A computational model for social learning mechanisms*

Manuel Lopes[†], Francisco S. Melo[‡†], Ben Kenward[♭], José Santos-Victos[†]

[†]Institute for Systems and Robotics, Instituto Superior Técnico, Lisboa, Portugal
[‡]School of Computer Science, Carnegie Mellon University, Pittsburgh, USA
[♭]Department of Psychology, Uppsala University, Uppsala, Sweden

Abstract. In this paper we propose a computational model for learning from demonstration. By adequate adjustment of a few parameters, our model is able to produce different learning behaviours, taking into account different elements of the demonstration. In particular, our model takes into consideration the *actions* of the demonstrator, its *effects* on the environment/surroundings, the demonstrator's *inferred goals*, and the interests and preferences of the learner itself. We present results where we show that our model can reproduce (in simulation) several well-known results from standard experimental paradigms in developmental psychology and also an application to a real robotic imitation learning task.

1 Introduction

In social learning, a learner makes use of information provided by an *expert* to improve its learning or acquire new skills. For example, an individual that observes the actions of a second individual can bias its exploration of the environment, improve its knowledge of the world or even reproduce parts of the observed behaviour. Two such social learning mechanisms have raised particular interest among the research community, these being *imitation* and *emulation* [1]. Imitation describes novel action acquisition arising from adhering to the goal of the demonstration as inferred by the learner, while replicating the observed actions and effects on the environment. In contrast with imitation, *emulation* is focused on the reproduction of the observed effects on the environment.

In [2], social learning is described as arising from considering three fundamental sources of information (see Fig. 1): the *actions* observed during a demonstration, the *effects* caused by such actions and the *goals* that led to such actions. Experiments in children and in chimpanzees also showed that they can understand the purpose of an action [3,4,5], even when the action fails [6,7,4]; the dynamics of the world [8]; and the restrictions of the demonstrator [9]. In [10], several learning mechanisms are described in terms of *social influence* and *social learning*, making them distinct from what is called "imitative behaviour". In this work we propose a new computational model for imitation-like behaviours. Our model is inspired by the taxonomy depicted in Fig. 1 and is able to reproduce several social learning behaviours such as imitation or emulation. Many models from the robotics community focused on *goal-directed imitation* are closer to *emulation* than it is to actual *imitation*, as no inference on the intentions of the demonstrator takes place during the learning process. Other models, that rely on the replication of some observed effect in the environment without taking into account the corresponding actions, can be found in [11,12,13,14]. More recent works have provided models in which a learner does infer the goal of the demonstrator and adopts this goal as

* This work was supported in part by the the EU projects RobotCub (IST-004370) and Contact (EU-FP6-NEST-5010) and also by FCT Programa Operacional Sociedade de Informação (POSC) in the frame of QCA III, the Carnegie Mellon-Portugal Program and the project PTDC/EEA-ACR/70174/2006.

