

Measure Of Belief Change as an Evaluation of Persuasion

Pierre Andrews and Suresh Manandhar¹

Abstract. In the field of natural argumentation and computer persuasion, there has not been any clear definition of the persuasiveness of a system trying to influence the user. In this paper, we describe a general evaluation task that can be instantiated on a number of domains to evaluate the beliefs change of participants. Through the use of a ranking task, we can measure the participant's change of beliefs related to a behaviour or an attitude. This general metric allows a better comparison of state of the art persuasive systems.

1 Motivation and Related Work

In novel fields of research, researchers often want to compare their approaches and the efficiency of their research. Thus, alongside the new field a research movement is created to develop robust evaluation frameworks that can provide comparative results and fair evaluations of research output for the field. For example, in the Information Retrieval field, the researchers have long studied different techniques of evaluation and selected the precision/recall measures, thus creating a framework of measures that can be used by all researchers and create evaluation campaigns such as the Text REtrieval Conferences (TREC <http://www.trec.nist.gov>).

The field of automated persuasion is attracting a growing interest in the research community, with new conferences and workshops every year [16, 19]. However, there has yet not been an agreed method for evaluating and comparing persuasive systems' output.

Existing research already provides examples of evaluation techniques for persuasion. For instance [4] uses a long term evaluation procedure to follow the change of students' behaviour when trying to persuade them to walk more. The measure of persuasiveness introduced by the authors is computed from the evolution of steps count for each participant, showing the change in walking behaviour of the students over one month. In this experimental setup, the researchers need a large amount of resources and time to provide pedometers to students, motivate them to use the system on a long term basis and wait for results; this amount of resources are not always available to all researchers. In addition, following a long term behaviour change is not an atomic setup and it is difficult to control for every external factors that can influence the user's behaviour.

[21] describes a smoking cessation program that tries to persuade participants to stop smoking through tailored letters. The users are asked if they think they will stop smoking in the month or six months following the reading of the letter. Participants were also asked if they had actually quit six month after the intervention. In this experiment, the authors show that there is no difference in the change of behaviour between the control group and the group that reads the tailored letters. The authors acknowledge that their experiment and trial was too small to show any statistical evidence. It is in fact difficult

with such binary observation to extract enough data and it is a general problem to be able to find enough participants to follow on such a long term experiment.

In behavioural medicine, many measures have been developed to evaluate the changes in different mental constructs associated to behaviour change. [11], for example, proposes a questionnaire to evaluate the stage of change (see [20]) of participants within the pain treatment domain, while [15] develops a scale to measure the evolution of self efficacy in the domain of arthritis treatment.

Other persuasive system researches take a more concise approach by evaluating a change during the persuasive session of an external representation of the user's intentions towards a behaviour. For instance [8] evaluates an embodied conversational agent simulating a real estate agent by comparing the amount of money that clients are prepared to spend on a house before and after the interaction with the estate agent. The estate agent tries to convince users to buy a house fifty percent more expensive than what they are actually ready to spend. The persuasiveness of the system is evaluated by looking at the increase in the user's budget after the interaction. The measure is between zero percent to 100% increase relative to the target increase chosen by the system.

[18] tries to evaluate the effect of distance over persuasion for computer mediated communication. The author uses a setup following the desert survival scenario [14] where participants have to rank a set of items relating to their survival in the desert. After having given an initial ranking, the participants are then faced with persuasive messages relevant to these items and finally give a ranking of the same items after the persuasive session. The author uses as a measure of persuasion the distance between the participant's final ranking and the ranking of the persuader. [7] introduces a variation of the ranking task in the domain of house buying; instead of having to rerank a full list of items (houses in this case), the participants are persuaded to insert a new item in their initial ranking. This evaluation measures how many users actually chose the new alternative and where they ranked it in the initial ranking. These measures allow the authors to evaluate the persuasion and the effectiveness of the tailoring of the arguments.

We believe that a ranking task such as the one used by [18] can apply to different domains and be used as a common evaluation metric to compare persuasive systems. In this paper, we ground the validity of this ranking task in theory of persuasion and describe a formalisation of the ranking task that provides an evaluation metric for controlled experiments that can be more robust to external factor. It also provides a standard measure available in many domains and that can be compared between researches. We also conclude that there is a need for more research in persuasion evaluation frameworks to help the development of the automatic persuasion and natural argumentation field.

