A Developmental Bayesian Model of Trust in Artificial Cognitive Systems

Massimiliano Patacchiola
Centre for Robotics and Neural Systems
Plymouth University
PL4 8AA Plymouth, United Kingdom
Email: massimiliano.patacchiola@plymouth.ac.uk

Angelo Cangelosi
Centre for Robotics and Neural Systems
Plymouth University
PL4 8AA Plymouth, United Kingdom
Email: a.cangelosi@plymouth.ac.uk

***Abstract*—In human-robot interaction trust is one of the main factors to take into account for enabling effective interaction. Limited models exist that delineate the development of trust in real world scenarios. Reshaping one of these models we show how a probabilistic framework based on Bayesian Networks (BNs) can incorporate the reliability of information sources into the decisional process of artificial systems. Furthermore, using a developmental approach we gain some insight on how children estimate people's reliability and how some aspects of the Theory of Mind (ToM) can affect that estimation. To test the model we reproduced a developmental experiment in a computational simulation and we embedded the BNs inside an artificial agent. The simulation results are in line with the real data, and confirm that BNs have the potential for being included as trust evaluator modules in robotic systems.**

I. Introduction

Including a model of trust in a robotic system can improve its interaction with humans in many ways. The robot's characteristics, and in particular performance factors, have a major influence on perceived trust in Human-Robot Interaction (HRI) [1]. Coupling a model of trust with an internal performance estimator would enable the robot to predict how much a person has been trusting him. Alternatively, the robot could use the model to evaluate the human's performance and discern reliable users from unreliable ones. This second use would be particularly significant in robot care for the elderly, where the reliability of the person can fluctuate due to mental illness. By having an estimate of the person's reliability the system can decide the degree of assistance to provide, and in case of anomalous activity it can call an operator for further assistance. Moreover the robot can give different weight to simultaneous commands received from different users, preferring the ones that have been reliable in the past. Although our approach can be applied in both scenarios, we focused on the latter, because the objective is to integrate the model in artificial cognitive systems, and to join it with recent findings in humanoid robots credibility [2], [3].

Our work is based on a well-defined framework called developmental (or epigenetic) robotics [4]–[7]. The aims of developmental robotics are on the one hand to study the cognitive processes in babies by testing theories on robotic platforms, on the other hand to embed the underlying principles of human cognition in artificial agents. We will show how this approach can be helpful and how it can shed light on some developmental mechanisms involved in trust building. The model of trust we are introducing can be used as a stand-alone module as we did in our simulation, but we prefer to consider it as part of a wider framework. In fact our next step will be to embed it in a humanoid robot and to make it interact with multiple users in a real environment. This premise is fundamental and must be kept in mind in order to understand the final objective we are facing.

A. Defining Trust

Trust has been defined as reliance on or confidence in the dependability of someone or something [8]. To expand this generic definition we can take into account also the temporal dimension because for building reliance and confidence we need to consider past interactions. This is especially true when we have to integrate discordant information from different sources; in this case we tend to prefer the source that has been reliable in the past. From the developmental point of view there are many studies that confirm how by four years of age children are able to track the reliability of informants preferring the most accurate source [9], [10]. However a lack of cognitive skills may interfere with the process of source selection generating some errors in inference. In the context of trust one of the most important skills is the ability to read others' belief, known as Theory of Mind (ToM) [11]. Several studies focused on the link between trust and ToM finding that an immature ToM may cause errors when estimating the informant's reliability [12]–[14]. We followed this body of research reshaping an existing model of ToM [15] into a new one that integrates ToM and trust into a unified scheme. Using Bayesian Networks (BNs) we showed that in some particular conditions children use a cause-effect strategy to learn and predict future events. Our work is not focused in explaining how ToM is acquired nor do we want to explain the underlying mechanism that impairs it, instead we aim to incorporate aspects of ToM inside a model of trust to analyse how it affects decision making and belief estimation.

B. Related Work

Because we used the framework of developmental robotics we wanted to define some constraints to our literature review,

cutting off a line of research involving the so called e-trust. E-trust occurs in digital contexts among artificial agents of a distributed system and does not concern studying the human mental processes. We decided to take into account only those analyses that endorse a psychological and developmental approach. In the psychological literature there are works that investigate causal inference in adults, showing how people are sensitive to the accuracy, certainty and self-knowledge of others [16], [17]. These models are good descriptors of adult behaviour but they have not been tested on children. Some models of epistemic trust in children's reasoning have been produced in the past years [18]. However such models do not seem to discriminate children with mature and immature ToM, whereas research consistently seems to indicate a close relationship with trust [12]–[14]. As far as we know there are no probabilistic models that link ToM and selective trust. Because of this uniqueness our work can be considered of particular relevance not only in HRI but also in developmental psychology.