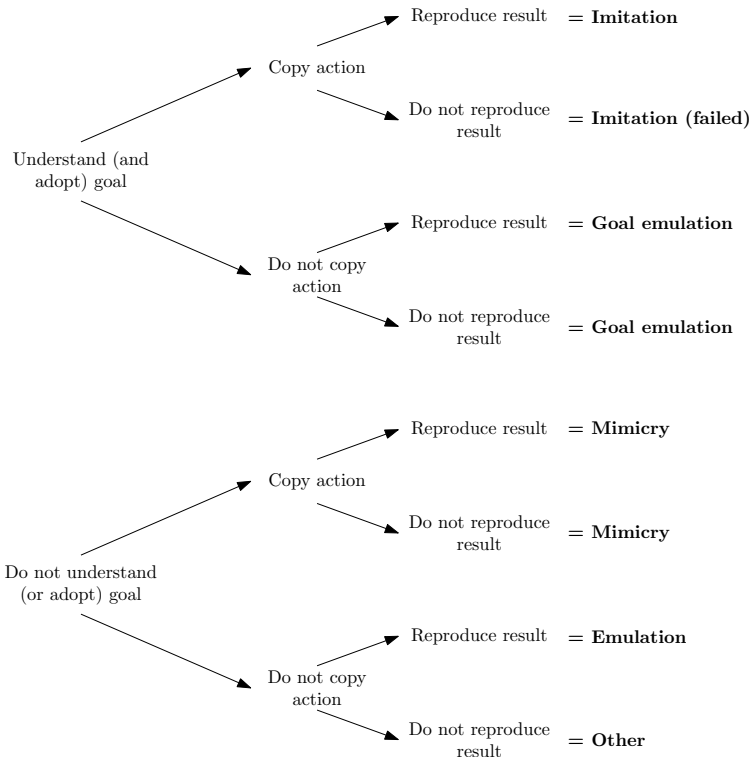


Fig. 1: Behaviour classification in terms of goal, actions and effects (reproduced from Call and Carpenter [2].

its own [15,16,17,18]. This class of behaviour is closer to imitation as defined in [2]. Finally, several works contrast with those referred above in that they are able to generate multiple social-learning behaviours [19,20,21].

In this paper we propose a general model that is able to integrate different sources of information from a demonstration, exhibiting different classes of *imitative behaviour*. Our model replicates the behaviour observed in several well-known experiments [22,9,8].

The paper is organized as follows. In Section 2 we introduce our new learning model and proceed in 3 by presenting the (simulated) results obtained with out model in well-known scenarios from the developmental biology literature. We also present the results obtained in a robotic imitation learning task. Finally, we conclude on Section 4.

2 Social learning algorithm

In this section we describe the fundamental process by which the learner perceives the task to be learned after observing the demonstration by another agent (*e.g.*, a human). In our computational model, we make use of different sources of information to compute a utility function that the learner will then use to determine its behaviour. We show that, by assigning weighting in different ways these sources of information, the learner can exhibit fundamentally different behaviours. In the continuation, we formalize how each of these sources can be explored separately by going over the general approach and then providing the full details.

– The first source of information is the learner's *preference between actions* in terms of the respective energetic costs. This preference translates the natural inclinations of the learner

and can be formalized as a *preference relation over the action repertoire*, represented by the corresponding utility function Q_A.

- The second source of information corresponds to the desire of the learner to replicate the *effects* observed in the demonstration. For example, the learner may wish to reproduce the change in the surroundings observed during the demonstration, or to replicate some particular transition experienced by the teacher. This can be translated in terms of a utility function by assigning a positive value to the desired effect and propagating that value to all states and actions. The utility function thus obtained will be denoted as Q_E.
- The third source of information is related to the desire of the learner to pursue the same *goal* as the teacher. This means that the learner makes some inference about the underlying intention of the teacher. Inferring this intention from the demonstration is achieved by a teleological argument [23]: the goal of the demonstrator is perceived as the one that more rationally explains its actions. Note that the goal cannot be reduced to the final effect only, since the means to reach this end effect may also be part of the demonstrator's goal. We denote the corresponding utility function by Q_G.

Given the three sources of information, formalized in terms of the utility functions Q_A, Q_E and Q_G, the learner will then adhere to the decision-rule obtained by combining the three functions. In particular, the learner will adhere to the decision-rule associated with the function

$$Q^* = \lambda_A Q_A + \lambda_E Q_E + \lambda_G Q_G, \tag{1}$$

with $\lambda_A + \lambda_E + \lambda_G = 1$. By resorting to a convex combination as in Eq. 1, there is an implicit tradeoff between the different sources of information. As a simple example, by considering $\lambda_A = \lambda_E = 0$ and $\lambda_G = 1$, the learner will focus on pursuing the same goal as the teacher, disregarding both its natural interests and any desire to reproduce any particular effect observed in the demonstration.

It is only to be expected that the use of different values for the parameters λ_A, λ_E and λ_G will lead to different behaviours from the learner. This is actually so, as illustrated by the results in our experiments. We also emphasize that Q_E greatly depends on the world model of the *learner* while Q_G greatly depends on the world model of the *teacher*.[1]

Formalism Now we proceed with the details about the underlying model. At each time instant, the learner must choose an action from its repertoire of action primitives $\mathcal{A}$, depending on the state of the environment. We represent the state of the environment at time t by X_t and let $\mathcal{X}$ be the (finite) set of possible environment states. This state evolves according to the transition probabilities

$$\mathbf{P}\left[X_{t+1} = y \mid X_t = x, A_t = a\right] = \mathsf{P}_a(x, y), \tag{2}$$

where A_t denotes the learner's action primitive at time t. The action-dependent transition matrix P thus describes the dynamic behaviour of the process $\{X_t\}$.

