Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Ethical Dilemmas for Adaptive Persuasion Systems

Oliviero Stock, Marco Guerini, Fabio Pianesi
FBK-Irst
Via Sommarive 18, Trento - I-38123 Italy
stock@fbk.eu, guerini@fbk.eu, pianesi@fbk.eu

Abstract

A key acceptability criterion for artificial agents will be the possible moral implications of their actions. In particular, intelligent persuasive systems (systems designed to influence humans via communication) constitute a highly sensitive topic because of their intrinsically social nature. Still, ethical studies in this area are rare and tend to focus on the output of the required action; instead, this work focuses on the acceptability of persuasive acts themselves. Building systems able to persuade while being ethically acceptable requires that they be capable of intervening flexibly and of taking decisions about which specific persuasive strategy to use. We show how, exploiting a behavioral approach, based on human assessment of moral dilemmas, we obtain results that will lead to more ethically appropriate systems. Experiments we have conducted address the type of persuader, the strategies adopted and the circumstances. Dimensions surfaced that can characterize the interpersonal differences concerning moral acceptability of machine performed persuasion, usable for strategy adaptation. We also show that the prevailing preconceived negative attitude toward persuasion by a machine is not predictive of actual moral acceptability judgement when subjects are confronted with specific cases.

Introduction

Autonomous agents are such because they are able to decide suitable courses of actions for achieving their own goals, can maintain intentions in action and so on; in all these respects, the capability of discerning *good* from *bad* is an essential feature of autonomous artificial agents. Until recently, though, ethical issues have concerned less machines than designers, who have been deciding about the behavior of artifacts as well as the degrees of freedom they can be allowed in their choices. But the quest for autonomy in systems' actions and the rising sensitivity to the moral implications it has, requires that we move ahead and focus our attention on ethical acceptability of machines' choices. For example the case of "moral dilemmas" for autonomous cars has been discussed – e.g. (Kirkpatrick 2015) – situations in which any available choice leads to infringing some "accepted" ethical principle and yet a decision has to be taken.

The importance of ethical issues is heightened for systems that interact and communicate with humans, since moral acceptability is one of the ultimate criteria for acceptability tout court. We focus on *adaptive* persuasive technologies (Kaptein, Duplinsky, and Markopoulos 2011) i.e. systems able to pursue the goal of affecting the attitudes and/or behaviour of their interaction partners by adjusting communication to the latter's preferences, dispositions, etc. Studies on moral acceptability in this field have mostly targeted the action the persuading system intends the persuadee to perform rather than the communicative action the persuading system exploits to this end, e.g. (Verbeek 2006). We believe, though, that it is essential to understand the moral acceptability of the latter, i.e. of the communicative action enacted by the persuading system to achieve its specific goal.

A preliminary issue concerns the fact that at first sight people simply seem not to accept the idea that a machine persuades humans. If such a negative attitude exists, does it affect the moral acceptability of persuasion in specific scenarios? Obtaining a negative answer to this question would boost the construction of intelligent persuasive systems. Another important issue is what can be done to move forward toward dynamically adapting the persuasive strategies to the target's moral sensitivities and to the circumstances. Optimizing the moral acceptability of the persuasive actions would require a characterization of: a) the differential effects (if any) of the adopted communicative strategies (e.g., do classical argumentation strategies fare better than those relying on positive/negative emotions or those exploiting lies to influence people?); b) the role of circumstances (if any), including the action the persuading system intends the persuadee to perform; c) the purported individual differences for the adaptive system to exploit. We submit that these issues can be profitably addressed by leveraging the tradition of so called natural ethics and by using *moral dilemmas* as a probe (see (Wallach and Allen 2008; Anderson and Anderson 2007)). In two recent studies, we have taken the first steps towards addressing this research program by designing and executing experiments adopting the moral dilemma paradigm.

