Applying Psychology of Persuasion to Conversational Agents through Reinforcement Learning: an Exploratory Study

Francesca Di Massimo[1], **Valentina Carfora**[2], **Patrizia Catellani**[2] and **Marco Piastra**[1]
[1]Computer Vision and Multimedia Lab, Università degli Studi di Pavia, Italy
[2] Dipartimento di Psicologia, Università Cattolica di Milano, Italy
francesca.dimassimo01@universitadipavia.it
valentina.carfora@unicatt.it
patrizia.catellani@unicatt.it
marco.piastra@unipv.it

Abstract

This study is set in the framework of *task-oriented conversational agents* in which *dialogue management* is obtained via *Reinforcement Learning*. The aim is to explore the possibility to overcome the typical end-to-end training approach through the integration of a quantitative model developed in the field of persuasion psychology. Such integration is expected to accelerate the training phase and improve the quality of the dialogue obtained. In this way, the resulting agent would take advantage of some subtle psychological aspects of the interaction that would be difficult to elicit via end-to-end training. We propose a theoretical architecture in which the psychological model above is translated into a probabilistic predictor and then integrated in the reinforcement learning process, intended in its *partially observable* variant. The experimental validation of the architecture proposed is currently ongoing.

1 Introduction

A typical conversational agent has a multi-stage architecture: spoken language, written language and dialogue management, see Allen et al. (2001). This study focuses on dialogue management for task-oriented conversational agents. In particular, we focus on the creation of a dialogue manager aimed at inducing healthier nutritional habits in the interactant.

Given that the task considered involves psychosocial aspects that are difficult to program directly, the idea of achieving an effective dialogue manager via machine learning techniques, *reinforcement learning* (RL) in particular, may seem attractive. At present, many RL-based approaches involve training an agent end-to-end from a dataset of recorded dialogues, see for instance Liu (2018). However, the chance of obtaining significant results in this way entails substantial efforts in both collecting sample data and performing experiments. Worse yet, such efforts ought to rely on the even stronger hypothesis that the RL agent would be able to elicit psychosocial aspects on its own. As an alternative, in this study we envisage the possibility to enhance the RL process by harnessing a model developed and accepted in the field of social psychology to provide a more reliable learning ground and a substantial accelerator for the process itself.

Our study relies on a quantitative, causal model of human behavior being studied in the field of social psychology (see Carfora et al., 2019) aimed at assessing the effectiveness of message *framing* to induce healthier nutritional habits. The goal of the model is to assess whether messages with different frames can be differentially persuasive according to the users' psychosocial characteristics.

2 Psychological model: Structural Equation Model

Three relevant psychosocial antecedents of behaviour change are the following: *Self-Efficacy* (the individual perception of being able to eat healthy), *Attitude* (the individual evaluation of the pros and cons) and *Intention Change* (the individual willingness of adhering to a healthy diet). These psychosocial dimensions cannot be directly observed and need to be measured as *latent* variables. To this purpose, questionnaires are used, each composed by a set of questions or *items* (i.e. *observed* variables). *Self-Efficacy* is measured with 8 items, each associated to a set of answers ranging from "not at all confident" (1)

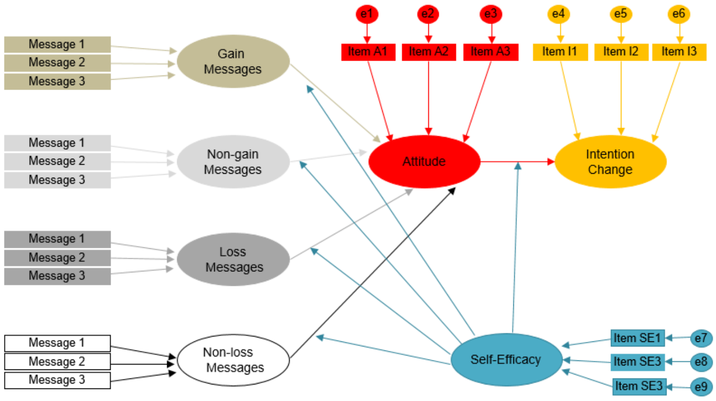

Figure 1: SEM simplified model for the case at hand.