We consider that the demonstration consists of a sequence $\mathcal{H}$ of state-action pairs

$$\mathcal{H} = \{(x_1, a_1), (x_2, a_2), \ldots, (x_n, a_n)\}.$$

Each pair (x_i, a_i) exemplifies to the learner the expected action (a_i) in each of the states visited during the demonstration (x_i). From this demonstration, the learning agent is expected to

[1] Clearly, the world model of the learner includes all necessary information relating the action repertoire for the learner and its ability to reproduce a particular effect. On the other hand, the world model of the teacher provides the only information relating the decision-rule of the teacher and its eventual underlying goal.

perceive what the demonstrated task is and, eventually by experimentation, learn how to perform it optimally. A decision-rule determining the action of the learner in each state of the environment is called *a policy* and is denoted as a map $\delta : \mathcal{X} \longrightarrow \mathcal{A}$. The learner should then *infer the task* from the demonstration and *learn the corresponding optimal policy*, that we henceforth denote by δ^*.

In our adopted formalism, the task can be defined using a function $r : \mathcal{X} \longrightarrow \mathbb{R}$ describing the "desirability" of each particular state $x \in \mathcal{X}$. This function r works as a *reward* for the learner and, once r is known, the learner should choose its actions to maximize the functional

$$J(x, \{A_t\}) = \mathbf{E}\left[\sum_{t=1}^{\infty} \gamma^t r(X_t) \mid X_0 = x\right],$$

where γ is a discount factor between 0 and 1 that assigns greater importance to those rewards received in the immediate future than to those in the distant future. We remark that, once r is known, the problem falls back to the standard formulation of reinforcement learning [24].

The relation between the function r describing the task and the optimal behavior rule can be evidenced by means of the function V_r given by

$$V_r(x) = \max_{a \in \mathcal{A}} \left[r(x) + \gamma \sum_{y \in \mathcal{X}} \mathsf{P}_a(x, y) V_r(y) \right] \tag{3}$$

The value $V_r(x)$ represents the expected (discounted) reward accumulated along a path of the process $\{X_t\}$ starting at state x, when the optimal behavior rule is followed. The optimal policy associated with the reward function r is thus given by

$$\delta_r(x) = \arg\max_{a \in \mathcal{A}} \left[r(x) + \gamma \sum_{y \in \mathcal{X}} \mathsf{P}_a(x, y) V_r(y) \right]$$

The computation of δ_r (or, equivalently, V_r) given P and r is a standard problem and can be solved using any of several standard methods available in the literature [24].

Methodology In the formalism just described, the fundamental imitation problem lies in the estimation of the function r from the observed demonstration $\mathcal{H}$. Notice that this is closely related to the problem of *inverse reinforcement learning* as described in [25]. We adopt the method described in [19], which is a basic variation of the *Bayesian inverse reinforcement learning* (BIRL) algorithm in [26].

For a given r-function, the *likelihood of a pair* $(x, a) \in \mathcal{X} \times \mathcal{A}$ is defined as

$$L_r(x, a) = \mathbf{P}\left[(x, a) \mid r\right] = \frac{e^{\eta Q_r(x,a)}}{\sum_{b \in \mathcal{A}} e^{\eta Q_r(x,b)}},$$

where $Q_r(x, a)$ is defined as

$$Q_r(x, a) = r(x) + \gamma \sum_{y \in \mathcal{X}} \mathsf{P}_a(x, y) V_r(y)$$

and V_r is as in (3). The parameter η is a user-defined *confidence parameter* that we describe further ahead. The value $L_r(x, a)$ translates the plausibility of the choice of action a in state x when the underlying task is described by r. Given a demonstration sequence