The model presented in this article is based on [15]. In [15] a Random Markov Field was used in order to show how a probabilistic approach can model some aspects of the ToM. This work is particularly relevant because it also shows some interesting applications in a multi-robot scenario and in a gaze following task. To show the potential of the Bayesian approach we applied a modified version of [15] to trust estimation. To achieve this goal we had to reshape [15]. Markov Random Fields are undirected graphs meaning that they do not incorporate the cause-effect principle, since instead of conditional probabilities they use joint probabilities to define relations between nodes. In our work we decided to use BNs instead of Markov Random Fields because we think that they better describe the mental processes of the child. We base our statement on consistent research which is part of the theory-theory approach to the ToM [19]. The theory-theory used causal probabilistic models to formalise the child reasoning in mathematical way. Because we want to use our model in different conditions we decided to keep it constrained to a tested developmental framework. This choice makes our model more robust and links it to a long series of high quality experiments [20]–[24].

The major issue we had to overcome in order to use the approach discussed in [15] is due to an assumption made by the authors. The model is explicitly based on the hypothesis that agents do not deceive each other. As a consequence there is no distinction between inferred state and actual state of others. Such an assumption is reasonable when the agents act cooperatively, but it is not possible when it is necessary to estimate trust. Implicit in the idea of trust there is the possibility that the agents deceive and that their mental states are different from their actions. The authors assert that people are unable to access a human's actual mental state. This is only partially true, because given a record of past actions we can infer others' mental state. For example, if a person was unreliable in past situations we can use this record to infer her behaviour in a similar situation. In our model we took into account the mental state of the others' and we introduced an inference step for estimating their mental states. Such an inference was ignored in [15] and was in fact impossible due to the limitation implicit in the model design. Due to all these considerations our revision of [15] should be considered a substantial improvement.

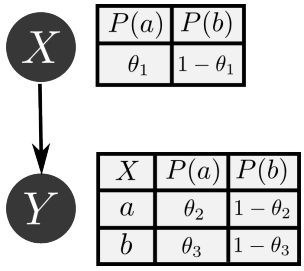

Fig. 1. A typical Bayesian network with two nodes, showing both the topology and the conditional probability tables. The unknown $\theta_1, \theta_2, \theta_3$ are the values to be estimated.

Another novelty we are going to introduce in this work is the use of Maximum Likelihood Estimation (MLE) for adjusting the network parameters. We compared the MLE to the familiarisation phase during which the child acquires information about the informants. As we showed in the next section, this method can be used to collect data and it can use these data to update the BN parameters.

II. Trust and Bayesian Networks

BNs are probabilistic graphical models that represent the conditional dependencies between a set of random variables. Our analysis will be limited to the case of finite and countable values. In this case the random variables are called discrete, and the associated probability distribution is called probability mass function. We used capital letters to indicate random variables and small letters to indicate the possible state or event of the variables. For example, the random variable X is discrete and it can assume four states $\{a, b, c, d\}$. When the outcome of a random variable depends on the outcome of another one, we have a conditional dependence. If the random variable Y depends on X we can calculate the probability of Y given a particular outcome of X which is denoted by $P(Y|X)$. When the two random variables are conditionally independent $P(Y|X) = P(Y)$ are also indicated with $Y \perp X$.

In a BN we represent random variables as nodes and conditional dependencies as edges. The conditional dependence incorporates a principle of causality. If X causes Y we have a BN with a direct edge from X to Y like in Figure 1.

The probabilities associated with the conditional dependence between X and Y in a BN are represented with a conditional distribution, and in the case of discrete random variables can be described with a table, called conditional probability table. In Figure 1 the conditional probability tables associated with X and Y are reported near each node. Each row of a conditional probability table sum to one, and contains the conditional probability of each value for a conditional case. A conditional case is a combination of values for the parent

nodes. The first column of the table associated to Y in Figure 1 represents the conditional values a and b for the parent X.

In our model we denoted X as representing the beliefs and Y as representing actions. Furthermore X_C denotes the belief associated to the child where X_R and X_U denote the belief associated to the reliable and unreliable informant. We used the same subscripts for naming the variables Y.