¹ University of York, United Kingdom, email: pierre.andrews@gmail.com; suresh@cs.york.ac.uk

2 Behaviour and Belief Change

When thinking of persuasion, the immediate indicator of success is a change in the behaviour of the persuaded subject; in fact, [17] proposed the following definition of *Persuasive Communication*:

“Any message that is intended to shape, reinforce or change the responses of another or others.” from [22], p. 4

It is generally accepted that the “*changed responses*” refers to either a change in behaviour or a change in the attitude towards the behaviour (see [22]). However, behaviours can take many forms and the method of evaluating a change in behaviour will be different for every application domain. For instance [4] tries to evaluate a change in walking behaviour and uses the number of steps a user performs as a measure of behaviour change. In another health advice domain, [21] tries to convince participants to stop smoking, the evaluation output is thus the number of participants that stopped smoking. [4] describes a continuous evaluation value for each participant that is hard to port to other domains whereas [21] describes a binary value that does not provide powerful data for analysis but is easy to understand. Both evaluation methods consider a change of behaviour and provide the authors with a tool to demonstrate the persuasiveness of their system. However, it is difficult for the reader to make a comparison between the approaches’ performances.

However, research in sociology and persuasive communication shows that intentions towards a behaviour can be modelled as a function of the user’s beliefs about such behaviour and the social norms influencing the user. For instance, [1] presents the Theory of Reasoned Action that is designed to predict one’s *intention* (I_B) to perform a particular *behaviour* (B) as a function f of one’s *attitude* toward this behaviour (A_B) and of the *subjective norms* the behaviour is exposed to (SN_B). Equation (1) represents this influence, where W_1 and W_2 are the personal importance attributed to each component:

The attitude is defined by (2) where b_i is the “*belief value*” and e_i is the “*evaluation*” of that belief,

The subjective norms is defined by (3) where b'_i is the “*normative belief value*”. i.e. the reference belief of the group the receiver considers himself in – and m_i the “*motivation*” to follow the group beliefs.

$$I_B = f(W_1 \times A_B + W_2 \times SN_B) \quad (1)$$

$$A_B = \sum b_i \times e_i \quad (2)$$

$$SN_B = \sum b'_i \times m_i \quad (3)$$

The standard example provided in persuasive communication lecture books ([22] for example) relates to the act of filing and paying taxes. The belief b_i would then be “I should file taxes” and the final intention I_B “I will file my taxes”. The usual attitude A_B towards the behaviour is very low as its evaluation in the person’s mind is low, however, the social norms, influenced by laws and peer pressure, are high. Thus, at the end, the intention towards the behaviour is still high and the taxes will be filed and paid.

A similar representation of human reasoning was developed within the Belief-Desire-Intention (BDI) model [5]. This model describes the actual intention of realising an action – or a behaviour – that is linked to someone’s desires about this behaviour and the relying world representation contained in its beliefs. However, [5] does not provide a model as strict as the one proposed by [1] to describe the

relations between these layers of practical reasoning. [6] tries to rationalise this relationship between beliefs and goals – or intentions – in the practical reasoning models close to BDI.

Evaluating a change of behaviour can thus be done, according to this model, by evaluating a change in the beliefs linked to the behaviour or a change in the influences of social norms. In a controlled experiment, one can choose to evaluate one or the other independently.

In particular, in a controlled experiment were the change in social norms’ influence is controlled for – on a short term evaluation for example –, researchers can evaluate a change in beliefs and evaluate the persuasion as a change in the attitude towards a behaviour instead of direct behavioural observation.

Beliefs can be linked to the judgement of a behaviour, but also to some external representation. For example [8] uses such a technique to evaluate the persuasiveness of their embodied conversational agent where instead of measuring the actual buying behaviour to see if the system is persuasive, the authors use a view of the attitude towards this behaviour given by the amount of money participants are ready to spend. However, this measure stays limited to the domain.

