Towards a Computational Model of the Self-attribution of Agency

Koen Hindriks[1], Pascal Wiggers[1], Catholijn Jonker[1], and Willem Haselager[2]

[1] Delft University of Technology
[2] Radboud University Nijmegen

Abstract. In this paper, a first step towards a computational model of the self-attribution of agency is presented, based on Wegner's theory of apparent mental causation. A model to compute a *feeling of doing* based on first-order Bayesian network theory is introduced that incorporates the main contributing factors to the formation of such a feeling. The main contribution of this paper is the presentation of a formal and precise model that can be used to further test Wegner's theory against quantitative experimental data.

1 Introduction

The difference between falling and jumping from a cliff is a significant one. Traditionally, this difference is characterized in terms of the contrast between something happening to us and doing something. This contrast, in turn, is cashed out by indicating that the person involved had mental states (desires, motives, reasons, intentions, etc.) that produced the action of jumping, and that such factors were absent or ineffective in the case of falling. Within philosophy, major debates have taken place about a proper identification of the relevant mental states and an accurate portrayal of the relation between these mental states and the ensuing behavior (e.g. [1,2]). In this paper, however, we will focus on a psychological question: how does one decide that oneself is the originator of one's behavior? Where does the feeling of agency come from? Regarding this question we start with the assumption that an agent generates explanatory hypotheses about events in the environment, a.o. regarding physical events, the behavior of others and of him/herself. In line with this assumption, in [3] Wegner has singled out three factors involved in the self-attribution of agency; the principles of priority, consistency and exclusivity. Although his account is detailed, both historically and psychologically, Wegner does not provide a formal model of his theory, nor a computational mechanism. In this paper, we will provide a review of the basic aspects of Wegner's theory, and sketch the outlines of a computational model implementing it, with a focus on the priority principle. Such a model of self-attribution can be usefully applied in interaction design to establish whether a human attributes the effects of the interaction to itself or to a machine.

The paper is organized as follows: Section 2 provides an outline of Wegner's theory and introduces the main contributing factors in the formation of an experience of will. In section 3, it is argued that first-order Bayesian network theory

K.G. Mehrotra et al. (Eds.): IEA/AIE 2011, Part I, LNAI 6703, pp. 295–305, 2011.

is the appropriate modeling tool for modeling the theory of apparent mental causation and a model of this theory is presented. In section 4, the model is instantiated with the parameters of the *I Spy* experiment as performed by Wegner and the results are evaluated. Finally, section 5 concludes the paper.

2 Apparent Mental Causation

Part of a theory of mind is the link between an agent's state and its actions. That is, agents describe, explain and predict actions in terms of underlying mental states that cause the behavior. In particular, human agents perceive their intentions as causes of their behavior. Moreover, intentions to do something that occur prior to the corresponding act are interpreted as reasons for doing the action. This understanding is not fully present yet in very young children.

It is not always clear-cut whether or not an action was caused by ones own prior intentions. For example, when one finds someone else on the line after making a phone call to a friend using voice dialing, various explanations may come to mind. The name may have been pronounced incorrectly making it hard to recognize it for the phone, the phone's speech recognition unit may have mixed up the name somehow, or, alternatively, one may have more or less unconsciously mentioned the name of someone else only recognizing this fact when the person is on the line. The perception of agency thus may vary depending on the perception of one's own mind and the surrounding environment.

In the self-attribution of agency, intentions play a crucial role, but the conscious experience of a feeling that an action was performed by the agent itself still may vary quite extensively. We want to gain a better understanding of the perception of agency, in particular of the attribution of agency to oneself. We believe that the attribution of agency plays an important role in the interaction and the progression of interaction between agents, whether they are human or computer-based agents. As the example of the previous paragraph illustrates, in order to understand human interaction with a computer-based agent it is also important to understand the factors that play a role in human self-attribution of agency. Such factors will enhance our understanding of the level of control that people feel when they find themselves in particular environments. One of our objectives is to build a computational model to address this question which may also be useful in the assessment by a computer-based agent of the level of control of one of its human counterparts in an interaction.

As our starting point for building such a model, we use Wegner's theory of apparent mental causation [4]. Wegner argues that there is more to intentional action than forming an intention to act and performing the act itself. A causal relation between intention and action may not always be present in a specific case, despite the fact that it is perceived as such. This may result in an illusion of control. Vice versa, in other cases, humans that perform an act do not perceive themselves as the author of those acts, resulting in more or less automatic behavior (automatisms). As Wegner shows, the causal link between intention and action cannot be taken for granted.