In our simulations all the random variables were discrete. In particular we used boolean random variables that assumed two states a and b. The states can be considered possible labels for an object, like in [12], or locations, like in [14].

The core of the model is the distinction between belief and actions whereas X represents beliefs and Y represents actions. An edge from X to Y indicates that the action Y is a direct consequence of her belief X like in Figure 1. This differentiation is very powerful because it can describe reliable and unreliable informants. When an informant is deceiving there is a difference between belief and action, and therefore the posterior distribution of X is significantly different from the one of Y.

A. Inference

The main advantage of using BN is the possibility to infer the posterior probability of the nodes given some observations. In particular we used inference for estimating the belief given an action, and vice versa. For example, if the belief is $X = b$ we can calculate the posterior probability of the action $Y = a$ given $X = b$. In the simplified case of Figure 1 the posterior value is contained in the probability table associated with the node Y. The inference process can be applied also from the effect to the cause. For example, if we observe the action $Y = b$ we can find the posterior probability of $X = a$ by means of the Bayes' Theorem:

$$P(X = a|Y = b) = \frac{P(Y = b|X = a)P(X = a)}{P(Y = b)} \quad (1)$$

However, computing the posterior distribution for complex networks can present some difficulties. When the network has only one root and each node has only one parent the network is called a tree. In this case the Pearl's message passing algorithm [25] computes the exact posterior distribution for each node. In case of multiple roots the network is called a poly-tree and another version of the Pearl's message passing algorithm must be used [26]. Networks with complex connections that cannot be classified as tree or poly-tree are called multi-connected. Exact inference is almost impossible to obtain with these networks, and approximation methods are generally used. The model described in this work is a poly-tree and exact inference methods can be used without problems.

B. Learning the Network Parameters

BNs are commonly used for creating expert systems, software that permits emulation of decision-making ability at the level of human experts on a tightly delineated problem [27]. In this context the network is built by hand with the help of domain experts. When the amount of knowledge required is huge it is possible to use a set of data instead of the experts for setting the network parameters, an approach known in literature as parameter estimation or parameter learning [28].

To estimate the network parameters we used the Maximum Likelihood Estimation (MLE). The MLE can take a dataset to adjust the BN parameters. In our case the dataset represents statistical information collected by the child during the interaction with the informants. The use of this technique for online learning in a developmental context seems to be an innovation because as far as we know there are no other cases of its use in such a context. The implicit hypothesis in our approach is that children can collect statistical information for tracking the reliability of others, and that this process is similar to how MLE sets the parameters of a BN. This is in line with recent research [29] which showed that young children can use statistical information, particularly a violation of random sampling, to infer the preferences of an adult for certain type of toys. This result was observed also in 20-month-old infants, confirming that it is an early developmental mechanism.

To illustrate the insight behind MLE we describe a simplified interaction where a child is observing a caregiver choosing two boxes called a and b. The child wants to predict the behaviour of the caregiver in a future trial. Observing the caregiver for a certain period of time the child can collect a dataset, specifically a set of outcomes where each element can be a or b. We denote N_a and N_b as the number of times the caregiver chooses a and b. To describe this toy example in the BN language we need only one discrete node Y representing the action of the caregiver. Defining a parameter θ we can formalise the probabilities that the caregiver chooses the two boxes this way:

$$P_Y(a) = \theta$$
$$P_Y(b) = 1 - \theta$$

Knowing the parameter θ allows predicting the behaviour of the caregiver in a future trial. Evaluating different hypotheses and choosing the one that better predicts the data, the MLE can find an estimation of the parameter θ using the equation:

$$\hat{\theta} = \frac{N_a}{N_a + N_b} \quad (2)$$

During the learning phase the MLE adds to its internal counters a value for each observed event, eventually reaching N_a and N_b. When the dataset is small enough that some events can have not yet been observed a problem can arise. The MLE assigns zero probability to those events, negatively affecting inference. To solve this issue it is often recommended to initialise the count of each event to one instead of zero [30]. This solution does not perturb the posterior distributions in any significant way, we then decided to use it in our simulation given the small size of our dataset. All the parameters of the networks have been set during the learning phase without any external intervention.