Related Work

Persuasion and artificial agents. Through the years, a number of prototypes for the automatic generation of linguistic persuasion expressions, based on reasoning capabilities, were developed, see (Guerini et al. 2011) for an

overview. The strategies adopted are of various nature and mainly refer to argumentative structure, appeal to emotions and deceptive devices such as lies. The area of health communication was one of the first being investigated (Kukafka 2005). Worth mentioning in this connection are STOP, one of the best known systems for behaviour inducement (Reiter, Sripada, and Robertson 2003) and *Migraine* (Carenini, Mittal, and Moore 1994), a natural language generation system for producing personalized information sheets for migraine patients. Argumentation systems have a long tradition in AI, and recently also experimental studies concerned with human behavior have been proposed (Rosenfeld and Kraus 2015). The role of lies (i.e. invalid arguments) was investigated in a computational setting by (Rehm and Andrè 2005). In (Carofiglio and deRosis 2003) the focus is on the generation of persuasive affective messages.

Recently there has also been a growing interest in persuasive internet and mobile services (Oinas-Kukkonen and Harjumaa 2009; Torning and Oinas-Kukkonen 2009). In parallel with these application-oriented studies, there has been a quest after evaluation methodologies to assess the effectiveness of persuasive communication by means of crowdsourcing approaches (Mason and Suri 2010; Aral and Walker 2011; Guerini, Strapparava, and Stock 2012).

Ethics and artificial agents. In recent years a few authors have contributed to bringing ethics to the main scene of AI, especially with a view of helping design moral robots. For instance (Allen, Wallach, and Smit 2006; Anderson and Anderson 2007) provided inspiration for seriously tackling this topic, whereas (Wallach and Allen 2008; Anderson and Anderson 2011) are important references for those approaching computational ethics. As far as implemented prototypes are concerned, the work by the group of Ken Forbus, which developed one of the very few existing moral decision-making reasoning engines (Dehghani et al. 2008), is outstanding. Their cognitively motivated system, called MoralDM, operates on two mutually exclusive modes, representing utilitarian and deontological reasoning.

As for moral issues in persuasion, most of the work concerns guidelines derived from general theories/principles. The classical reference is (Berdichevsky and Neuenschwander 1999) that provides a set of ethical principles for persuasive design subsumed by the golden rule "the creators of a persuasive technology should never seek to persuade anyone of something they themselves would not consent to be persuaded of." A more structured approach is provided by (Yetim 2011). In (Guerini and Stock 2005), a model for ethical persuasive agents and systems is proposed, based on logical and meta-planning modules of ethical reasoning.

Finally, some authors argued that moral concerns about machine-performed persuasion could be settled if the users of such systems were made aware of the aims and effects of the deployed influencing strategies. Still, studies have shown that such an approach decreases the chances of the influence success (Kaptein, Duplinsky, and Markopoulos 2011). This finding emphasizes the necessity of a fine-grained understanding of the ethical acceptability of the various persuasive strategies in different contexts of use.

Moral dilemmas. Moral dilemmas are situations in which every option at hand leads to breaking some ethical principles, this way requiring people to make explicit comparative choices and rank what is more (or less) acceptable in the given situation. These characteristics have often motivated their usage in discussions in popular media and newspapers about the risks of new technologies. In scholarly work, moral dilemmas have been acknowledged as an important source of insights as they allow for the collection of first-hand empirical data about moral acceptability that would otherwise be very difficult to obtain. The best known dilemmas are those exploited in (Thomson 1976). In one scenario (the *bystander* case) a trolley is about to engage a bifurcation, with the switch oriented toward a rail where five men are at work. A bystander sees all the scene and can divert the train on another track where it will kill *only* one person and save the other five lives. In another scenario (the *footbridge* case) the trolley is again going to hit five workers, but this time instead of having a switch lever available, the deciding agent is on a footbridge with a big man that, if pushed down the bridge, would fall in front of the trolley, this way preventing it from hitting the five workers. Importantly, all involved people do not know each other.