In this paper, we discuss beliefs that can be linked to behaviour’s intentions as well as to a ranking between a set of items, which we believe can be applied to various domains and can provide a measure for comparison between researches. [18, 3] use the desert scenario task to provide a ranking task to participants: they are told that they are stranded in the desert after a plane crash and should rank a set of items (compass, map, knife, etc.) according to their usefulness for the participants’ survival. The resulting ranking provides an external representation of the set of beliefs each participant has formed about the utility of each item.

The ranking does not provide a detailed view on every internal belief that the user holds about the situation, however, if the user changes this ranking, this change represents a measurable change in the internal beliefs. According to the Theory of Reasoned Action this change in beliefs has an impact on the user’s intention towards the behaviour, and we can assume that the measured persuasion has an influence on the behaviour too.

3 Measuring Persuasiveness

The ranking task provides an observation of the user’s beliefs that can be used to extract a metric evaluation that can be shared and compared between research domains. In this section, we present the general metric measure that can be used and consider different issues in implementing a ranking task and applying the persuasiveness metric.

When participating in a ranking task, the participants first give their preferred initial ranking R_i of items (for example, in the desert scenario task: knife, compass, map, ...) and then engage in with the persuasive system which attempts to change the participants’ ranking to a different ranking R_s ; at the end of the persuasion session, the participants can change their items choice to a final ranking R_f (see figure 1).

The persuasiveness of the session is measured as the evolution of the distance between the user’s rankings (R_i , R_f) and the system’s goal ranking (R_s). If the system is persuasive, it changes the user’s beliefs about the items ranking towards a ranking similar to the system’s ranking. The change of beliefs is reflected by the evolution of the distance between rankings as defined by equation (4).

$$P = \Delta(d(R_f, R_s), d(R_i, R_s)) \quad (4)$$

3.1 Swap measure

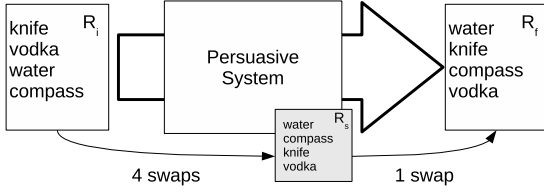


Figure 1. Desert Scenario Ranking Task Example

There exist a number of distance measures between rankings [13, 10, 12]. The Kendall τ coefficient is generally used to measure a difference between rankings. However, this measure is not a metric and is not always straightforward to interpret. A part of the Kendall τ coefficient is however a metric and provides an intuitive measure in the ranking task. The “Kendall τ permutation metric” [13] is used to compute the pairwise disagreement between two rankings; measuring the number of swaps between adjacent items to get from one ranking to the other ranking. The Kendall τ permutation metric between the rankings R_1 and R_2 is defined in Equation (5)² where P_{airs} is the set of all possible pairs of items of R_1 and R_2 .

$$K_{\tau}(R_1, R_2) = \sum_{\{i,j\} \in P_{airs}} \bar{K}_{i,j}(R_1, R_2) \quad (5)$$

$$\bar{K}_{i,j}(R_1, R_2) = \begin{cases} 0 & \text{if the pair of items } i \text{ and } j \text{ are in the same} \\ & \text{order in the two rankings,} \\ 1 & \text{otherwise} \end{cases} \quad (6)$$

Equation (7) defines the evolution of the Kendall τ permutations metric during the persuasive session and provides a metric evaluation of the system’s persuasiveness.

$$\begin{aligned} \bar{P}_{persuasiveness} &= \Delta(K_{\tau}(R_i, R_s), K_{\tau}(R_f, R_s)) \\ &= K_{\tau}(R_i, R_s) - K_{\tau}(R_f, R_s) \end{aligned} \quad (7)$$

For example, if the user’s initial ranking of the items is $R_i = \text{map} > \text{flashlight} > \text{compass}$ and the system goal ranking is $R_s = \text{compass} > \text{flashlight} > \text{map}$. The Kendall τ permutations metric is calculated with the table of pairs:

R_i	R_s	$\bar{K}(R_i, R_s)$
map > compass	map < compass	1
map > flashlight	map < flashlight	1
flashlight > compass	flashlight < compass	1
$K_{\tau}(R_i, R_s)$		3

