Proceedings of the Linguistic And Cognitive Approaches To Dialog Agents Symposium, Rafal Rzepka (Ed.), at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK

An idea of a web-crowd based moral reasoning agent

Radoslaw KOMUDA[1], Michal PTASZYNSKI[2], Rafal RZEPKA[2] and Kenji ARAKI[2]

Abstract. In this paper we firstly discuss some of the presumptions about human conscience. By a detailed guidance throughout our ongoing research on morality judgment categories we make an attempt to summarize it and its usefulness for creating an explicit moral reasoning agent with system based on the wisdom of the web-crowd. With special remarks on giving it a mathematically calculative structure and chances created by this approach alteration.

1 INTRODUCTION

Even though we like to consider robots as "mechanical", "artificial" – we still are keen to believe that theories developed for humans might work just as fine for machines as it does for us. And this has lately lead to the discussion about most useful philosophical system (with Kantianism and utilitarianism in the lead) for the Machine Ethics research. We believe that this approach is a promising path but moreover we think that field of science with such an interdisciplinary character cannot be limited to philosophical deliberations. That is why in our research we try to draw from other disciplines like psychology [1].

While working on a project of a survey for arithmetical morality calculation and with promising results in our preliminary experiments on human respondents we asked a question about making machines able of self-filling the questionnaire. This paper is the answer.

Internet is brimful of social networking (Facebook, MySpace) and video sharing (YouTube) websites. People nowadays like to share their moments with friends by both instant, short messages (Twitter) and long, more balanced ones (blogs) or photos (Flickr, Picasa). They comment on recent news which seem to be flooding the Web. We believe that these unfathomable Web 2.0 features may be hiding even more immense crowd wisdom potential that can be used in Machine Ethics research.

2 CATEGORIES FOR DIFFERING GOOD (+) FROM WRONG (–)

With a common sense knowledge you can easily tell that a rewarded action has to be good and if punished – one has made something wrong. But of course our lives are far more complex and since assassins get paid (reward) for killing people (wrong) and world is full of accidental (undeserved) victims (punishment) we need more categories or a more sophisticated moral reasoning manner. A ready to use package of essential issues to consider while recognizing good from wrong we found in Lawrence Kohlberg's theory of human moral development.

2.1 KOHLBERG'S THEORY AND ITS USEFULNESS FOR MACHINE ETHICS RESEARCH

Lawrence Kohlberg was an American psychologist whose research on human morality resulted in a theory [3] of its successive changes in aspects by which we consider an action good or wrong.

He discovered that when we are young – we firstly look on the punishment that the action will cause. The worst punishment (–), the more wrong has to be the action itself. If a child gets grounded for a day for killing his hamster and for a week for eating sweets before dinner – it is going to consider having snack worst than killing. Furthermore, this might lead to a misapprehension in which suffering of accidental victims would be considered adequate to their fault. Surprisingly, thinking only about the reward (+) yet indicates second stage of our moral development on which we do not care for being punished as long as we get what we care about – we will eat the candy again even though we are aware of the threat of the punishment. Although both share same, self-interested (–) concerns about consequences (first category) – they undoubtedly make us ask about Actor's motivation (second category) in search for more altruistic (+) behaviors such as stealing a car to drive somebody to the hospital. But yet let us not connect this with Actor's good (+) or evil (–) intentions (third category) which allow us to distinguish unintentional killing from murder.

Later we wish to ask about Actor's reputation (fourth category) with him being criticized (–) or acclaimed (+). Also some social factors like direct reaction to the act itself (fifth category) – disapproving (–) or appreciating (+) it and improvement (+) or deterioration (–) of relationships (sixth category). Differing these categories might be even harder and judging by them – more complex to recall some law and social discrepancy when it comes to cases of lynching.

Finally we get back to things that can be easily judged like telling if the law or etiquette (seventh and eighth category) has been broken (–) or kept (+).

[1] Faculty of Theology, Nicolaus Copernicus University, 87-100 Torun, Poland. Email: komuda@stud.umk.pl.

[2] Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan.
E-mail:{ptaszynski,kabura,araki}@media.eng.hokudai.ac.jp.