Wegner interprets the self-attribution of agency as an experience that is generated by an interpretive process that is fundamentally separate from the mechanistic process of real mental causation [3]. He calls this experience the *feeling of doing* or the *experience of will*. The fact that Wegner's theory explains the feeling of doing as the result of an interpretive process is especially interesting for our purposes. It means that this theory introduces the main factors that play a role in interpreting action as caused by the agent itself retrospectively. It thus provides a good starting point for constructing a computational model that is able to correctly attribute agency to a human agent.

Wegner identifies three main factors that contribute to the experience of a feeling of doing: (i) An intention to act should have been formed just before the action was performed. That is, the intention must appear within an appropriately small window of time before the action is actually performed. Wegner calls this the *priority principle*. (ii) The intention to act should be consistent with the action performed. This is called the *consistency principle*. (iii) The intention should exclusively explain the action. There should not be any other prevailing explanations available that would explain the action and discount any intention, if present, as a cause of the action. This is called the *exclusivity principle*.

A crucial factor in assessing the contribution of the priority principle to the feeling of doing is the timing of the occurrence of the intention. In [5] it is experimentally established that the experience of will typically is greatest when the intention is formed about 1 second before the action is performed. As Wegner argues, the priority principle does not necessarily need to be satisfied in order to have a feeling of doing. *People may sometimes claim their acts were willful even if they could only have known what they were doing after the fact* [3]. Presumably, however, an agent that makes up an intention after the fact to explain an event will (falsely) *believe* that it occurred prior to that event.

The contribution of the consistency principle to the experience of will *depends [...] on a cognitive process whereby the thoughts occurring prior to the act are compared to the act as subsequently perceived. When people do what they think they were going to do, there exists consistency between thought and act, and the experience of will is enhanced* [3]. The comparison of thought and action is based on a semantic relation that exists between the content of the thought and the action as perceived. The thought may, for example, name the act, or contain a reference to its execution or outcome. The mechanism that determines the contribution of the consistency principle to a feeling of doing thus relies on a measure of how strongly the thought and action are semantically related. Presumably, the contribution of the consistency principle is dependent on the priority principle. Only thoughts consistent with the act that occurred prior to the perceived act, within a short window of time, contribute to a feeling of doing.

The contribution of the exclusivity principle to the experience of will consists in the weighing of various possible causes that are available as explanations for an action. The principle predicts that when the own thoughts of agents do not appear to be the exclusive cause of their action, they experience less conscious will; and, when other plausible causes are less salient, in turn, they experience

more conscious will [3]. People discount the causal influence of one potential cause if there are others available [6]. Wegner distinguishes between two types of competing causes: (i) internal ones such as: emotions, habits, reflexes, traits, and (ii) external ones such as external agents (people, groups), imagined agents (spirits, etc.), and the agent's environment. In the cognitive process which evaluates self-agency these alternative causes may discount an intention as the cause of action. Presumably, an agent has background knowledge about possible alternative causes that can explain a particular event in order for such discounting to happen. Wegner illustrates this principle by habitual and compulsive behavior like eating a large bag of potato chips. In case we know we do this because of compulsive habits, any intentions to eat the chips are discounted as causes.

3 Computational Model

One of our aims is to provide a computational model in order to validate and explicate Wegner's theory of apparent mental causation. This theory defines the starting point for the computational model. But the theory does not describe the functioning of the affective-cognitive mechanisms that lead to a feeling of doing at the level of detail which is required for achieving this goal. We thus have to make some modeling choices in order to specify *how* a feeling of doing is created. Here a computational model is introduced that provides a tool for simulating the feeling of doing. In the next section the model is instantiated with an experiment performed by Wegner as a means to validate that the model also fits some of the empirical evidence that Wegner presents to support his theory.