If the final user ranking is $R_f = \text{flashlight} > \text{compass} > \text{map}$, the table of pairs is:

R_f	R_s	$\bar{K}(R_f, R_s)$
compass > map	compass > map	0
flashlight > map	flashlight > map	0
flashlight > compass	flashlight < compass	1
$K_{\tau}(R_f, R_s)$		1

At the beginning of the persuasive session, the distance is maximum between the two rankings – three swaps are needed – whereas, at the end of the session, only one swap is required. The persuasiveness metric is then: $\bar{P}_{persuasiveness} = 3 - 1 = 2$.

For an n items ranking, the range of the persuasiveness metric is thus

$$\left[-\frac{n \times (n-1)}{2}, \frac{n \times (n-1)}{2} \right]$$

To be able to compare different persuasive systems that can rely on heterogeneous ranking task with different numbers of items, we need to normalise this persuasiveness measure as defined by equation (8).

$$P_{persuasiveness} = \frac{2 \times (K_{\tau}(R_i, R_s) - K_{\tau}(R_f, R_s))}{n \times (n-1)} \quad (8)$$

3.2 Interpretation and Constrains

In this general approach to the ranking task, the normalised persuasiveness metric will have a minimum of -1 and a maximum of +1.

- The *minimum* corresponds to the case where the participants actually made the maximum number of swaps **away** from the system’s ranking between the initial and the final ranking.
- A *null* $P_{persuasiveness}$ means that the participant did not change the ranking and that the system was not persuasive.
- The *maximum* $P_{persuasiveness}$ corresponds to a successful persuasion of the system as the participants will have done the maximum number of swaps **towards** the system’s ranking and $R_f = R_s$.

In this general setup of the ranking task, there is however a issue for the interpretation of the results. What does it mean for the persuasive system that the users change their beliefs **away** from the persuasive goals that the system was seeking? Was the system extremely bad? is $P_{persuasiveness} < 0$ worst than $P_{persuasiveness} = 0$? It is actually difficult to interpret the $P_{persuasiveness}$ metric in its negative range.

[9] discusses “arguments that backfire”, where the use of fallacy lowers the audience’s trust in the speaker and thus lowers the effectiveness of the argumentation. This might make the whole persuasion “backfire”, yielding negative results that will make the audience go against the speaker persuasive goals, even if they shared initial beliefs. This will explain negative $P_{persuasiveness}$ results as the shared beliefs represented in the initial ranking will be lost and the participant will provide a final ranking further away from the system’s goal ranking than the initial ranking. The negative results are thus valid in their interpretation and can help detect backfiring argumentation strategies that alienate the audience.

However, an additional issue with this setup of the ranking task makes it hard to compare between different domains instantiation. For example, in an extreme case of this general view of the ranking task, the user can enter an initial ranking R_i that is the same as the ranking R_s chosen by the system. In this case, the task of the persuasive system is not to persuade users but to keep the same ranking and

² from [10]

the only evolution of ranking that can be observed are swaps **away** from the system’s ranking. In this extreme case, it is not interesting to compare the persuasiveness metric with another persuasive session where $R_i \neq R_s$ and where the system would have had to actually do some persuasion *effort*.

To be able to compare different persuasive systems with this ranking task, the persuasion task, with regards to this ranking task, should be of comparable *effort*. Normalising the $P_{persuasiveness}$ allows to compare different persuasive task that have a different number of items, but does not protect from comparing a system persuading the user to do a little relative number of swaps with a system that has to persuade the user of a large relative number of swaps.

A solution to get a uniform $P_{persuasiveness}$ metric, which can be compared between systems, is to guarantee that each system will have a comparable persuasive *effort*. This can be guaranteed by choosing the system’s goal ranking R_s to always maximise the persuasive *effort* by maximising the number of swaps needed to go from R_i to R_s . This is guaranteed by choosing R_s as the invert ranking of R_i as shown in the example given above. In this case, the initial distance between rankings is $\frac{n \times (n-1)}{2}$ where n is the number of items in the ranking.