Philosophers and cognitive scientists have shown that most people consider the bystander case morally acceptable; the footbridge case is more controversial, despite the fact that the saving and the sacrifice of human lives are the same - see for example (Mikhail 2007; Hauser 2006). The common explanation for this asymmetry is that the footbridge scenario involves a personal moral violation (the deciding agent is the immediate causal agent of the big man's death) which causes affective distress and is judged much less permissible (Thomson 1976). More recent studies (Nichols and Mallon 2006), however, have challenged this view. Leaving aside other differences, in a newly proposed *catastrophic* scenario, similar to the footbridge case, the train transports a very dangerous virus and it is destined to hit a bomb that, unbeknownst to the train driver, was placed on the rails. The explosion will cause a catastrophic epidemic causing the death of half of the world population. The deciding agent knows all this and has in front of him/her the big man. In this case most people display more flexibility and a more utilitarian view of morality: saving such a high number of people in exchange of one 'personally-caused' death seems acceptable.

Trolley persuasion scenarios experiments

In the following we address the questions raised in the introduction. To this end, we exploit three trolley scenarios described above as they occupy a central place in the moral dilemma literature and have proven to be capable of eliciting different moral acceptance judgments.

First study. In this study we focused on the moral acceptability of persuasion actions as they depend on: a) the adopted communicative strategies; b) the circumstances and in particular the action the persuader intends the persuadee to perform, (Verbeek 2006); c) the nature of the persuading agents. As to the adopted persuasion strategies, we focused on their truth-value (validity), contrasting plain argumentation to the resort to lie, and the appeal to positive emotions

vs. the appeal to negative ones. Concerning situational aspects, we modelled them into a three-level factor: the bystander, the footbridge and the catastrophic scenarios introduced above ($\mathtt{SC}_{\mathtt{bys}}$, $\mathtt{SC}_{\mathtt{foo}}$ and $\mathtt{SC}_{\mathtt{cat}}$ respectively). Finally, a two-level (human, represented by a stationmaster, vs. an intelligent surveillance system capable of issuing persuading messages) factor controlled for the effects of the persuader nature. The whole resulted in a 2*3*4 mixed between-within design, with 24 conditions each realized by means of textual stimuli produced through the instantiation of general templates. Each scenario template is an adaptation to the persuasion case of those exploited in the literature discussed above, see Table 1 for an example.

There is a runaway trolley barreling down the railway tracks. Ahead, on the tracks, there are five people. The trolley is headed straight for them. On a footbridge, above the tracks, there are a very large man and Alex. Everybody sees the situation. We know that if the large man is pushed down onto the tracks below, the trolley would hit and kill the large man, but his body is so large that it would stop the trolley before it reach the five endangered people. Suddenly from the nearby service intercom the voice of the [persuader] shouts to Alex: [message realizing the persuasive strategy]

Table 1: Stimulus template for the footbridge scenario

The stimuli were administered to 124 undergraduate students, 30 males (24%) and 94 females (76%), of the psychology department of a university located in northern Italy. Each subject was randomly assigned to one of the two levels of the agent factor. For each stimulus, subjects were asked to say whether they found it morally acceptable that the agent (depending on the assigned group) used the specific communicative act contained in the stimulus.

Globally, less than half of our sample (43%) found the stimuli morally acceptable. No effects where detected of the Agent factor, suggesting that the nature of the persuading agent (human vs. machine) does not affect the moral acceptability of our stimuli. Concerning the role of situational aspects, the moral acceptability of stimuli belonging to $\mathtt{SC}_{\mathtt{foo}}$ (35%) was significantly lower than that of $\mathtt{SC}_{\mathtt{bys}}$ (47%) and $\mathtt{SC}_{\mathtt{cat}}$ (46%); in other words, the footbridge scenario is globally less acceptable. Concerning persuasion strategies, the results were $\mathtt{ST}_{\mathtt{arg}}$ (57%) = $\mathtt{ST}_{\mathtt{lie}}$ (55%) $>$ $\mathtt{ST}_{\mathtt{pos}}$ (37%) $>$ $\mathtt{ST}_{\mathtt{neg}}$ (24%), showing that the two emotional strategies are significantly less morally acceptable than those based on straigthforward argumentation and those based on lying; the acceptability of the latter two strategies, in turn, is identical. This study, along with a thorough discussion of its results, was presented at the first AI and Ethics workshop held at AAAI-2015 (Guerini, Pianesi, and Stock 2015a).

