illustrated in the following example traces.

| *time*<br>*trace 1* | *time point 0* | *time point 1* | *time point 2* | *time point 3* | *time point 4* | *time point 5* |
|---|---|---|---|---|---|---|
| *input* | food at p2<br>screen | food at p2<br>screen | food at p2<br>screen | cup at p2<br>screen | cup at p2<br>no screen | cup at p2<br>no screen |
| *internal* | | b1 | b1 | b1 | b1 | b1 |
| *output* | | | | | | go to p2 |

| *time*<br>*trace 2* | *time point 0* | *time point 1* | *time point 2* | *time point 3* | *time point 4* | *time point 5* |
|---|---|---|---|---|---|---|
| *input* | food at p2<br>screen | food at p2<br>screen | no food at p2<br>screen | no food at p2<br>screen | cup at p2<br>no screen | cup at p2<br>no screen |
| *internal* | | b1 | b1 | | | |
| *output* | | | | | | |

## References

Ashby, W.R. (1952). *Design for a Brain*. Wiley and Sons, New York.

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

Maslin, K.T. (2001). *An Introduction to the Philosophy of Mind*. Polity Press/Blackwell Publishers, Cambridge, UK.

Morgan, C.L. (1894). *An introduction to comparative psychology*. London: Scott, 1894.

Müller, J. P. (1996). *The design of intelligent agents: a layered approach*. Lecture Notes in Artificial Intelligence, Vol. 1177, 1996.