If the ranking task is defined with this constrain, then we can write the $\bar{P}_{persuasiveness}$ as defined by equation (9) which implies that the persuasiveness range is $[0, \frac{n \times (n-1)}{2}]$ and the normalised persuasiveness, defined by equation (10), has a range of $[0, 1]$. If the participant is not persuaded by the system, then $R_i = R_f$ and $P_{persuasiveness} = 0$ but if the system is persuasive, then the participant has done the maximum number of swaps **towards** the system ranking and $P_{persuasiveness} = 1$ as $R_f = R_s$.

$$\bar{P}_{persuasiveness} = \frac{n \times (n - 1)}{2} - K_{\tau}(R_f, R_s) \quad (9)$$

$$P_{persuasiveness} = 1 - \frac{2 \times K_{\tau}(R_f, R_s)}{n \times (n - 1)} \quad (10)$$

When designing the persuasive experiment and setting the ranking task, the researcher should therefore be very attentive that the chosen system’s goal ranking is always the invert of the user’s ranking. The system must also be able to achieve such a persuasion.

The non maximised setup of the ranking task is helpful in detecting “backfiring” argumentation which will move the user’s beliefs away from the system’s goal belief. This provides a good insight of the argumentation process but is not usable for comparing different systems’ performances to change the user’s belief. The second measure, can be used for this purpose as it guarantees the maximisation of the persuasion, however, nuances of the belief change will be lost as, in this setup, there is no option for the participant to *disagree* more with the system. The goal of the experiment should thus set the measure to use:

- if the experiment is designed to evaluate the persuasive strategies of the system, then it is interesting to leave space for the participants to *disagree* with the system and the first measure should be preferred.
- if the experiment is designed to compare the system’s effectiveness to change the user’s beliefs between system, then it is recommended to use the second “maximised disagreement” measure that removes the bias of the initial belief choice.

4 Sample Experiment and Results

[18, 3] used the desert survival scenario [14] ranking task to evaluate the persuasiveness of dialogue sessions but did not formalise a general persuasiveness metric. In our research, we have used a similar ranking task based on a different scenario to evaluate a persuasive system with the formal $P_{persuasiveness}$ metric described earlier. In this section, we report initial observations on the use of this metric as well as an example of a different scenario where the ranking task can be used.

Our research was evaluating a human-computer dialogue system able to discuss with users to persuade them. The domain chosen to evaluate this dialogue system was similar to a restaurant recommendation scenario. Twenty-eight participants were told that they would discuss with an automated dialogue system simulating one of their friend in a chat about a shortlist of restaurants where they could bring mutual friends for dinner.

After having been explained the scenario, the participants are presented with a list of ten restaurants described by five attributes (food quality, cuisine type, cost, service quality and decor) and are asked to choose three restaurants they would like to propose as possible alternatives to their group of friends. They can choose any three restaurants and rank them in their order of preference.

The actual dialogue system has access to a database of around one thousand restaurants³, but asking the user to evaluate, in a short time, all of these restaurants is not realistic. In the same way, asking them to rank the full list of ten restaurants is not possible and would not correspond to a natural task that the participants would perform in real life.

After having given information about their restaurants preference and a specific restaurants choice, the participants are faced with a dialogue session with a system simulating a friend that tries to persuade them to keep the same selection of three restaurants, but to choose a different preference order. In this case, to ensure maximum persuasion *effort*, the system always chooses a ranking of restaurants that is the invert of the user’s choice.

At the end of the dialogue, the participants are asked to give their final ranking of restaurants reflecting their agreement with the simulated friend. This is used as the final ranking to measure the persuasiveness of the system. The participants are also asked to fill in a questionnaire relating to different aspects of the dialogue system. In this experiment, to evaluate the fitness of our evaluation metric, the participants are asked to rate the statement “*The other user was persuasive*” on a five points likert scale: “*Strongly disagree, Disagree, Neither agree nor disagree, Agree, Strongly Agree*”.

This statement evaluates the persuasion perceived by the participants during the dialogues. The persuasiveness metric applied in this case shows that there is a significant correlation between the user perception and the persuasion measured through the ranking task (Spearman $\rho = 0.70$, $p < 0.01$)⁴. This confirms that the measure is at least as good as asking the question directly to the users. However, getting such direct measure might bias the answer of the users.