The MLE can also be used in more complex scenarios. Let's suppose that the child wants to predict both actions and

beliefs of the caregiver. To describe this new example in term of BN we need the node X (representing beliefs) in addition to the node Y, like in Figure 1. Because the node Y has a parent, its conditional probability table has four entries. We call θ_2 and θ_3 the parameters associated to Y, and θ_1 the parameter associated to X. In this scenario the child has to collect a new set of information to effectively predict the belief of the caregiver. When the caregiver performs an action, the child has to estimate the related belief and incorporate it for future prediction. In our model the MLE can estimate $\theta_1, \theta_2, \theta_3$ counting the number of events and using Equation 2. The posterior distributions are then estimated through inference.

III. Simulation

To test our model we reproduced a developmental experiment [14] in a simulated environment. The approach in [14] is particularly relevant because it directly investigated the correlation between trust and ToM. The experiment concerned children with mature and immature ToM who had to deal with two kinds of informant: helpers and trickers. The children assisted at two different scenes in which an adult indicated to a protagonist the location of a sticker hidden inside one of two boxes. The helper always revealed the correct location of the sticker, whereas the tricker always gave wrong advice. After observing the scene it was the child's turn to be the finder. The helper and the tricker gave conflicting advice to the child who had to guess where the sticker was. The Theory of Mind Scale [31] was used in order to estimate the maturity of the child's ToM. The Theory of Mind Scale is a five-item scale containing tasks that measure the developmental progression of children's mental state understanding. The tasks ask children to reason about situations in which a protagonist has different preferences than their own. Five factors are investigate: diverse-desire, diverse-beliefs, knowledge-ignorance, false belief, false emotions. In order to investigate the relation between trust and ToM further, the authors used also metacognitive questions. The children observed a new pointer helping two finders and another pointer tricking two finders. After observing each new pointer, the children answered four forced-choice questions each one investigating a different factor: intention judgment, same-context prediction, trait judgment, different-context prediction. For more details about the Tehory of Mind Scale and the metacognitive questions we refer the reader to [31] and [14].

A. Methods

To reproduce [14] we created a simulated environment with two different artificial agents. The first agent represented children with mature ToM and the second agent represented children with immature ToM. Because in the original experiment helpers and trickers act in separate contexts we embedded two separate BNs into the agents' cognitive systems. The first BN modelled the interaction between the agent and the helper, whereas the second BN modelled the interaction between the agent and the tricker. As we said this is consistent with the original experiment which the same child never received suggestions from both informants at the same time. A graphical illustration of the two BNs can be observed in Figure 2. The two nodes X_C and Y_C are beliefs and actions of the agent. The posterior distribution of the node Y_C allows the agent to choose one action among all the possible outcomes. In our case there are only two possible actions: choose box a and choose box b. For example, if $P_{Y_C}(a) = 0.8$ and $P_{Y_C}(b) = 0.2$ the agent will choose the box a because the associated action has a higher probability. The connections between Y_U and Y_C, and between Y_R and Y_C represent the influence that the opinions of the informants have on the agent's action. The action of the agent is then a consequence of its own belief X_C and the informant action Y_R or Y_U. Because we were dealing with two categories of agents (Mature ToM VS Immature ToM) and two kinds of pointer (Helper VS Tricker) we had a total of four BNs with four related datasets: Mature ToM and Reliable Pointer, Mature ToM and Unreliable Pointer, Immature ToM and Reliable Pointer, Immature ToM and Unreliable Pointer. Following the procedure in the original experiment we split the simulation into three parts: familiarization, decision making, and belief estimation.

1) Familiarisation: The familiarisation consisted of learning the parameters of each BN using the MLE and the associated dataset. The Bayesian approach makes it possible to make inferences with limited data. Each dataset consisted of only six trials, where each trial represented an interaction between the child and the informant. The number of trials is exactly the same as in the original experiment. In [14] there was a session where the child watched the helper interacting with two finders for a total of six trials. In a second session the child watched the tricker interacting with two different finders, also in this case for six trials. As for the original experiment we counterbalanced the box selection, 50% of the time the helper suggested the box a and 50% the box b. The suggestions were always correct revealing the correct position of the stickers. In the other session a tricker always gave wrong suggestions. The dataset associated with the tricker contained six entries, with each box recommended half of the time. The recommendations were always incorrect. During the familiarisation phase the information acquired by the agent with immature ToM differed substantially from the information acquired by the agent with mature ToM. The agent with immature ToM associated the action Y_U to the wrong belief X_U, whereas the agent with mature ToM identified the deception and associated to Y_U the real belief X_U. Because of this deficit in reading the informants' intention the agent with immature ToM collected wrong statistical data and the inference was then distorted in subsequent phases. This bias is what divides a mature ToM from an immature ToM. It is well documented in the literature but the exact mechanism behind it is still not well understood. We integrate the bias in our model as part of the familiarisation phase but it is behind our scope to explain its origin. We refer the reader to [32] for an exhaustive meta-analysis of the phenomenon.