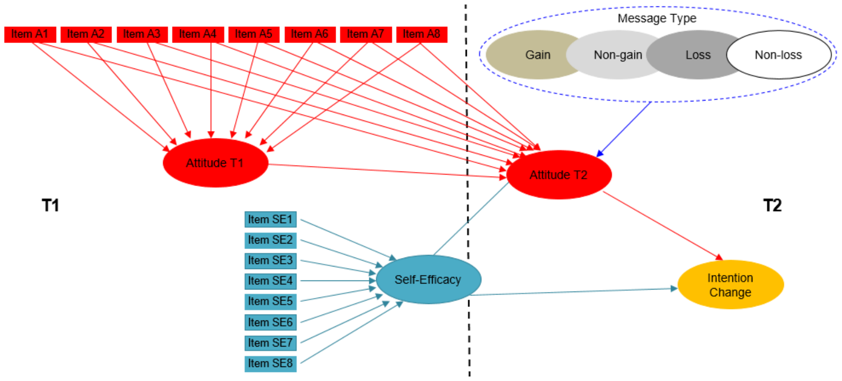

Figure 2: DBN translation of the SEM shown in Figure 1.

to "extremely confident" (7). *Attitude* is assessed through 8 items associated to a differential scale ranging from 1 to 7 (the higher the score, the more positive the attitude). *Intention Change* is measured with three items on a Likert scale, ranging from 1 (“definitely do not”) to 7 (“definitely do”). See Carfora et el. (2019).

In our study, the psychosocial model was assessed experimentally on a group of volunteers. Each participant was first proposed a questionnaire (Time 1 – T1) for measuring *Self-Efficacy*, *Attitude* and *Intention Change*. In a subsequent phase (i.e. *message intervention*), participants were randomly assigned to one of four groups, each receiving a different type of persuasive message: *gain* (i.e. positive behavior leads to positive outcomes), *non-gain* (negative behavior prevents positive outcomes), *loss* (negative behavior leads to negative outcomes) and *non-loss* (positive behavior prevents negative outcomes) (Higgins, 1997; Cesario et al., 2013). In a last phase (Time 2 - T2), the effectiveness of the message intervention was then evaluated with a second questionnaire, to detect changes in participants’ *Attitude* and *Intention Change* in relation to healthy eating.

The overall model is described by the *Structural Equation Model* (SEM, see Wright, 1921) in Figure 1. For simplicity, only three items are shown for each latent variable. Besides allowing the description of latent variables, SEMs are *causal* models in the sense that they allow a statistical analysis of the strength of causal relations among the latents themselves, as represented by the arrows in figure. SEMs are linear models, and thus all causal relations underpin linear equations.

Note that latent variables in a SEM have different roles: in this case *gain/non-gain/loss/non-loss* messages are *independent variables*, *Intention Change* is a *dependent variable*, *Attitude* is a *mediator* of the relationship between the independent and the dependent variables, and *Self-Efficacy* is a *moderator*, namely, it explains the intensity ot the relation it points at. *Intention Change* was measures at both T1 and T2, *Attitude* was measured at both T1 and T2, and *Self-Efficacy* was measured at T1 only. Note that the time transversality (i.e. T1 → T2) is implicit in the SEM depiction above.

3 Probabilistic model: Bayesian Network

Once the SEM is defined, we aim to translate it into a probabilistic model, so as to obtain the probability distributions needed for the learning process. We resort to a graphical model, and in particular to a *Bayesian Network* (BN, see Ben Gal, 2007), namely a graph-based description of both the observable and latent random variables in the model and their conditional dependencies. In BNs, nodes represent the variables and edges represent dependencies between them, whereas the lack of edges implies their independence, hence a simplification in the model. As a general rule, the joint probability of a BN can be inferred as follows:

$$P(X_1, \ldots, X_N) = \prod_{i=1}^{N} P(X_i \mid parents(X_i)),$$

where $X_1, \ldots, X_N$ are the random variables in the model and $parents(X_i)$ indicate all the nodes having an edge oriented towards X_i.

In the case at hand, a temporal description of the model, accounting for the time steps T1 and T2, is necessary as well. For this purpose, we use a *Dynamic Bayesian Network* (DBN, see Dagum et al., 1992). The DBN thus obtained is shown in Figure 2.