Observation of the answers from the user also shows the need for a side measure of persuasiveness. In the seven participants that answered that they “neither agree nor disagree” to the statement, an outlying participant that does not perceive a strong persuasion but is still persuaded more than the other. In this case, the side measure of

³ provided by M.A. Walker from [23]

⁴ in a similar setup with 52 participants, the same question was asked and also yields a significant correlation with the persuasiveness measure (Spearman $\rho = 0.38$, $p < 0.01$)

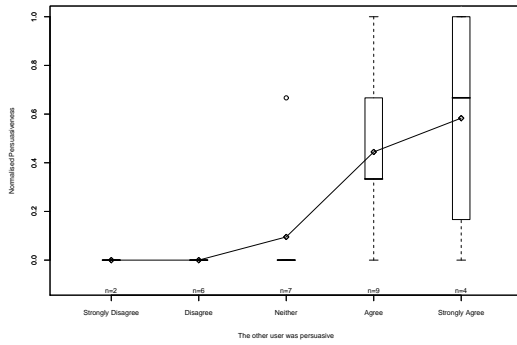


Figure 2. Correlation Between the Perceived Persuasion and the Measured Persuasion. n is the number of participants that gave this particular answer.

the rank change allows to see that users were persuaded even if they did not perceive a strong persuasion from the system.

Similarly, the “strongly agree” answers show that there is a distribution of the persuasiveness measure along the whole axis: some of the participants that perceived a strong persuasion from the system did not actually change their ranking⁵. This illustrates the case where users do actually feel that they are persuaded but might not have changed their beliefs accordingly. In which case, the system cannot be said to be persuasive.

Thus, a side measure of persuasion, that does not directly rely on participants’ self evaluation can show more information about the actual persuasion process while staying a good indicator of the system’s persuasive performances.

5 Ordering of Beliefs vs. Ordering from Beliefs

In belief revision literature, in particular within the AGM model [2], someone’s belief set is represented as a set of consistent axioms on which operations can be performed to revise or update the beliefs. Axioms can be added or removed from the set at each revision to maintain the consistency of the belief set. However, in someone’s mind, not all beliefs are equal as some are said to be more *entrenched*, and they are harder to remove from the person’s belief set.

This entrenchment affects the possible revisions of the belief set and can be seen as a preference ordering of beliefs in the person’s mind. Beliefs higher in one’s preferences will be harder to change and remove from the belief set than lower beliefs.

Belief revision, which is the base of the persuasion discussed in this paper is thus seen as an operation on a set of ordered beliefs that can be extended, reduced or reordered. This *ordering of beliefs* could be seen as similar to the ranking task proposed in this paper: the ranking represents the entrenchment ordering of the user’s belief and the system’s task is to make the user revise such ordering.

However, the ranking task is actually less abstract and each item of the ranking does not need to directly map to a belief in the participant’s mind. For example, in the ranking task of the desert survival scenario, each item does not map to one of the participant’s belief (or an axiom representing such belief).

For instance, two items are available in the desert survival scenario: an airmap and a compass. Most participants have the belief

⁵ note that this could also be due to a misunderstanding of the instructions by some of these participants.

that they can use these two items to find their way out of the desert. If these two items are ranked high in the initial ranking, the system can assume that the participant holds the following beliefs:

- “I can walk across the desert to find rescue.”
- “I can find my way to rescue on the map.”
- “I can use the compass for orientation on the map.”

The ranking of the compass and the map over a flashlight for example does not represent a direct preference ranking over beliefs but that the participant sees more use for these items than for the flashlight, *because* of his current beliefs.

In the restaurant domain, the ranking represents the users preference towards the restaurants, these preferences are not a direct mapping to an entrenchment ordering, but is still related to this concept. If a user ranked a *Pizzeria* over a *Grill*, this might map to a set of preference ordering over the cuisine type. However, it might also be that the Pizzeria is cheaper than the Grill.