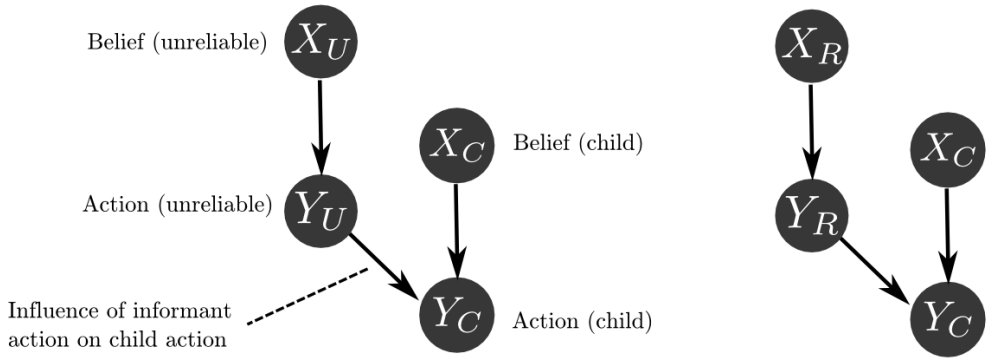

Fig. 2. The BNs integrated in the agent's cognitive system. The network on the left represents the relation between the child and the unreliable informant (tricker), the network on the right represent the relation between the child and the reliable informant (helper). Because the two networks have the same structure and the same number of nodes, we used the subscript U (unreliable) and R (reliable) to distinguish the informants' belief and action. The conditional probability tables have not been reported due to space constraints.

2) Decision Making: In the decision making phase the agent had to choose one of two boxes given the informant's opinion. To test this condition we set as evidence the informants' action node $P_{Y_R}(a) = 1.0$ and $P_{Y_U}(a) = 1.0$. Such a configuration corresponds to an informant indicating the box a as the one containing the sticker. In a similar way we could set the network to model a situation in which the informant indicates the box b but for the sake of clarity we omitted that result in our tables. After this preliminary phase the message passing algorithm [26] was used to compute the posterior distributions for each node.

3) Belief Estimation: The last phase of the simulation was the belief estimation. In [14] the children were asked some metacognitive questions in order to investigate their perception of the two informants. These questions were used to determine whether the children could correctly identify the tricker as unreliable and the helper as reliable. To test the network on this task we set the action and belief nodes as evidence: $P_{Y_C}(a) = 1.0$ and $P_{X_C}(a) = 1.0$. Such a configuration represents the agent inferring actions and beliefs of the informants given evidence that the sticker is in the box a. Like in the previous phase after this preliminary step the message passing algorithm was used to compute the posterior distributions of the network.

B. Results

The final results in [14] showed that only the children with mature ToM distinguished between helpers and trickers. The children's score on the finding task was positively correlated with their performance on the Theory of Mind Scale, $r(87) = .339$, $p = .001$. This result seems to confirm the fact that children's reasoning about whom to trust is strongly correlated with their understanding of mental life. Other research is consistent with this conclusion [33], [34]. Our simulation was coherent with these results. The agent representing children with mature ToM recognized helpers and trickers. When the informant was a helper the agent accepted the suggestion, when the informant was a tricker the agent rejected the suggestion. The agent representing children with immature ToM did not predict the behaviour of the tricker, as observed in the experiment.

To understand the results we obtained we can directly examine the output of each node in the BNs after the inference phase. We decided to report four tables which illustrate the posterior distributions of each BN. In the following sections we are going to discuss these tables. The use of a probabilistic approach made it possible to have a closer look at the internal mechanics of the model, and we wanted to take advantage of this in the discussion of the results. By studying the posterior distributions we obtain a clear understanding of the decision taken by the agents in our simulation.