It is clear that any model of the theory of apparent mental causation must be able to account for the varying degrees or levels in the experience of a feeling of doing, the variation in timing of intention and action, the match that exists between those, and the competition that may exist between various alternative causes. Neither one of these factors nor the feeling of doing itself can be represented as a two-valued, binary state, since humans can experience more or less control over particular events. As observed in [3], even *our conscious intentions are vague, inchoate, unstudied, or just plain absent. We just don't think consciously in advance about everything we do, although we try to maintain appearances that this is the case.*

Given the considerations above, it seems natural to use a probabilistic approach to model the degrees of priority, and consistency and to weigh the various competing alternative explanations. Moreover, the cognitive process itself that results in an experience of will is an interpretive or inferential process. Given the various inputs relating to time and perceived action, a cause that explains the action is inferred which may or may not induce a feeling of doing. A natural choice to model such dependencies is to use Bayesian networks. Bayesian networks [7] have been used extensively to model causal inference based on probabilistic assessments of various sorts of evidence (see for examples of this in research on a *theory of mind* e.g. [8,9]). Bayesian networks also allow us to use symbolic representations of the thoughts formed and the actions performed by an agent,

which need to be compared in order to compute a feeling of doing in the theory of apparent mental causation.

In this paper, Multi-Entity Bayesian Network (MEBN) Theory is used [10]. MEBN is *a knowledge representation formalism that combines the expressive power of first-order logic with a sound and logically consistent treatment of uncertainty.* An MEBN Theory consists of several MEBN fragments that together define a joint probability distribution over a set of first order logic predicates. Figure 1 shows two MEBN fragments, each depicted as a rounded rectangle, that model the priority principle. A fragment contains a number of nodes that represent random variables. In accordance with the mathematical definition, random variables are seen as functions (predicates) of (ordinary) variables.

The gray nodes in the top section of a fragment are called *context nodes*; they function as a *filter* that constrains the values that the variables in the fragment can take. In contrast to the nodes in the bottom section of a fragment, context nodes do not have an associated probability distribution but are simply evaluated as true or false. Another perspective on these nodes is that they define what the network is about. The context nodes labeled with the $IsA(t, v)$ predicate define the type t of each of the variables v used. In our model, we distinguish intentions, events, opportunities, and time intervals in which the former may occur. Intentions are *mental states* which are to be distinguished from events, which are temporally extended and may change the state of the world. Opportunities are states which enable the performance of an action. In the model, the probabilities associated with each of these nodes should be interpreted as the likelihood that the agent attaches to the occurrence of a particular state, event or other property (e.g. causal relationship) given the available evidence.

Dark nodes in the bottom section of a fragment are called *input nodes* and are references to nodes that are defined in one of the other fragments. In Figure 1, the node in the right fragment labeled $Exists(a, t_a)$ is an input node. To ensure that the model defines a proper probability distribution, a node can be defined in a single fragment only, in which it is said to be *resident*. The node labeled $Exists(a, t_a)$ is resident in the left fragment in Figure 1.

As usual, the links between nodes represent dependencies. Every resident node has a conditional probability table attached that gives a probability for every state of the node given the states of its parent nodes. Prior distributions are attached to resident nodes without parents. Essentially, every fragment defines a parameterized Bayesian network that can be instantiated for all combinations of its variables that satisfy the constraints imposed by its context nodes.

In order to be able to compute a feeling of doing, the prior probability distributions are assumed to be given in this paper. The computational model presented does not explain how explanatory hypotheses about perceived events are generated, nor does it include an account of the perception of these events. Even though the model assumes this information somehow has already been made available, it is setup in such a way that it already anticipates an account for computing at least part of this information. In particular, the mechanism approach of [6] to explain causal attribution has played a guiding role in

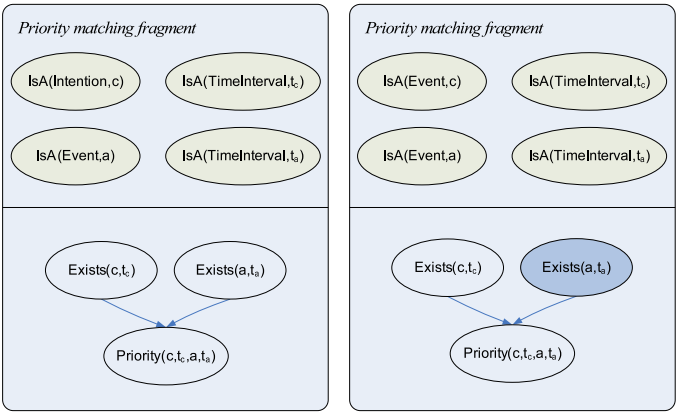

Fig. 1. Priority Fragments

defining the model. The basic idea of this approach is that *causal attribution involves searching for underlying mechanism information (i.e. the processes underlying the relationship between the cause and the effect)*, given evidence made available through perception and introspection. Assuming that each mechanism defines a particular covariation (or joint probability distribution) of the contributing factors with the resulting outcome, the introduction of separate probability distributions for each particular event that is to be explained can be avoided. As a result, the number of priority and causality fragments needed is a function linear in the number of mechanisms instead of the number of events.