Another example is the smoking cessation program, the ranking items could be directly mapped to a set of beliefs related to why the participant is smoking, such as: “smoking makes me feel better”, “smoking makes me look cool”, “smoking will kill me”, etc. However, these might be hard to change as some of the beliefs might be too entrenched. A different, indirect, ranking task could evaluate the change of beliefs about smoking while avoiding too much entrenchment bias; for example, the participants could be asked to rank a set of items they would buy first if they had a limited amount of money, such a set could contain “a bottle of water”, “a pack of cigarettes”, “a newspaper”, “a lighter”, etc. The reranking of the items relating to smoking, while not ensuring that the participants will stop smoking, will still show a change in their attitude towards the smoking behaviour.

The choice of the ranking items should thus not be directly mapped to a set of ordered beliefs or preferences, but to a set of items that represent, in practice, a set of knowledge and of preferences about the domain. The ranking will be guided by the user’s belief: a *ranking from beliefs*, but might not directly map to the *ranking of beliefs* in the user’s mind.

6 Conclusion and Discussion

In this paper, we have introduced the different approaches of evaluating systems’ persuasion through the state of the art of automated persuasion. We have also formalised a framework providing a reusable persuasiveness metric that could be used by other researchers to compare different automated persuasion approaches.

The applications illustrated in the paper are short term setups that cannot evaluate the long term impact on the participants, but actually, this ranking task can also be used as a measure in long term evaluations. For example, in the case of the smoking cessation problem [21], the use of a ranking task might have provided more insight in the beliefs change of the users after the first intervention; six months later, the same ranking task without extra intervention might have been used to evaluate the beliefs that remained of the persuasion, even if the participants did not stop smoking. Such ranking task would have thus given more insight on why the system was not effective.

This paper provides sample results that show that the proposed persuasive measure is at least as good as directly asking the user about the persuasion while providing a *hidden* measure that does not bias the participants. It remains to be shown if this measure correlates with actual behaviour change. This was not in the scope of our

research but will have to be evaluated if researchers have to use the measure in more complex domains where the user's personal beliefs might not have a strong weight in comparison to the social norms affecting the behaviour.

In addition, the change in the ranking evaluates a change in the user's attitude but might not be directly linked to persuasion as such a change might come from coercion and threats. Thus a measure of coercion is required to ensure that the change measured by the proposed metric comes actually from persuasion. For example, to evaluate coercion, in the reported sample experiment, the user was directly asked the question: "The other user was not forceful in changing your opinion" which did not show to correlate with the persuasiveness metric.

We have shown that the ranking task can be applied to different domains, however, to use such a task, the persuasion must be performed on a domain where behaviours or attitudes can be mapped to a ranked set of items. It is clear that not all persuasive domains can be reduced to a ranking task. In addition, doing such reduction might limit artificially the scope of research on automated persuasion.

We believe that the formalisation of the ranking task as a framework for evaluating persuasive systems is a first step towards finding an appropriate evaluation methodology for comparing persuasive systems. It is important for the development of the field of automatic persuasion and natural argumentation that researchers extend their work on a set of standard evaluation frameworks that can be used to evaluate and compare systems on long and short term changes in the user's beliefs, attitude and behaviours. In addition, this paper only discussed the problem of evaluating the existence and ranking of beliefs linked to a behaviour, but the problem remains to find a task to evaluate the social norms influencing the behaviour.