1) Mature ToM: Table I illustrates the output of the BNs for the agent with mature ToM. In the decision making task the reliable pointer indicated the box a when the sticker was in that box. The suggestion was accepted by the agent as demonstrated by the inequality $P_{Y_C}(a) > P_{Y_C}(b)$. When the informant was unreliable the same suggestion was rejected. The network output showed the rejection in the form $P_{Y_C}(a) < P_{Y_C}(b)$, which means that the agent selected the box b. To test the belief estimation we assumed the agent knew which box the sticker was in and it had to guess the informants' belief. In our query to the BN the stickers were inside the box a. After the computation of the posterior distributions we observed the inequalities $P_{Y_R}(a) > P_{Y_R}(b)$ and $P_{X_R}(a) > P_{X_R}(b)$, which showed that the model successfully anticipated the actions and the belief of the reliable pointer. A query similar to the previous one was sent to the BN that modelled the interaction with the unreliable informant. The computed posterior distribution was $P_{Y_U}(a) < P_{Y_U}(b)$ and $P_{X_U}(a) > P_{X_U}(b)$. These inequalities imply that the agent identified a discrepancy between belief and action, and predicted the malevolent intentions of the tricker.

2) Immature ToM: Table II illustrates the output of the BNs for the agent with an immature ToM. In the decision

TABLE I

AGENT WITH MATURE TOM. THE TABLE ON THE LEFT REPRESENTS THE OUTPUTS OF THE BN AFTER THE INTERACTIONS WITH THE RELIABLE POINTER (HELPER), THE TABLE ON THE RIGHT REPRESENT THE OUTPUTS AFTER THE INTERACTIONS WITH THE UNRELIABLE POINTER (TRICKER). THE ROWS OF THE TABLES REPRESENT THE POSTERIOR PROBABILITY DISTRIBUTIONS ASSOCIATED WITH EACH NODE OF THE NETWORKS, FOR BOTH DECISION MAKING (DM) AND BELIEF ESTIMATION (BE) TASKS.

	DM		BE	
	a	b	a	b
X_C	0.5	0.5	1.0	0.0
Y_C	0.65	0.35	1.0	0.0
X_R	0.8	0.2	0.57	0.43
Y_R	1.0	0.0	0.62	0.38

Helper

	DM		BE	
	a	b	a	b
X_C	0.5	0.5	1.0	0.0
Y_C	0.35	0.65	1.0	0.0
X_U	0.2	0.8	0.57	0.43
Y_U	1.0	0.0	0.38	0.62

Tricker

TABLE II

AGENT WITH IMMATURE TOM. THE TABLE ON THE LEFT REPRESENTS THE OUTPUTS OF THE BN AFTER THE INTERACTIONS WITH THE RELIABLE POINTER (HELPER), THE TABLE ON THE RIGHT REPRESENT THE OUTPUTS AFTER THE INTERACTIONS WITH THE UNRELIABLE POINTER (TRICKER). THE ROWS OF THE TABLES REPRESENT THE POSTERIOR PROBABILITY DISTRIBUTIONS ASSOCIATED WITH EACH NODE OF THE NETWORKS, FOR BOTH DECISION MAKING (DM) AND BELIEF ESTIMATION (BE) TASKS.

	DM		BE	
	a	b	a	b
X_C	0.5	0.5	1.0	0.0
Y_C	0.65	0.35	1.0	0.0
X_R	0.8	0.2	0.57	0.43
Y_R	1.0	0.0	0.62	0.38

Helper

	DM		BE	
	a	b	a	b
X_C	0.5	0.5	1.0	0.0
Y_C	0.65	0.35	1.0	0.0
X_U	0.8	0.2	0.57	0.43
Y_U	1.0	0.0	0.62	0.38

Tricker

making task the reliable pointer indicated the box a when the sticker was in the box a. The suggestion was accepted as demonstrated by the inequality $P_{Y_C}(a) > P_{Y_C}(b)$. The same suggestion was given by the unreliable informant but the sticker in that case was in the box b. Because of the lack of ToM the agent could not recognise the deception and accepted the advice. We observed the result in the posterior distribution $P_{Y_C}(a) > P_{Y_C}(b)$. In the belief estimation task the model produced the right posterior distributions for the helper but the wrong distributions for the tricker: $P_{Y_U}(a) > P_{Y_U}(b)$ and $P_{X_U}(a) > P_{X_U}(b)$. These inequalities show that the agent cannot predict the malevolent intent of the tricker returning a wrong distribution for the action node Y_U.

IV. CONCLUSION

In this article we wanted to introduce a module for estimating trust in HRI. Starting from an existing probabilistic model of ToM [15] we successfully integrated trust as a main component of the model. Following a developmental approach we used BNs as a probabilistic approach for integrating trust and ToM into a common scheme. The MLE was used to set the network parameters. We illustrated how the MLE can be considered a mathematical extension of the process that allows the children to collect their own statistical information from the others'. To verify the reliability of the model we reproduced a developmental experiment [14]. In this experiment children with mature and immature ToM selected the advice of reliable and unreliable informants regarding the location of some stickers. The results showed that children with mature ToM could identify the unreliable informant, whereas children with immature ToM could not. In our simulated environment we reproduced those results.