3.1 Priority Fragments

The priority principle is implemented by the Priority fragments in Figure 1. Though these fragments are structurally similar, two fragments are introduced in line with the idea that different causal mechanisms may associate different time frames with a cause and its effect. For reasons of space and simplicity, Figure 1 only depicts two fragments, one associated with intentional mechanisms leading to action and a second one for other causal events. The exact time differences depend on the mechanism involved. For example, when moving the steering wheel of a car one expects the car to respond immediately, but a ship will react to steering with some delay.

The *Exists* random variables model that an agent may be uncertain whether a particular state or event has actually taken place at a particular time (also called the *existence condition* in [11]). If there is no uncertainty these nodes will have value true with probability one. The probability associated with the *Priority* random variable is non-zero if the potential cause occurs more or less in the right time frame before the event that is explained by it and the associated probability that the relevant events actually occurred is non-zero. In line with [5], the probability associated with the intentional mechanism increases as the time

difference decreases to about one second. As one typically needs some time to perform an action, the probability starts to decrease again for time intervals less than one second. Each fragment may be instantiated multiple times, illustrated in Section 4, depending on the number of generated explanatory hypotheses.

3.2 Causality Fragments

Figure 2 depicts two fragments corresponding respectively with the intentional mechanism (left) and another type of mechanism (right) that may explain an event. In this case, the fragments are structurally different in two ways. First, even though both fragments require that cause c and effect a are consistent with the mechanism associated with the fragment, the consistency nodes are different. The type of consistency associated with the intentional fragment, called *intentional consistency*, is fundamentally different in nature from that associated with other mechanisms as it is based on the degree of *semantic* relatedness of the content of intention c and the event a (represented as a probability associated with the node). This reflects the fact that one of Wegner's principles, the consistency principle, is particular to intentional explanations. Second, an additional context node representing an opportunity o to act on the intention is included in the fragment corresponding with the intentional mechanism. An intention by itself does not result in action if no opportunity to act is perceived. In line with common sense and philosophical theory [2], the intentional mechanism leads to action given an intention and the right opportunity as input. The model entails that the presence of multiple opportunities increases the probability that a relevant intention is the cause of an event. Additional detail is required to model this relation precisely, but for reasons of space we refer to [12] for a formal model.

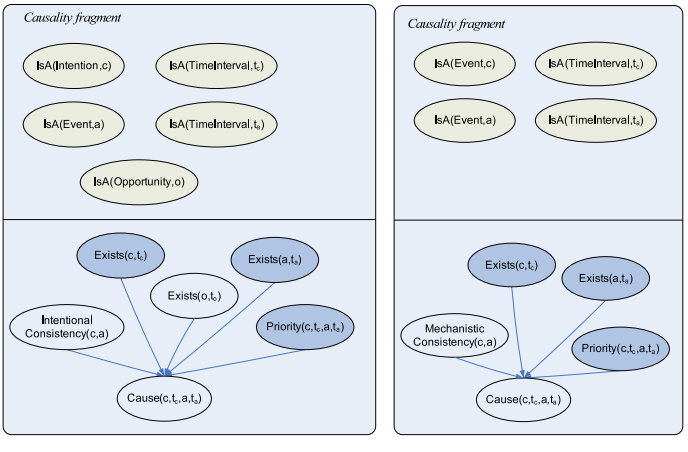

Fig. 2. Causality Fragments

The node labeled $Cause(c, t_c, a, t_a)$ in the intentional fragment models the *feeling of doing*. The associated probability of this node represents the probability that the intention c of an agent has caused event a. In other words, it represents the level of self-attribution of agency for that agent. The probability associated with the node depends on the priority and consistency as well as on the presence (i.e. existence) of both c and a. Obviously, if either c or a is not present, $Cause(c, t_c, a, t_a)$ will be false with probability 1. Additionally, in the intentional fragment an opportunity o must exist.

3.3 Exclusivity Fragment

In order to model the exclusivity principle, an exclusivity fragment is introduced as depicted in Figure 3. In general, if there are multiple plausible causes for an event, exclusivity will be low. Technically, this is modeled as an exclusive-or relation between the competing causes. The value of the random variable *Exclusivity* is set to true to enforce exclusivity. As a result, given two causes of which only one is very likely, the posterior probability of the unlikely cause is reduced. This effect is known as the *discounting effect*, also called *explaining away* [7], and has been studied extensively (e.g. [6]).