$$\mathcal{H} = \{(x_1, a_1), (x_2, a_2), \ldots, (x_n, a_n)\}.$$

the corresponding likelihood is

$$L_r(\mathcal{H}) = \prod_{i=1}^{n} L_r(x_i, a_i).$$

The method uses MCMC to estimate the distribution over the space of possible r-functions (usually a compact subset of $\mathbb{R}^p, p > 0$), given the demonstration [26]. It will then choose the maximum *a posteriori* r-function. Since we consider a uniform prior for the distribution, the selected reward is the one whose corresponding optimal policy "best matches" the demonstration. The confidence parameter η determines the "trustworthiness" of the method: it is a user-defined parameter that indicates how "close" the demonstrated policy is to the optimal policy [26].

Some important remarks are in order. First of all, to determine the likelihood of the demonstration for each function r, the algorithm requires the transition model in P. If such transition model is not available, then the learner will only be able to *replicate particular aspects of the demonstration*. However, as argued in [19], the imitative behaviour obtained in these situations may not correspond to actual imitation.

Secondly, it may happen that the transition model available is *inaccurate*. In this situation (and unless the model is significantly inaccurate) the learner should still be able to perceive the demonstrated task. Then, given the estimated r-function, the learner may only be able to determine a *sub-optimal policy* and will need to resort to *experimentation* to improve this policy.

3 Experiments

In this section we start by comparing the simulation results obtained using our proposed model with those observed in well-known biological experiments in children and primates. We also illustrate the application of our imitation-learning framework in a task with a robot.

3.1 Application to human imitative tasks

In a simple experiment described in [22], several infants were presented with a demonstration in which an adult turned a light on by pressing it with the head. One week later, most infants replicated this peculiar behaviour, instead of simply using their hand. Further insights were obtained from this experiment when, years later, a new dimension to the study was added by including task constraints [9]. In the new experiment, infants were faced with an adult turning the light on with the head but having the hands restrained/occupied. The results showed that, in this new situation, children would display a more significant tendency to use their hands to turn the light on. The authors suggest that infants understand the goal and the restriction and so when the hands are occupied they emulate because they assume that the demonstrator did not follow the "obvious" solution because was of the restrictions. Notice that, according to Fig. 1, using the head corresponds to *imitation* while using the hand corresponds to (goal) *emulation*.

We conducted two experiments. In the first experiment, we disregarded the preferences of the learner (*i.e.,* we set $\lambda_A = 0$) and observed how the behaviour changed as we assigned more importance to the replication of the observed effect (*i.e.,* as λ_E goes from 0 to 1). The results are depicted in Figure 2. Notice that, when faced with a restricted teacher, the learner switches to an "emulative" behaviour much sooner, replicating the results in [9].

On a second test, we disregarded the observed effect (*i.e.,* we set $\lambda_E = 0$) and observed how the behaviour of the learner changed as we assigned more importance to the demonstration,

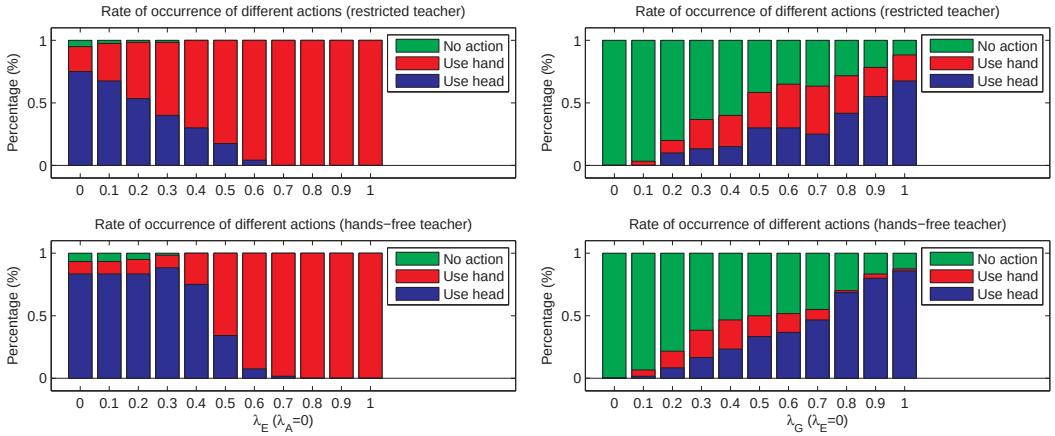


Fig. 2: (left) Change in behaviour as the learner increasingly focuses on replicating the observed effect. (right) Change in behaviour as the learner focuses less on its individual interests and more on the demonstration (goal, actions and effects). Each bar corresponds to a trial of 2,000 independent runs.