It is important to point out that the model we propose can be extended to more complex situations. For example, similarly to [15] it can take into account the contemporary influence of two informants, and consider the weigh of each opinion based on past reliability. Thanks to the flexibility of BNs it is straightforward to reorganise nodes and edges for representing this new scenario. When two informants give advices at the same time, a single network is sufficient to integrate them in the decisional process of the child. In this network the action nodes of the two informants Y_U and Y_R have a direct connection to the child node Y_C. The posterior distribution of Y_C is influenced by the internal beliefs and the two external actions. Using the same approach it is possible to integrate more than two sources. Moreover the model is not limited to situations where the tricker cheats every time, it can integrate partially reliable and unreliable users. The posterior distribution $P_1(a)$ of an informant who always told the truth, will be higher than the distribution $P_2(a)$ of a second informant that sometimes cheated. In this case we have that $P_1(a) > P_2(a)$ with $P_1(a) > P_1(b)$ and $P_2(a) > P_2(b)$, meaning that the first informant is more reliable than the second and that both are indicating the right box.

Finally, given the growing presence of autonomous systems

in our society it is necessary to implement a module that permits the estimate of reliability. The model presented in this work can be considered a good candidate for such a module because it allows an artificial cognitive system to estimate others' belief. The simulation presented here must be considered as a premise. The next step of our research will be to integrate the module in a humanoid robot, in order to estimate the reliability of different users in some daily activity.

ACKNOWLEDGMENT

This material is based upon work supported by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF under Award No. FA9550-15-1-0025.

REFERENCES

[1] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction." *Human Factors*, vol. 53, pp. 517–527, 2011.

[2] I. Gaudiello, E. Zibetti, S. Lefort, M. Chetouani, and S. Ivaldi, "Trust as indicator of robot functional and social acceptance. an experimental study on user conformation to icub answers," *Computers in Human Behavior*, vol. 61, pp. 633–655, 2016.

[3] D. Zanatto, M. Patacchiola, J. Goslin, and A. Cangelosi, "Priming anthropomorphism: Can the credibility of humanlike robots be transferred to non-humanlike robots?" in *Proceedings of the Eleventh Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts.*, Christchurch, New Zealand, 2016, pp. –.

[4] M. Asada, K. F. MacDorman, H. Ishiguro, and Y. Kuniyoshi, "Cognitive developmental robotics as a new paradigm for the design of humanoid robots," *Robotics and Autonomous Systems*, vol. 37, pp. 185–193, 2001.

[5] M. Asada, K. Hosoda, Y. Kuniyoshi, H. Ishiguro, T. Inui, Y. Yoshikawa, M. Ogino, and C. Yoshida, "Cognitive developmental robotics: A survey," *IEEE Transactions on Autonomous Mental Development*, vol. 1, pp. 12–34, 2009.

[6] A. Cangelosi and M. Schlesinger, *Developmental Robotics, From Babies to Robots*. Cambridge, MA: MIT Press, 2015.

[7] M. Lungarella, G. Metta, R. Pfeifer, and G. Sandini, "Developmental robotics: a survey," *Connection Science*, vol. 15, pp. 151–190, 2003.

[8] G. R. VandenBos, *APA Dictionary of Psychology*. Washington, DC: Maple Press, 2015.

[9] M. Fusaro, K. H. Corriveau, and P. L. Harris, "The good, the strong, and the accurate: Preschoolers' evaluations of informant attributes," *Journal of Experimental Child Psychology*, vol. 110, pp. 561–574, 2011.

[10] V. K. Jaswal and L. A. Neely, "Adults don't always know best: Preschoolers use past reliability over age when learning new words," *Psychological Science*, vol. 17, pp. 757–758, 2006.

[11] D. G. Premack and G. Woodruff, "Does the chimpanzee have a theory of mind?" *Behavioral and Brain Sciences*, vol. 1, pp. 515–526, 1978.

[12] M. A. Koenig and P. L. Harris, "Preschoolers mistrust ignorant and inaccurate speakers," *Child Development*, vol. 76, pp. 1261–1277, 2005.