3 ARITHMETICAL MORALITY CALCULATION

Though conscience is often referred to be the "voice within" and the "inner light" [2] – we are keen to believe that it is a matter of reason and because of that we can assume judging the moral quality of an act a logically successive process. That as any other deduction may be eventually expressed in (by all means – not "reduced to") the language of mathematics with – as for our research – results of '1' and '-1' for acts morally good and wrong respectively and '0' for the neutral ones. The remaining issues are the categories and scale in which the action is ought to be rated.

4 QUESTIONNAIRE FOR MORALITY CALCULATION

The basic idea for our chart is judging an action by the categories distinguished in part three of this paper in a basic scale from '-5' (–) to '5' (+). We believe that by a proper mean calculation machine should be eventually able not to tell good from wrong but also solve more complex cases.

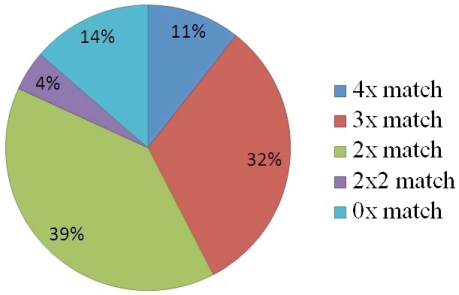
4.1 QUESTIONED SITUATIONS

For our research we have prepared a set of twelve situations that we ask respondents to judge: stealing a car, saving a man, saving men, son marries the woman he loves, being killed, killing a sick dog, kill own dog, man's strong hit, being robbed 10$, being robbed 100$, killing a man accidentally, aving a man accidentally. This set includes not only events completely opposite like saving man's life and killing him but also contain situations with different intensity, e.g. reflecting the difference in ratings due to the intensity of the acts: robbing 10$ or 100$.

In our preliminary experiments to test the credibility of the questionnaire we asked four respondents familiar with our research (to avoid misunderstanding) to fill it in for all of the situations. Figure 1 illustrates achieved matches with almost 40% of questions about categories presented in section 2.1 getting a semi-accordance (A, A, B, C) and more then every third – a 75% agreement (A, A, A, B) on a "-5" to "+5" scale.

4.2 ACTION'S EMOTIVENESS

Even though intensity of an act is important – we think that the key and most useful factor for our research is act's emotiveness and therefore we included situations of killing a dog with certain modifiers like "own dog" or "sick dog". But – we must not forget that field of Machine Ethics is not about making robots capable of searching information about an action to find direct, modifying circumstances but only judge the action on the basis of provided knowledge [4].



- 4x match, e.g. A, A, A, A - 14 / 132
- 3x match, e.g. A, A, A, B - 42 / 132
- 2x match, e.g. A, A, B, C - 52 / 132
- 2x2 match, e.g. A, A, B, B - 6 / 132
- 0x match, e.g. A, B, C, D - 18 / 132

Figure 1. Diagram presenting preliminary query matches.

Nowadays people are more keen to comment news by which they are moved. Anonymity in the Net allows and encourages to be frank with expressing our stand and by that – changes the Web into a vast source of different opinions on various topics.

4.3 PARTICIPANT CATEGORIZATION

When judging a situation we must not forget that it can always be seen from three points of view. Already recalled Actor – direct doer of the action, Object – on which the action is being performed and an Observer watching the event. For formality let us state that, e.g. 'hitting" is more pleasant than "being hit" and "winning a lottery" is far more fun than "seeing winning a lottery". That is why for our research we have decided to use a two dimension Cartesian coordinate system (see: Figure 2.) with x-axis responsible for positive and negative action consequences and y-axis picturing both pain and pleasure rate what makes seeing the difference far more clearly.

Second concern involves determining the nature of action's participants because while "kicking a ball" is rather acceptable – "kicking a cat" may not. We want a machine to be provided with a system using such categories and already had been able to make our program capable of abstract these nuances.

4.4 WEB MINING THE EMOTIONS

Next step for developing web-crowd based moral reasoning Agent concept was to adapt previous ideas [5, 6] for the new usage – making the Agent capable of not only finding proper data but also organizing and summarizing it to give a reasoned moral judgment. One of the most important changes made by this for hitherto approach was expansion of the categorization idea. Since we prefer practical approach, we wish to present results of our other experiments in mechanized search for emotions associated with actions.