focusing less on its personal preferences (*i.e.,* as λ_G goes from 0 to 1). The results are depicted in Figure 2. Notice that, in this test, we set λ_E to zero, which means that the agent is not explicitly considering the observed effect. However, when combining its own interests with the observed demonstration (that includes goals, actions and effects), the learner tends to *replicate the observed effect* and disregard the observed actions, thus displaying emulative behaviour. This is particularly evident in the situation of a restricted teacher.

We emphasize that the difference in behaviour between the restricted and non-restricted teacher is due only to the *perceived difference on the ability of the teacher to interact with the environment.*

3.2 Application to robot imitation learning

We now present an application of our imitation learning methodology in a sequential task. We used BALTAZAR [27], a robotic platform consisting of a humanoid torso with one anthropomorphic arm and hand and a binocular head (see Figure 4). To implement the imitation learning algorithm in the robot we considered a simple recycling game, where the robot must separate different objects according to their shape (Figure 3). In front of the robot are two slots (Left and Right) where 3 types of objects can be placed: Large Balls, Small Balls and Boxes. The boxes should be dropped in a corresponding container and the small balls should be tapped out of the table. The large balls should be touched upon, since the robot is not able to efficiently manipulate them. Every time a large ball is touched, it is removed from the table by an external user. Therefore, the robot has available a total of 6 possible actions: Touch Left (TcL), Touch Right (ThR), Tap Left (TpL), Tap Right (TpR), Grasp Left (GrL) and Grasp Right (GrR).

For the description of the process $\{X_t\}$ for the task at hand, we considered a state-space consisting of 17 possible states. Of these, 16 correspond to the possible combinations of objects in the two slots (including empty slots). The 17th state is an invalid state that accounts for the situations where the robot's actions do not succeed (for example, when the robot drops the ball in an invalid position in the middle of the table).