REFERENCES

- [1] I. Ajzen and M. Fishbein, *Understanding attitudes and predicting social behaviour*, Prentice-Hall, Englewood Cliffs, New Jersey, 1980.
- [2] C. E. Alchourrón, P. Gärdenfors, and D. Makinson, 'On the logic of theory change: Partial meet contraction and revision functions', *Journal of Symbolic Logic*, **50**, 510–530, (1985).
- [3] Pierre Andrews, Suresh Manandhar, and Marco De Boni, 'Argumentative human computer dialogue for automated persuasion', in *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pp. 138–147, Columbus, Ohio, (June 2008). Association for Computational Linguistics.
- [4] Timothy W. Bickmore and Rosalind W. Picard, 'Establishing and maintaining long-term human-computer relationships', *ACM Transaction of Human-Computer Interaction*, **12**(2), 293–327, (June 2005).
- [5] Michael E. Bratman, *Intention, Plans, and Practical Reason*, Cambridge University Press, March 1999.
- [6] F. Paglieri C. Castelfranchi, 'The role of beliefs in goal dynamics: Prolegomena to a constructive theory of intentions', *Synthese*, **155**, 237–263, (2007).
- [7] Giuseppe Carenini and Johanna D. Moore, 'An empirical study of the influence of argument conciseness on argument effectiveness', in *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ed., Hitoshi Iida, pp. 150–157, Hong Kong, (October 2000).
- [8] J. Cassell and T. W. Bickmore, 'Negotiated collusion: Modeling social language and its relationship effects in intelligent agents', *User Modeling and Adaptive Interfaces*, **13**(1-2), 89–132, (February 2002).
- [9] Daniel H. Cohen, 'Arguments that backfire', in *The Uses of Argument: Proceedings of a conference at McMaster University*, ed., David Hitchcock, pp. 58–65. Ontario Society for the Study of Argumentation, (April 2005).
- [10] Ronald Fagin, Ravi Kumar, and D. Sivakumar, 'Comparing top k lists', in *SODA '03: Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 28–36, Philadelphia, PA, USA, (2003). Society for Industrial and Applied Mathematics.
- [11] M. P. Jensen, W. R. Nielson, J. M. Romano, M. L. Hill, and J. A. Turner, 'Further evaluation of the pain stages of change questionnaire: is the transtheoretical model of change useful for patients with chronic pain?', *Pain*, 255–264, (June 2000).
- [12] M. G. Kendall, 'A new measure of rank correlation', *Biometrika*, **30**(1/2), 81–93, (June 1938).
- [13] Maurice Kendall and Jean D. Gibbons, *Rank Correlation Methods*, A Charles Griffin Title, fifth edn., September 1990.
- [14] J. C. Lafferty and P. M. Eady, *The desert survival problem*, Plymouth, Michigan: Experimental Learning Methods, 1974.
- [15] K. Lorig, R. L. Chastain, E. Ung, S. Shoor, and H. R. Holman, 'Development and evaluation of a scale to measure perceived self-efficacy in people with arthritis', *Arthritis and rheumatism*, **32**(1), 37–44, (1989).
- [16] *Symposium on Persuasive Technology, in conjunction with the AISB 2008: Convention Communication, Interaction and Social Intelligence*, eds., Judith Masthoff, Chris Reed, and Floriana Grasso, Aberdeen, April 2008.
- [17] Gr Miller, 'On being persuaded: Some basic distinctions.', in *Persuasion: New directions in theory and research*, eds., M. E. Roloff and G. R. Miller, 11–28, SAGE Publications, (January 1980).
- [18] Youngme Moon, 'The effects of distance in local versus remote human-computer interaction', in *CHI '98: Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 103–108, New York, NY, USA, (1998). ACM Press/Addison-Wesley Publishing Co.
- [19] *Persuasive Technology. Proceedings of the Third International Conference, PERSUASIVE 2008*, eds., H. Oinas-Kukkonen, P. Hasle, M. Harjumaa, K. Segerstahl, and P. Öhrström, volume 5033 of *Lecture Notes in Computer Science: Information Systems and Applications, incl. Internet/Web, and HCI*, Springer, Oulu, Finland, July 2008.
- [20] J. O. Prochaska and Carlo Diclemente, 'Stages of change in the modification of problem behavior', *Progress in Behavior Modification*, **28**, 183–218, (1992).
- [21] Ehud Reiter, Roma Robertson, and Liesl M. Osman, 'Lessons from a failure: generating tailored smoking cessation letters', *Artificial Intelligence*, **144**(1-2), 41–58, (2003).
- [22] James B. Stiff and Paul A. Mongeau, *Persuasive Communication*, The Guilford Press, second edn., October 2002.
- [23] M. A. Walker, S. J. Whittaker, A. Stent, P. Maloor, J. Moore, M. Johnston, and G. Vasireddy, 'Generation and evaluation of user tailored responses in multimodal dialogue', *Cognitive Science: A Multidisciplinary Journal*, **28**(5), 811–840, (2004).