Notice that the messages are only significant at T2, as they have not been sent yet at T1. We gathered message in the one node *Message Type*, assuming it can take four, mutually exclusive values. The mediator *Attitude* is measured at both time steps while the moderator *Self-Efficacy* is constant over time, as suggested in Section 2. *Intention Change* has relevance at T2 only since, as we will mention in Section 5, it will be used to estimate a reward function once the final time step is reached.

4 Learning the BN

The collected data are as follows. The analysis was conducted on 442 interactants, divided in four groups, each one receiving a different type of messages[1]. The answers to the items of the questionnaire always had a range of 7 values. However, this induces a combinatory esplosion, making it impossible to cover all the subspaces ($7^8 = 5.764.801$ different combinations for *Attitude*, for instance). We thus decide to aggregate: $low :=$ (1 to 2); $medium :=$ (3 to 5); $high :=$ (6 to 7).

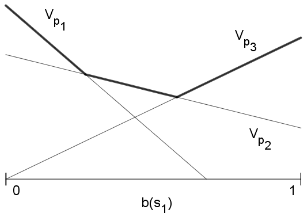

Figure 3: Basic example of computation of V_π in a case where $\mathcal{S} = \{s_1, s_2\}$. p_1, p_2, p_3 are three possible policies.

Our aim is to learn the *Joint Probability Distribution* (JPD) of our model, as that would make us able to answer, through marginalizations and conditional probabilities, any query about the model itself. The conditional probability distributions to be learnt in the case in point are then the following:

- $P(\textit{Item } Ai)$, for $i = 1, \ldots, 8$;
- $P(\textit{Item } SEi)$, for $i = 1, \ldots, 8$;
- $P(\textit{Message Type})$;
- $P(\textit{Attitude } T1 \mid \textit{Item } Ai, i = 1, \ldots, 8)$;
- $P(\textit{Self-Efficacy} \mid \textit{Item } SEi, i = 1, \ldots, 8)$;
- $P(\textit{Attitude } T2 \mid \textit{Item } Ai, i = 1, \ldots, 8, \textit{Message Type}, \textit{Self-Efficacy})$;
- $P(\textit{Intention Change} \mid \textit{Attitude } T2, \textit{Self-Efficacy})$.

The first three can be easily inferred from the raw data as relative frequencies. As for the following four, even aggregating the 7 values as mentioned, a huge amount of data would still be necessary ($3^8 \cdot 2^4 \cdot 3 = 314.928$ subspaces for *Attitude T2*, for instance). As conducting a psychological study on that amount of people would not be feasible, we address the issue with an appropriate choice of the method. To allow using *Maximum Likelihood Estimation* (MLE) to learn the BN, we resort to the *Noisy-OR* approximation (see Oniśko, 2001). According to this, through a few appropriate changes (not shown) to the graphical model, the number of subspaces can be greatly reduced (e.g. $3 \cdot 2 \cdot 3 = 18$ for *Attitude T2*).

5 Reinforcement Learning: Markov Decision Problems

The translation into a tool to be used for reinforcement learning is obtained in the terms of *Markov*

[1] The original study included also a control group, which we do not consider here.

Decision Processes (MDPs), see Fabiani et al. (2010).

Roughly speaking, in a MDP there is a finite number of situations or *states* of the environment, at each of which the agent is supposed to select an action to take, thus inducing a state transition and obtaining a *reward*. The objective is to find a *policy* determining the sequence of actions that generates the maximum possible cumulative reward, over time. However, due to the presence of latents, in our case the agent is not able to have complete knowledge about the state of the environment. In such a situation, the agent must build its own estimate about the current state based on the memory of past actions and observations. This entails using a variant of the MDPs, that is *Partially Observable Markov Decision Processes* (POMDPs, see Kaelbling 1998). We then define the following, with reference to the variables mentioned in Figure 2:

$$\mathcal{S} := \{\text{states}\} = \{\textit{Attitude } T2, \textit{Self-Efficacy}\};$$

$$\mathcal{A} := \{\text{actions}\} = \{\text{ask } A1, \ldots, \text{ask } A8\} \cup \{\text{ask } SE1, \ldots, \text{ask } SE8\} \cup \{G, NG, L, NL\},$$

where Ai denotes the question for $Item\ Ai$, SEi denotes the question for $Item\ SEi$ and G, NG, L, NL denote the action of sending Gain, Non-gain, Loss and Non-loss messages respectively;

$$\Omega := \{\text{observations}\} = \{\textit{Item } A1, \ldots, \textit{Item } A8, \textit{Item } SE1, \ldots, \textit{Item } SE8\}.$$