[13] L. J. Moses and D. A. Baldwin, "What can the study of cognitive development reveal about children's ability to appreciate and cope with advertising?" *Journal of Public Policy and Marketing*, vol. 24, pp. 186–201, 2005.

[14] K. E. Vanderbilt, D. Liu, and G. D. Heyman, "The development of distrust," *Child Development*, vol. 82, pp. 1372–1380, 2011.

[15] J. Butterfield, O. C. Jenkins, D. M. Sobel, and J. Schwertfeger, "Modeling aspects of theory of mind with markov random fields," *International Journal of Social Robotics*, vol. 1, pp. 41–51, 2009.

[16] D. Buchsbaum, S. Bridgers, A. Whalen, E. Seiver, T. L. Griffiths, and A. Gopnik, "Do i know that you know what you know? modeling testimony in causal inference," in *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, 2012.

[17] A. R. Landrum, B. S. E. Jr, and P. Shafto, "Learning to trust and trusting to learn: a theoretical framework," *Trends in Cognitive Sciences*, vol. 19, pp. 109–111, 2015.

[18] P. Shafto, B. Eaves, D. J. Navarro, and A. Perfors, "Epistemic trust: Modeling children's reasoning about others' knowledge and intent," *Developmental Science*, vol. 15, pp. 436–447, 2012.

[19] A. Gopnik and A. N. Meltzoff, *Words, thoughts, and theories*. Cambridge, MA: MIT Press, 1997.

[20] A. Gopnik, D. M. Sobel, L. E. Schulz, and C. Glymour, "Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation," *Developmental Psychology*, vol. 37, pp. 620–629, 2001.

[21] A. Gopnik, C. Glymour, D. M. Sobel, L. E. Schulz, and T. Kushnir, "A theory of causal learning in children: Causal maps and bayes nets," *Psychological Review*, vol. 111, pp. 3–32, 2004.

[22] A. Gopnik and L. E. Schulz, "Mechanisms of theory formation in young children," *TRENDS in Cognitive Sciences*, vol. 8, pp. 371–377, 2004.

[23] N. D. Goodman, C. L. Baker, E. B. Bonawitz, V. K. Mansinghka, A. Gopnik, H. Wellman, and J. B. Tenenbaum, "Intuitive theories of mind: A rational approach to false belief," in *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Society*. Vancouver, Canada: R. Sun, 2006, pp. 1382–1387.

[24] D. M. Sobel, J. B. Tenenbaum, and A. Gopnik, "Childrens causal inferences from indirect evidence: Backwards blocking and bayesian reasoning in preschoolers," *Cognitive Science*, vol. 28, pp. 303–333, 2004.

[25] J. Pearl, "Reversed bayes on inference engines: A distributed hierarchical approach," in *Proceedings of the Second National Conference on Artificial Intelligence.* Menlo Park, California: AAAI Press, 1982, pp. 133–136.

[26] J. H. Kim and J. Pearl, "A computational model for combined causal and diagnostic reasoning in inference systems," in *Proceedings of the IJCAI-83.*, Karlsruhe, Germany, 1983, pp. 190–193.

[27] D. J. Spiegelhalter, A. P. Dawid, S. L. Lauritzen, and R. G. Cowell, "Bayesian analysis in expert systems," *Statistical Science*, vol. 8, pp. 219–247, 1993.

[28] D. Koeller and N. Friedman, *Probabilistic Graphical Models*. Cambridge, MA: MIT Press, 2009.

[29] T. Kushnir, F. Xu, and H. M. Wellman, "Young children use statistical sampling to infer the preferences of other people," *Psychological Science*, vol. 21, pp. 1134–1140, 2010.

[30] S. Russell and P. Norvig, *Artificial Intelligence a Modern Approach*, 3rd ed. Pearson.

[31] H. M. Wellman and D. Liu, "Scaling of theory-of-mind tasks," *Child Development*, vol. 75, pp. 523–541, 2004.

[32] H. M. Wellman, D. Cross, and J. Watson, "Meta-analysis of theory-of-mind development: The truth about false belief," *Child Development*, vol. 72, pp. 655–684, 2001.

[33] K. Lee, C. A. Cameron, J. Doucette, and V. Talwar, "Phantoms and fabrications: Young childrens detection of implausible lies," *Child Development*, vol. 73, pp. 1688–1702, 2002.

[34] C. DiYanni and D. Kelemen, "Using a bad tool with good intention: Young children's imitation of adults' questionable choices," *Journal of Experimental Child Psychology*, vol. 101, pp. 241–261, 2008.