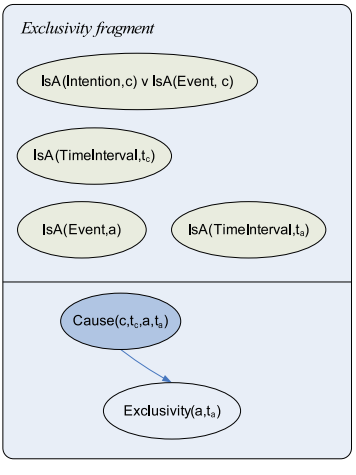

Fig. 3. Exclusivity Fragment

Given an event to be explained and a number of generated explanatory hypotheses (including all contributing factors associated with a particular mechanism), each of the fragments discussed is instantiated accordingly, taking into account the context conditions. To obtain a single, connected Bayesian network, all of the resulting fragments are connected by merging the reference nodes with their resident counterparts. Using this network, the *feeling of doing* can be computed by performing probabilistic inference and querying the $Cause(c, t_c, a, t_a)$

variable in the intentional fragment given the values of the other nodes in the network. By querying other *Cause* variables we can find by means of comparison which of the potential causes is the most plausible one. As a result, only when the node representing the feeling of doing has a high associated probability an agent would explain the occurrence of an event as caused by itself.

4 Simulation of the *I SPY* Experiment

In this section, an instantiation of the model that corresponds with an experiment performed by Wegner is presented. In [5] the results of the *I Spy* experiment are presented that tested whether participants report an experience of agency for something that is most likely the result of someone else's action. In the experiment two participants are seated on opposite sides of a table. On the table a square board that is attached to a computer mouse is located and both participants are asked to put their fingertips on the board and to move the mouse by means of the board in slow sweeping circles. By doing so, a cursor is moved over a computer screen showing a photo from the book *I Spy* [13], hence the name of the experiment, picturing about 50 small objects. The subjects had to move the mouse for about 30 seconds after which they would hear a 10 second clip of music through headphones and within this period they had to stop moving the mouse and then rate on a continuous scale whether they allowed the stop to happen or intended the stop to happen. In addition to the music, subjects would occasionally hear words over the headphones. Participants were told that they would hear different bits of music and different words. One of the participants however did not hear music at all, but was a confederate who received instructions from the experimenter to stop on a particular picture or to let the other participant determine the stop. The forced stops were timed to occur at specific intervals from when the participant heard a corresponding word that was intended to prime a thought about items on the screen. By varying timing, priority was manipulated. For unforced stops the words heard by the participant corresponded about half of the time to an object on the screen.

It turned out that in initial experiments in which the confederate did not force stops the mean distance between stops and the pictures that were primed by words was not significantly different from the mean distance in trials in which the prime word did not refer to an object on the screen. These initial experiments were performed to confirm that participants would not stop the cursor on an object simply because of hearing the word. In consecutive experiments, however, where the stops were forced by the confederator, participants tended to perceive the stops as more or less intended, dependent on the time interval between the hearing of the prime word and the actual stop. If the word occurred between 5 and 1 seconds before the stop, a significant increase in self-attribution was observed.

Based on the description of the *I Spy* experiment and the results presented in [5], an instantiation of the computational model has been derived. Due to space limitations we cannot provide all details.

The resulting model gives the same results as those reported in [5]: If the a priori probability associated with the *Priority* variables is higher (corresponding

to the time interval between 5 to 1 seconds), then a significantly higher feeling of doing is produced than otherwise. The second column of Table 1 shows the posterior probability of the $Cause(I_p, t_p, S, t_s)$ node that models the feeling of doing for several a priori probabilities of the *Priority* variable. For a probability of 0.85 for priority the probability of *Cause* corresponds to the feeling of doing for a time difference of about 1 second as described in [5]. Similarily, the values obtained with a probability for priority of 0.8 and 0.35 correspond to the feeling of doing reported in [5] for respectively 5 seconds and 30 seconds time difference between the prime word and the stop of the cursor.

In [5], also the variance in feeling of doing observed in the experiment is reported. One would expect that a person's personality influences his feeling of doing. Various people, for example, might be more or less sensitive to priming or might have a strong or weak tendency to claim agency in a setup such as in the *I Spy* experiment. We tested the model with different values of priority with a moderated a priori probability for the existence of intention of 0.45 and with a high a priori probability of 0.65 for the existence of an intention. The corresponding posterior probabilities of the cause node are shown in Table 1. These probabilities adequately correspond with the variance reported by Wegner, which gives some additional support for the proposed computational model.