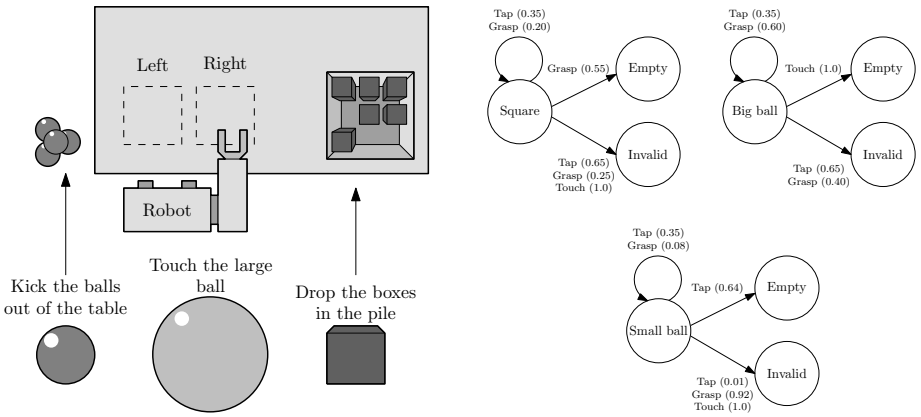


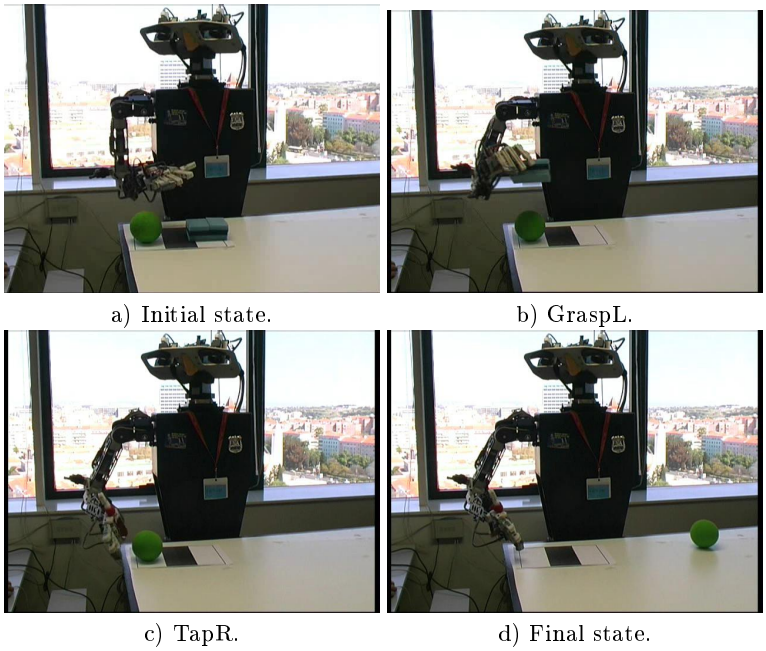
Fig. 3: (left)Recycling game (right)Transition diagrams describing the transitions for each slot/object.

To test the imitation, we first provided the robot with an error-free demonstration of the optimal behaviour rule. As expected, the robot was successfully able to reconstruct the optimal policy. We also observed the learned behaviour when the robot was provided with *two* different demonstrations, both optimal. The results are described in Table 1. Each state is represented as a pair (S_1, S_2) where each S_i can take one of the values "Ball" (Big Ball), "ball" (Small Ball), "Box" (Box) or $\emptyset$ (empty). The second column of Table 1 then lists the observed actions for each state and the third column lists the learned policy. Notice that, as before, the robot was able to reconstruct an optimal policy, by choosing one of the demonstrated actions in those states where different actions were observed.

Table 1: Demonstration 1: Error free demonstration. Demonstration 2: Inaccurate and incomplete demonstration, where the boxed cells correspond to the states not demonstrated or in which the demonstration was inaccurate. Columns 3 and 5 present the learned policy for Demo 1 and 2, respectively.

State	**Demo 1**	**Learned Pol.**	**Demo 2**	**Learned Pol.**
$(\emptyset, \text{Ball})$	TcR	TcR	$\boxed{\text{-}}$	TcR
$(\emptyset, \text{Box})$	GrR	GrR	GrR	GrR
$(\emptyset, \text{ball})$	TpR	TpR	TpR	TpR
$(\text{Ball}, \emptyset)$	TcL	TcL	TcL	TcL
(Ball, Ball)	TcL,TcR	TcL,TcR	$\boxed{\text{GrR}}$	TcL
(Ball, Box)	TcL,GrR	GrR	TcL	TcL
(Ball, ball)	TcL	TcL	TcL	TcL
$(\text{Box}, \emptyset)$	GrL	GrL	GrL	GrL
(Box, Ball)	GrL,TcR	GrL	GrL	GrL
(Box, Box)	GrL,GrR	GrR	GrL	GrL
(Box, ball)	GrL	GrL	GrL	GrL
$(\text{ball}, \emptyset)$	TpL	TpL	TpL	TpL
(ball, ball)	TpL,TcR	TpL	TpL	TpL
(ball, Box)	TpL,GrR	GrR	TpL	TpL
(ball, ball)	TpL	TpL	TpL	TpL

We then provided the robot with an *incomplete and inaccurate* demonstration. As seen in Table 1, the action at state (∅, Ball) was never demonstrated and the action at state (Ball, Ball) was *wrong*. The last column of Table 1 shows the learned policy. Notice that in this particular case the robot was able to recover the *correct policy*, even with an incomplete and inaccurate demonstration.



a) Initial state. b) GraspL. c) TapR. d) Final state.

Fig. 4: Execution of the learned policy in state (Box, SBall).

In Figure 4 we illustrate the execution of the optimal learned policy for the initial state (Box, SBall).[2]

To assess the sensitivity of the imitation learning module to the action recognition errors, we tested the learning algorithm for different error recognition rates. For each error rate, we ran 100 trials. Each trial consists of 45 state-action pairs, corresponding to three optimal policies. The obtained results are depicted in Figure 5.