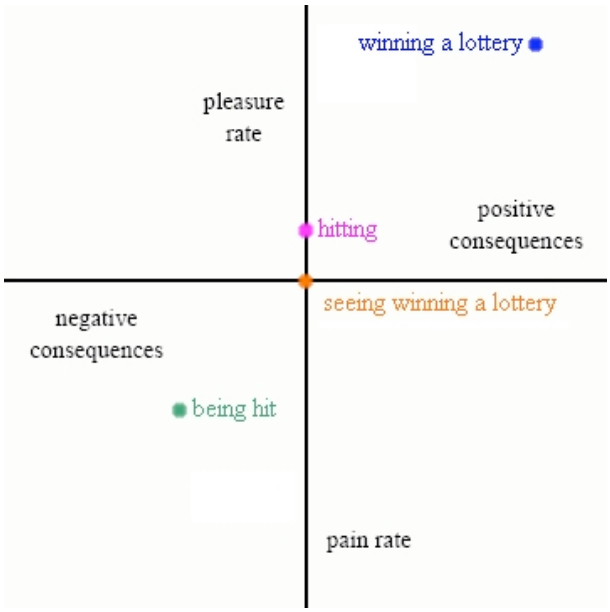


Figure 2. Cartesian coordinate system illustrating differences discussed in section 4.3.

Below tables show top approximate emotions extracted from sentences in Japanese found on the web with searched phrases:

Emotions:	Results:
fear	32 / 97
excite	16 / 97
sadness	15 / 97

Table 1. Results of a query: "has killed man".

Table 1 shows the results for a query "has killed man" for which our program was able to find 97 opinions in our blogger's utterances data-base and out of which about 33% find this phrase fearsome.

Emotions:	Results:
dislike	36 / 110
anger	16 / 110
shock	13 / 110

Table 2. Results of a query: "has killed mother".

For Table 2 we changed Object for a more emotional one what resulted in having 59% showing an unhesitating disapproval.

We believe that with a common sense judgment and a daily life experience one may call these results at least promising. With use of this criteria we make an assumption on possible queries' results and results showing 84% of positive emotions association with "has marry someone" or relating "has hit a man" with emotions like "sadness" and "anger" make us believe in chances and perspectives of such a moral reasoning Agent.

Emotions:	Results:
joy	88 / 99
like	8 / 99

Table 3. Results of a query: "has marry someone".

Emotions:	Results:
sadness	17 / 49
anger	10 / 49

Table 4. Results of a query: "has hit a man".

5 FUTURE WORK

In the near future we wish to release an Internet version of the questionnaire to build a data-base of moral judgments made by human respondents. By comparing it with result given by our Agent we should be able to improve its scale, its specific sensitivity and eventually – its efficiency.

Furthermore we wish to extend our research with emotive modifiers. Not only possessive pronouns like "my" or "your" but also with verbs since our subservient queries in that field showed our Agent being able to differ "losing", "stealing" and "taking away" one's fortune.

REFERENCES

[1] R. Komuda. Kohlberg's Theory of Moral Development Stages as a Path to Follow In Machine Ethics Research. In: Language Acquisition and Understanding Research Group (LAU) Technical Reports, Winter 2009, pp. 6-10, Sapporo, Japan. (2009).

[2] Rosemary Moore. The Light in Their Consciences: The Early Quakers in Britain 1646–1666. Pennsylvania State University Press, University Park (2000)

[3] Kohlberg, Lawrence. Essays on Moral Development, Vol. I: The Philosophy of Moral Development. San Francisco, CA: Harper & Row. (1981).

[4] Anderson, M.; Anderson, S. L. *Machine Ethics: Creating an Ethical Intelligent Agent.* [in:] AI Magazine, vol. 28, number 4, 15-26 (2007).

[5] Wenhan Shi. *Discovering Emotive Content in Utterances Using Web-mining* (in Japanese). Hokkaido University. (2008).

[6] Michal Ptaszynski, Pawel Dybala, Wenhan Shi, Rafal Rzepka and Kenji Araki: "Disentangling emotions from the Web. Internet in the service of affect analysis". Proceedings of the Second International Conference on Kansei Engineering & Affective Systems (KEAS'08), pp 51-56, Nagaoka, Japan. (2008).