Table 1. Posterior probability of $Cause(I_p, t_p, S, t_s)$ for different a priori probabilities of $Priority(I_p, t_p, S, t_s)$ and $Exists(I_p, t_p)$

	$P(Exists(I_p, t_p))$		
$P(Priority)$	0.55	0.45	0.65
0.3	0.41	0.36	0.45
0.35	0.44	0.39	0.48
0.5	0.51	0.46	0.56
0.8	0.62	0.56	0.66
0.85	0.63	0.58	0.67

5 Conclusion and Future Work

In this paper, a first step towards a computational model of the self-attribution of agency is presented, based on Wegner's theory of apparent mental causation [3]. A model to compute a *feeling of doing* based on first-order Bayesian network theory is introduced that incorporates the main contributing factors (according to Wegner's theory) to the formation of such a feeling. The main contribution of this paper is the presentation of a formal and precise model that provides detailed predictions with respect to the self-attribution of agency and that can be used to further test such predictions against other quantitative experimental data. An additional benefit of the model is that given empirical, quantitative data parameters of the network can be learned, using an algorithm as in [10].

A number of choices had to be made in order to obtain a computational model of Wegner's theory of apparent mental causation. Not all of these choices

are explicitly supported by Wegner's theory. In particular, it has been hard to obtain quantitative values to define the probability distributions in our model. The report of the *I Spy* experiment in [5] does detailed information, but did not provide sufficient information to construct the probability distributions we need. Certain values had to be guessed in order to obtain outcomes corresponding with the results in [5]. The only validation of these guesses we could perform was to verify whether variation of some of the input values of our model could be said to reasonably correspond with the reported variations in the experiment in [5]. It is clear that more work needs to be done to validate the model. In future work, we want to design and conduct actual experiments to validate and/or refine the model of self-attribution.

To conclude, we want to remark that there are interesting relations here with other work. As is argued in [9], Bayesian networks are not sufficient as cognitive models of how humans infer causes. These networks are very efficient for computing causes, but are themselves instantiations from more general, higher-level theories. In a sense, this is also the case in our model since both the consistency fragment as well as the causality fragment in our first-order Bayesian theory of apparent mental causation need to be instantiated by other domain-specific theories in order to derive the right semantic relations between thoughts and actions, and to identify potential other causes of events. Additional work has to fill in these gaps in the model, starting from e.g. ideas presented in [6,9].

References

1. Anscombe, G.E.M.: Intention. Harvard University Press, Cambridge (1958)
2. Dretske, F.: Explaining behavior. MIT Press, Cambridge (1988)
3. Wegner, D.M.: The Illusion of Conscious Will. MIT Press, Cambridge (2002)
4. Wegner, D.M.: The mind's best trick: How we experience conscious will. Trends in Cognitive Science 7, 65–69 (2003)
5. Wegner, D.M., Wheatley, T.: Apparent mental causation: Sources of the experience of will. American Psychologist 54 (1999)
6. Ahn, W.K., Bailenson, J.: Causal attribution as a search for underlying mechanisms. Cognitive Psychology 31, 82–123 (1996)
7. Pearl, J.: Probabilistic Reasoning in Intelligent Systems - Networks of Plausible Inference. Morgan Kaufmann Publishers, Inc., San Francisco (1988)
8. Gopnik, A., Schulz, L.: Mechanisms of theory formation in young children. Trends in Cognitive Science 8, 371–377 (2004)
9. Tenenbaum, J., Griffiths, T., Niyogi, S.: Intuitive Theories as Grammars for Causal Inference. In: Gopnik, A., Schulz, L. (eds.) Causal Learning: Psychology, Philosophy, and Computation. Oxford University Press, Oxford (in press)
10. Laskey, K.B.: MEBN: A Logic for Open-World Probabilistic Reasoning. Technical Report C4I-06-01, George Mason University Department of Systems Engineering and Operations Research (2006)
11. Kim, J.: Supervenience and Mind. Cambridge University Press, Cambridge (1993)
12. Jonker, C., Treur, J., Wijngaards, W.: Temporal modelling of intentional dynamics. In: ICCS, pp. 344–349 (2001)
13. Marzollo, J., Wick, W.: I Spy. Scholastic, New York (1992)