As expected, the error in the learned policy increases as the number of wrongly interpreted actions increases. Notice, however, that for small error rates ($\leq 15\%$) the robot is still able to recover the demonstrated policy with an error of only 1%. In particular, if we take into account the fact that the error rates of the action recognition method used by the robot are between 10% and 15%, the results in Figure 5 guarantee a high probability of accurately recovering the optimal policy.

4 Discussion and concluding remarks

We presented a formalism that generates several social learning behaviours, namely, imitation and emulation. The model makes use of several sources of information such as goals, actions

[2] For videos showing additional experiences see http://vislab.isr.ist.utl.pt/baltazar/demos/

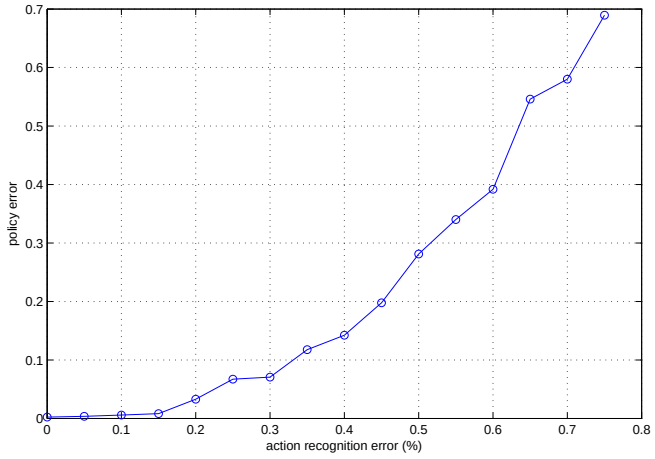


Fig. 5: Percentage of wrong actions in the learned policy as the action recognition errors increase.

and effects. Besides those three sources, suggested in [2], we also explicitly consider knowledge about the world model (affordances) of the demonstrator and of the learner [16]. The difference between them helps to explain which actions may be irrelevant. We also include the agents actions preferences (in terms of energy "cost"), to be able to quantitatively describe what is meant by an "inefficient" action.

The taxonomy in [2] reveals itself very interesting by providing a easy way to develop a mathematical model of social-learning behaviours, mainly imitation and (goal) emulation. Notwithstanding, we argue that the separation generally observed between behaviours might not translate into separable cognitive mechanisms. Under the experimental paradigm described previously ([9]) there is no third alternative: the system either performs one of the two actions and is labelled "emulator" or performs the other and is labelled "imitator". We suggest that when the complexity of the task increases, *i.e.,* when considering sequences of actions or more action possibilities, it will not be possible to label with precision if a particular behaviour is emulative or imitative, and maybe different percentages of action matching will be observed.

We also suggest that the difference between imitation and goal-emulation is only a matter of preference/motivation of the learner. As seen from our mathematical model, all cognitive capabilities might be active at the same time, but when faced with options the motivation to interact socially or to achieve the results as fast as possible will weight differently the different sources of information.

Finally, the proposed model can provide robotics with better "social skills", by being able to understand and predict to some extent the outcome of people's actions.

References

1. A. Whiten, V. Horner, C. A. Litchfield, and S. Marshall-Pescini, "How do apes ape?" *Learning & Behavior*, vol. 32, no. 1, pp. 36–52, 2004.
2. J. Call and M. Carpenter, "Three sources of information in social learning," in *Imitation in animals and artifact*. MIT Press Cambridge, MA, USA, 2002.
3. F. Bellagamba and M. Tomasello, "Re-enacting intended acts: Comparing 12- and 18-month-olds," *Infant Behavior and Development*, vol. 22, no. 2, pp. 277–282, 1999.