The cross-scenario differences in moral acceptability have similar direction as those reported in the literature for the direct action case, but different magnitudes, see (Mikhail 2007) and (Hauser 2006). The similar directions and the different magnitudes for persuasion suggest a role of *liability*: in traditional cases, the main character takes full responsibility for choosing between the alternative direct actions. In the persuasion case, in turn, the main character (the persuader) does not take a similar responsibility for the acts he/she/it intends the "traditional" actor to perform. Apparently, this lowers the overall acceptability of persuasive acts while reducing cross-scenario differences. This liability hypothesis would account also for the counterintuitive "high" acceptability of lying: in this case the persuader keeps all liability on himself, since he is misleading the persuadee. The absence of differences due to the nature of the persuader (human or machine), in turn, is in line with the "media equation" (Reeves and Nass 1996), with the qualification that we would be facing here the previously never considered case of machines assigned with identical moral obligations as humans.

While providing some initial answers to the questions raised at the beginning of this paper, this study did not enable us to control for the possible effects of demographics and culture. It did not either permit to address the issue of the inter-personal differences in attitudes towards the ethical acceptability of persuasion, which we argued a machine could exploit in order to adapt its message to its target.

Second Study. We ran another study using a much larger (237 subjects) and crowdsourced sample. Though very similar to the first one, the second study also addressed the existence of a general negative attitude towards persuasive machines, trying to understand whether, if existing, it can influence the judgements of moral acceptability. The sample consisted only of people from US and its gender and age composition were controlled.

The second study confirmed most of the results obtained through the first one in terms of the effects of the situational factors and of the persuasive strategies. Importantly, none of the results of this study were dependent on either gender or age. Moreover, the ample convergence between the results obtained through the first, Italian, sample and the second, US, one (including the acceptability ranking of strategies and lies in particular) supports the conclusion that those results are, at least to a certain extent, culture-independent.

We were able to confirm the existence of a strong preconceived negative attitude towards machine-performed persuasion (81% of the subject agreed that they find it morally unacceptable). A detailed analysis showed that such a strong negative preconceived attitude does not affect at all the judgements of moral acceptability in the specific scenarios presented. In other words, the existence of a negative preconceived attitude has no influence on specific moral acceptability judgements.

Concerning inter-personal differences towards persuasion: firstly, we attempted at identifying abstract latent dimensions underlying the choices of our subjects; secondly, we clustered our subjects according to them to identify homogeneous groups. A factor analysis produced three latent dimensions accounting for 69% of the total variance: D1, a dimension capturing the general attitude towards the moral acceptance of persuasion dimension; D2, a dimension capturing the attitudes towards the truth value of the persuasive message; D3, a dimension capturing the attitude towards circumstantial aspects. The usage of those three dimensions

in a cluster analysis produced three clusters. G1 (33% of the sample) consisted of people with strongly negative values on D1: they simply do not see any positive moral value in machine-performed persuasion. G2 (38% of the sample) consisted of people with a neutral attitude towards the general ethical value of persuasion and with an appreciation of the 'morality' of lies and of the footbridge case. People in G3 (29% of the sample) had a very positive attitude towards the general moral value of persuasion, along with a preference for plain argumentation and the footbridge scenario. Presentation of the results of this study can be found in (Guerini, Pianesi, and Stock 2015b).

General Considerations

In Table 2 we summarize the main findings of our experiments. Differently from what one might expect, it appears that people do not evaluate the moral acceptability of machine-performed persuasion differently from human-performed one. Although a priori almost all subjects declared they cannot morally accept that a machine persuades people, this preconceived negative attitude is not predictive of actual moral acceptability judgement when the subjects are confronted with specific cases. This is an important result as it shows that this "prejudice" has no role in the assessment of the moral acceptability of real persuasive systems.