Starting from an unknown initial state s_0 (often taken to be uniform over $\mathcal{S}$, as no information is available), the agent takes an action a_0, that brings it, at time step 1, to state s_1, unknown as well. There, an observation o_1 is made.

The process is then repeated over time, until a *goal* state of some kind has been reached. Hence, we can define the *history* as an ordered succession of actions and observations:

$$h_t := \{a_0, o_1, \ldots, a_{t-1}, o_t\}, h_0 = \emptyset.$$

As at all steps there is uncertainty about the actual state, a crucial role is played by the agent's estimate about the state of the environment, i.e. by the *belief state*. The agent's belief at time step t, denoted as $\mathbf{b}_t$, is driven by its previous belief $\mathbf{b}_{t-1}$ and by the new information acquired, i.e. the action taken a_{t-1} and observation made o_t. We then have:

$$b_{t+1}(s_{t+1}) = P(s_{t+1} \mid \mathbf{b}_t, a_t, o_{t+1}).$$

In the POMDP framework, the agent's choices about how to behave are influenced by its belief state and by the history. Thus, we define the agent's *policy*:

$$\pi = \pi(\mathbf{b}_t, h_t),$$

that we aim to optimize. To complete the picture, we define the following functions to describe the model evolution in time (the notation $'$ indicates a reference to the subsequent time step):

state-transition function:

$$T\colon (s, a) \mapsto P(s' \mid s, a) := T(s', s, a);$$

observation function:

$$O\colon (s, a) \mapsto P(o' \mid a, s') := O(o', a, s');$$

reward function:

$$R\colon (s, a) \mapsto \mathbb{E}\,[r' \mid s, a] := R(s, a).$$

These functions can be easily adapted to the specifics of the case at hand. It can be seen that, once the JPD derived from the DBN is completely specified, the reward is deterministic. In particular, it is computed by evaluating the changes in the values for the latent *Intention Change*.

As we are interested in finding an optimal policy, we now need to evaluate the goodness of each state when following a given policy. As there is no certainty about the states, we define the *value function* as a weighted average over the possible belief states:

$$V_\pi(\mathbf{b}_t, h_t) := \sum_{s_t} b_t(s_t) V_\pi(s_t, \mathbf{b}_t, h_t),$$

where $V_\pi(s_t, \mathbf{b}_t, h_t)$ is the *state* value function. The latter depends on the expected reward (and on a discount factor $\gamma \in [0, 1]$ stating the preference for fast solutions):

$$V_\pi(s_t, \mathbf{b}_t, h_t) := R(s_t, \pi(\mathbf{b}_t, h_t)) + \gamma \sum_{s_{t+1}} T(s_{t+1}, s_t, \pi(\mathbf{b}_t, h_t)) * \sum_{o_{t+1}} O(o_{t+1}, \pi(\mathbf{b}_t, h_t), s_{t+1}) V_\pi(s_{t+1}, \mathbf{b}_{t+1}, h_{t+1}).$$

Finally, we define the target of our seek, namely the *optimal value function* and the related *optimal policy*, as:

$$\begin{cases} V^*(\mathbf{b}_t, h_t) := \max_\pi V_\pi(\mathbf{b}_t, h_t), \\ \pi^*(\mathbf{b}_t, h_t) := argmax_\pi V_\pi(\mathbf{b}_t, h_t). \end{cases}$$

It can be shown that the optimal value function in a POMDP is always piecewise linear and convex, as