4. S. Johnson, A. Booth, and K. O'Hearn, "Inferring the goals of a nonhuman agent," *Cognitive Development*, vol. 16, no. 1, pp. 637–656, 2001.
5. H. Bekkering, A. Wohlschläger, and M. Gattis, "Imitation of gestures in children is goal-directed." *Quarterly J. Experimental Psychology*, vol. 53A, pp. 153–164, 2000.
6. S. Want and P. Harris, "Learning from other people's mistakes: Causal understanding in learning to use a tool," *Child Development*, vol. 72, no. 2, pp. 431–443, 2001.
7. A. N. Meltzoff, "Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children," *Developmental Psychology*, vol. 31, no. 5, pp. 838–850, 1995.
8. V. Horner and A. Whiten, "Causal knowledge and imitation/emulation switching in chimpanzees (pan troglodytes) and children (homo sapiens)," *Animal Cognition*, vol. 8, pp. 164–181, 2005.
9. G. Gergely, H. Bekkering, and I. Király, "Rational imitation in preverbal infants," *Nature*, vol. 415, p. 755, 2002.
10. T. R. Zentall, "Imitation in animals: Evidence, function, and mechanisms," *Cybernetics and Systems*, vol. 32, no. 1, pp. 53–96, 2001.
11. H. Kozima, C. Nakagawa, and H. Yano, "Emergence of imitation mediated by objects," in *2nd Int. Workshop on Epigenetic Robotics*, 2002.
12. L. Montesano, M. Lopes, A. Bernardino, and J. Santos-Victor, "Learning object affordances: From sensory–motor coordination to imitation," *IEEE Transactions on Robotics*, vol. 24, no. 1, pp. 15–26, Feb. 2008.
13. P. Fitzpatrick, G. Metta, L. Natale, S. Rao, and G. Sandini., "Learning about objects through action: Initial steps towards artificial cognition," in *IEEE International Conference on Robotics and Automation*, Taipei, Taiwan, 2003.
14. M. Lopes and J. Santos-Victor, "A developmental roadmap for learning by imitation in robots," *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics*, vol. 37, no. 2, April 2007.
15. B. Jansen and T. Belpaeme, "A model for inferring the intention in imitation tasks," in *15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN06)*, Hatfield, UK, 2006.
16. M. Lopes, F. S. Melo, and L. Montesano, "Affordance-based imitation learning in robots," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, USA, Nov 2007, pp. 1015–1021.
17. C. L. Baker, J. B. Tenenbaum, and R. R. Saxe, "Bayesian models of human action understanding," *Advances in Neural Information Processing Systems*, vol. 18, 2006.
18. R. Rao, A. Shon, and A. Meltzoff, *Imitation and social learning in robots, humans, and animals.* Cambridge University Press, 2007, ch. A Bayesian model of imitation in infants and robots.
19. F. Melo, M. Lopes, J. Santos-Victor, and M. I. Ribeiro, "A unified framework for imitation-like behaviors," in *4th International Symposium in Imitation in Animals and Artifacts*, Newcastle, UK, April 2007.
20. A. Alissandrakis, C. L. Nehaniv, and K. Dautenhahn, "Action, state and effect metrics for robot imitation," in *15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 06)*, Hatfield, United Kingdom, 2006, pp. 232–237.
21. M. Lopes and J. Santos-Victor, "Visual learning by imitation with motor representations," *IEEE Trans. Systems, Man, and Cybernetics - Part B: Cybernetics*, vol. 35, no. 3, 2005.
22. A. N. Meltzoff, "Infant imitation after a 1-week delay: Long-term memory for novel acts and multiple stimuli," *Developmental Psychology*, vol. 24, no. 4, pp. 470–476, 1988.
23. G. Csibra and G. Gergely, ""obsessed with goals": Functions and mechanisms of teleological interpretation of actions in humans," *Acta Psychologica*, vol. 124, pp. 60–78, 2007.
24. R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction.* USA: MIT Press, 1998.
25. P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *Proceedings of the 21st International Conference on Machine Learning (ICML'04)*, 2004, pp. 1–8.
26. D. Ramachandran and E. Amir, "Bayesian inverse reinforcement learning," in *20th Int. Joint Conf. Artificial Intelligence*, India, 2007.
27. M. Lopes, R. Beira, M. Praça, and J. Santos-Victor, "An anthropomorphic robot torso for imitation: design and experiments." in *International Conference on Intelligent Robots and Systems*, Sendai, Japan, 2004.