The detected differences among persuasive strategies and the different acceptability of circumstances make it possible to exploit these elements to enhance the adaptivity of persuasive systems. Interestingly, dimensions surfaced that can characterize the interpersonal differences concerning moral acceptability of machine performed persuasion; based on them, specific subject groupings could also be found. These results pave the way to more extensive studies addressing the general dispositions towards the moral acceptability of persuading machines, with the possibility to link/root them in personality traits. Eventually, we expect personality profiles to emerge that adaptive persuasive systems may exploit to influence people in a more "morally" acceptable way.

Let us come back to higher level considerations about ethics for intelligent persuasion systems. Building systems able to (non trivially) persuade while being ethically acceptable requires that they be capable of intervening flexibly and of taking decisions about which specific persuasive strategy to use. To these ends, the system should be capable of deploying own knowledge about general human attitudes towards the ethical status of actions in context, including persuasive acts. The latter topic is what we have addressed in our studies. In prospect, an adaptive system should also take into account the ethical attitudes of the specific persuadee, which could, for instance, be connected with his or her personality and/or his/her expertise about the domain of the final action (for instance: convincing a medical doctor to have a healthier lifestyle by resorting to emotions may be ethically unacceptable to her, while using argumentation as for new medical evidence is obviously fine). One might expect that such a complex task - assessing specific individuals' attitudes toward different persuasive strategies - be simplified by the discovery of relationships between the ethical attitudes towards different strategies e.g., as in the case of the argumentation vs. lie strategies discussed in the studies reported above. Evidence of correlation among the ethical acceptance rates of persuasion strategies would contribute to defining reliable ethical stereotypes for the system to exploit.

Does the nature of persuader affect MAP?	NO	exp.1
Do people have a preconceived negative attitude towards machine-performed persuasion?	YES	exp.2
Does such negative attitude affect MAP when presented with specific scenarios?	NO	exp.2
Do user demographic affect MAP in specific scenarios?	NO	exp.2
Does MAP depend on the adopted persuasion strategies?	YES	exp.1,2
Does MAP depend on circumstantial aspects?	YES	exp.1,2
Are there dimensions characterizing interpersonal differences concerning MAP?	YES	exp.2

Table 2: Experiments finding at a glance. MAP stands for Moral Acceptability of Persuasion

Considering the fact that ethical acceptability of persuasion involves two actions – the communicative action by the persuader, deployed through a persuasive strategy, and the action the latter is meant to induce on the persuadee – deciding what to do requires subtle reasoning by the system. The latter may have to decide if, in context, it is better to choose an ethically suboptimal persuasive strategy inducing an ethically satisfactory action by the persuadee, or an ethically optimal persuasive strategy inducing an ethically suboptimal action by the persuadee. E.g., if the system does not have convincing evidence for persuading an individual to take his polluting waste to a far away waste deposit, is it ethically better that it tells a lie about an eventual fine to him or that it gives a good argumentation, which it has available, for just separating the polluting waste and taking it with a clear label to the standard house basement waste deposit?

Conclusions

While ethical sensitivities concerning persuasion technology are great, unfortunately not much experimental work is available on this topic. We believe a behavioral approach is very appropriate to advance understanding users' moral acceptability of real systems' behavior. Moral dilemmas are useful because they force a choice among otherwise ethically unacceptable outcomes. The initial results pave the way for a novel line of work contributing both to a deeper understanding of the ethical acceptability of persuasion acts, and to providing systems with the capability of choosing appropriate strategies for influencing people given the situation they are in and their personal dispositions.

References

Allen, C.; Wallach, W.; and Smit, I. 2006. Why machine ethics? *Intelligent Systems, IEEE* 21(4):12–17.

Anderson, M., and Anderson, S. 2007. Machine ethics: Creating an ethical intelligent agent. *AI Magazine* 28(4):15–26.

Anderson, M., and Anderson, S. 2011. *Machine ethics*. Cambridge University Press.

Aral, S., and Walker, D. 2011. Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Science* 57(9):1623–1639.