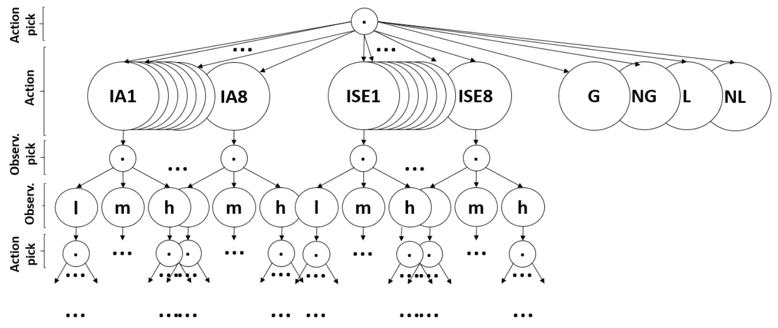

Figure 4: Expansion of the policy tree. l, m, h stand for *low, medium* and *high.*

exemplified in Figure 3. In other words, the optimal policy (in bold in Figure 3) combines different policies depending on their belief state values.

The next step is to use the POMDP to detect the optimal policy, that is the sequence of questions to ask to the interactant, in order to draw her/his profile, hence the message to send, which maximizes the effectiveness of the interaction. To this end, the contribution of the DBN is fundamental. From the JPD associated, in fact, we construct the probability distributions necessary to define the functions T, O, R that compose the value function.

6 Policy from Monte Carlo Tree Search

It is evident from Figure 4, describing the full expansion of the policy tree for the case in point, that the computational effort and power required for a brute-force exploration of all possible combinations is unaffordable.

Among all the policies that can be considered, we want to select the optimal ones, thus avoiding coinsidering policies that are always underperforming. In other words, with reference to Figure 3, we want to find V_{p_1}, V_{p_2}, V_{p_3} among those of all possible policies, and use them to identify the optimal policy V^*.

To accomplish this, we select the *Monte Carlo Tree Search* (MCTS) approach, see Chaslot et al. (2008), due to its reliability and its applicability to computationally complex practical problems. We adopt the variant including an *Upper Confidence Bound formula*, see Kocsis et al. (2006). This method combines *exploitation* of the previously computed results, allowing to select the game action leading to better results, with *exploration* of different choices, to cope with the uncertainty of the evaluation. Thus, using $V_\pi(s_t, \mathbf{b}_t, h_t)$ as defined before to guide the exploration, the MCTS method reliably converges (in probability) to optimal policies. These latter will be applied by the conversational agent in the interaction with each specific user, to adapt both the sequence and the amount of questions to her/his personality profile and selecting the message which is most likely to be effective.

7 Conclusions and future work

In this work we explored the possibility of harnessing a complete and experimentally assessed SEM, developed in the field of persuasion psychology, as the basis for the reinforcement learning of a dialogue manager that drives a conversational agent whose task is inducing healthier nutritional habits in the interactant. The fundamental component of the method proposed is a DBN, which is derived from the SEM above and acts like a predictor for the belief state value in a POMDP.

The main expected advantage is that, by doing so, the RL agent will not need a time-consuming period of training, possibly requiring the involvement of human interactants, but can be trained 'in house' – at least at the beginning – and be released in production at a later stage, once a first effective strategy has been achieved through the DBN. Such method still requires an experimental validation, which is the current objective of our working group.

Acknowledgments

The authors are grateful to Cristiano Chesi of IUSS Pavia for his revision of an earlier version of the paper and his precious remarks. We also acknowledge the fundamental help given by Rebecca Rastelli, during her collaboration to this research.

References

Allen, J., Ferguson, G., & Stent, A. 2001. *An architecture for more realistic conversational systems*. In Proceedings of the 6th international conference on Intelligent user interfaces (pp. 1-8). ACM.

Anderson, Ronald D. & Vastag, Gyula. 2004. *Causal modeling alternatives in operations research: Overview and application*. European Journal of Operational Research. 156. 92-109.

Auer, Peter & Cesa-Bianchi, Nicolò & Fischer, Paul. 2002Kocsis, Levente & Szepesvári, Csaba. 2006. *Bandit Based Monte-Carlo Planning. Finite-time Analysis of the Multiarmed Bandit Problem*. Machine Learning. 47. 235-256.

Bandura, A. 1982. *Self-efficacy mechanism in human agency*. American Psychologist, 37, 122-147.

Baron, Robert A. & Byrne, Donn Erwin & Suls, Jerry M. 1989. *Exploring social psychology, 3rd ed.* Boston, Mass.: Allyn and Bacon. 0205119085.