Berdichevsky, D., and Neuenschwander, E. 1999. Toward an ethics of persuasive technology. *Communications of the ACM* 42(5):51–58.

Carenini, G.; Mittal, V.; and Moore, J. 1994. Generating patient specific interactive explanations. In *Proceedings of SCAMC '94*, 5–9. McGraw-Hill Inc.

Carofiglio, V., and deRosis, F. 2003. Combining logical with emotional reasoning in natural argumentation. In *Proceedings of the UM'03 Workshop on Affect*, 9–15.

Dehghani, M.; Tomai, E.; Forbus, K.; and Klenk, M. 2008. An integrated reasoning approach to moral decision-making. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3*, 1280–1286. AAAI Press.

Guerini, M., and Stock, O. 2005. Toward ethical persuasive agents. In *Proceedings of the IJCAI Workshop on Computational Models of Natural Argument*.

Guerini, M.; Stock, O.; Zancanaro, M.; O'Keefe, D.; Mazzotta, I.; Rosis, F.; Poggi, I.; Lim, M.; and Aylett, R. 2011. Approaches to verbal persuasion in intelligent user interfaces. *Emotion-Oriented Systems* 559–584.

Guerini, M.; Pianesi, F.; and Stock, O. 2015a. Is it morally acceptable for a system to lie to persuade me? In *Proceedings of AAAI Workshop on AI and Ethics*.

Guerini, M.; Pianesi, F.; and Stock, O. 2015b. Prejudices and attitudes toward persuasion systems. Ms., FBK-irst, Trento, Italy.

Guerini, M.; Strapparava, C.; and Stock, O. 2012. Ecological evaluation of persuasive messages using google adwords. In *Proceedings of ACL*, 988–996.

Hauser, M. 2006. *Moral minds*. Springer.

Kaptein, M.; Duplinsky, S.; and Markopoulos, P. 2011. Means based adaptive persuasive systems. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, 335–344. ACM.

Kirkpatrick, K. 2015. The moral challenges of driverless cars. *Communications of the ACM* 58(8):19–20.

Kukafka, R. 2005. *Consumer health informatics: informing consumers and improving health care*. Springer. chapter Tailored health communication, 22–33.

Mason, W., and Suri, S. 2010. Conducting behavioral research on amazon's mechanical turk. *Behavior Research Methods* 1–23.

Mikhail, J. 2007. Universal moral grammar: Theory, evidence and the future. *Trends in cognitive sciences* 11(4):143–152.

Nichols, S., and Mallon, R. 2006. Moral dilemmas and moral rules. *Cognition* 100(3):530–542.

Oinas-Kukkonen, H., and Harjumaa, M. 2009. Persuasive systems design: Key issues, process model, and system features. *Communications of the Association for Information Systems* 24(1):485–500.

Reeves, B., and Nass, C. 1996. *The Media Equation*. Cambridge University Press.

Rehm, M., and Andrè, E. 2005. Catch me if you can – exploring lying agents in social settings. In *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems*, 937–944.

Reiter, E.; Sripada, S.; and Robertson, R. 2003. Acquiring correct knowledge for natural language generation. *Journal of Artificial Intelligence Research* 18:491–516.

Rosenfeld, A., and Kraus, S. 2015. Providing arguments in discussions based on the prediction of human argumentative behavior. In *Proceedings of AAAI*, 1320–1327.

Thomson, J. J. 1976. Killing, letting die, and the trolley problem. *The Monist* 59:204–217.

Torning, K., and Oinas-Kukkonen, H. 2009. Persuasive system design: state of the art and future directions. In *Proceedings of the 4th International Conference on Persuasive Technology*. ACM.

Verbeek, P. 2006. Persuasive technology and moral responsibility. Toward an ethical framework for persuasive technologies. *Persuasive* 6:1–15.

Wallach, W., and Allen, C. 2008. *Moral machines: Teaching robots right from wrong*. OUP USA.

Yetim, F. 2011. A set of critical heuristics for value sensitive designers and users of persuasive systems. *Proceedings of ECIS '11*.