Ben Gal I. 2007. *Bayesian Networks*. Encyclopedia of Statistics in Quality and Reliability. John Wiley & Sons.

Bertolotti, M., Carfora, V., & Catellani, P. 2019. *Different frames to reduce red meat intake: The moderating role of self-efficacy*. Health Communication, in press.

Carfora, V., Bertolotti, M., & Catellani, P. 2019. *Informational and emotional daily messages to reduce red and processed meat consumption*. Appetite, 141, 104331.

Cesario, J., Corker, K. S., & Jelinek, S. 2013. *A self-regulatory framework for message framing*. Journal of Experimental Social Psychology, 49, 238-249.

Chaslot, Guillaume & Bakkes, Sander & Szita, Istvan & Spronck, Pieter. 2008. *Monte-Carlo Tree Search: A New Framework for Game AI*. Bijdragen.

Dagum, Paul and Galper, Adam and Horvitz, Eric. 1992. *Dynamic Network Models for Forecasting*. Proceedings of the Eighth Conference on Uncertainty in Artificial Intelligence.

Dagum, Paul and Galper, Adam and Horvitz, Eric and Seiver, Adam. 1999. *Uncertain reasoning and forecasting*. International Journal of Forecasting.

De Waal, Alta & Yoo, Keunyoung. 2018. *Latent Variable Bayesian Networks Constructed Using Structural Equation Modelling*. 2018 21st International Conference on Information Fusion. 688-695.

Fabiani, Patrick & Teichteil-Königsbuch, Florent. 2010. *Markov Decision Processes in Artificial Intelligence*. Wiley-ISTE.

Gupta, Sumeet & W. Kim, Hee. 2008. *Linking structural equation modeling to Bayesian networks: Decision support for customer retention in virtual communities*. European Journal of Operational Research. 190. 818-833.

Heckerman, David. 1995. *A Bayesian Approach to Learning Causal Networks*.

Higgins, E.T. 1997. *Beyond pleasure and pain*. American Psychologist, 52, 1280-1300.

A. Howard, Ronald. 1972. *Dynamic Programming and Markov Process*. The Mathematical Gazette. 46.

Pack Kaelbling, Leslie & Littman, Michael & R. Cassandra, Anthony. 1998. *Planning and Acting in Partially Observable Stochastic Domains*. Artificial Intelligence. 101. 99-134.

Kocsis, Levente & Szepesvári, Csaba. 2006. *Bandit Based Monte-Carlo Planning*. Machine Learning: ECML 2006. Springer Berlin Heidelberg. 282-293.

Lai, T.L & Robbins, Herbert. 1985. *Asymptotically Efficient Adaptive Allocation Rules*. Advances in Applied Mathematics. 6. 4-22.

Liu, Bing. 2018. *Learning Task-Oriented Dialog with Neural Network Methods*. PhD thesis.

Murphy, Kevin. 2012. *Machine Learning: A Probabilistic Perspective*. The MIT Press. 58.

Pearl Judea. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Representation and Reasoning Series (2nd printing ed.). San Francisco, California: Morgan Kaufmann.

Oniśko, Agnieszka & Druzdzel, Marek J. & Wasyluk, Hanna. 2001. *Learning Bayesian network parameters from small data sets: application of Noisy-OR gates*. International Journal of Approximate Reasoning. 27.

Silver, David & Veness, Joel. 2010. *Monte-Carlo Planning in Large POMDPs*. Advances in Neural Information Processing Systems. 23. 2164-2172.

Matthijs T. J. Spaan. 2012. *Partially Observable Markov Decision Processes*. In: Reinforcement Learning: State of the Art. Springer Verlag. 387-414.

Sutton, Richard & G. Barto, Andrew. 1998. *Reinforcement Learning: An Introduction*. IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council. 9. 1054.

Wright, Sewall. 1921. *Correlation and causation*. Journal of Agricultural Research. 20. 557–585.

Young, Steve & Gasic, Milica & Thomson, Blaise & Williams, Jason. 2013. *POMDP-based statistical spoken dialog systems: A review*. Proceedings of the IEEE, 101. 1160-1179.