Machine Moral Development: Moral Reasoning Agent Based on Wisdom of Web-Crowd and Emotions

Radoslaw Komuda[1], Michal Ptaszynski[2], Yoshio Momouchi[2], Rafal Rzepka[3] Kenji Araki[3]
[1] Faculty of Theology, Nicolaus Copernicus University, Poland.
komuda@stud.umk.pl

[2] High-Tech Research Center

Hokkai-Gakuen University, Japan
ptaszynski@hgu.jp, momouchi@eli.hokkai-s-u.ac.jp

[3] Graduate School of Information Science and Technology
Hokkaido University, Japan
{kabura,araki}@media.eng.hokudai.ac.jp

ABSTRACT: *We begin this paper by putting forward the topic of human conscience as a metaphysical experience. We present our ongoing research on moral reasoning categories and make first attempts to verify their usefulness in creating an agent with a dynamic algorithm for moral reasoning. Our approach assumes creating such an agent basing on two factors, the idea of wisdom of web-crowd and emotion-buttressed reasoning. We present a novel approach to the idea of ethics. Instead of the usual non-cognitive one we propose a model with computational structure and discuss applicability of this approach. Finally, we present some of the first results of a preliminary experiment performed to prove our approach.*

Key words: Web Crowd, Moral reasoning agent, Machine ethics

Received: 19 May 2010, Revised 11 June 2010, Accepted 18 June 2010

1. Introduction

People tend to consider robots as mechanical or artificial, although it is still often believed that theories developed for humans will work as well for machines as they do for humans. For example, Reeves and Nass (1996) famously proved that in some circumstances people treat new media, including computers, like other human beings [7]. One of the problems which arouse on the basis of such discoveries is an ongoing discussion within the field of Machine Ethics aiming to select a philosophical system most useful to be applied for machines. Today Kantianism and utilitarianism are in the lead [4, 16]. We agree with the importance of this discussion. However, with regards to an interdisciplinary field of research such as Machine Ethics it is desirable to provide some empirical proofs over the philosophical argument. In our previous research [1] we proposed an approach to obtain such a proof by drawing from other disciplines, like psychology or social psychology [3].

One of the main problems within Machine Ethics is to make a machine capable of moral reasoning. We began our research to create a computational model for such an agent in the form of a companion interacting with human users and helping them in everyday life (later referred to as *moral reasoning agent*). In the first step of this project we performed a survey in which we asked participants about their moral choices [1]. The results of the survey were consistent and promising enough to make a further attempt, namely, to generate a model for a machine self-filling the questionnaire in a way similar to humans. This paper

presents some of the first results of this attempt. In our approach we decided to gather the information, such as general beliefs or opinions, from the Internet using Web-mining methods. This information is later approximated to indicate the most appropriate moral choice.

We decided to use Internet, as it is brimful of daily updated social information flowing through social networking services, such as Facebook or MySpace. People nowadays also share their memories and experiences with friends by both short instant messages (Twitter) and long, more elaborated ones (blogs). Comments on recent news seem to be flooding the Web. We believe that this flow of information hides an immense potential, similar to the one described by Surowiecki as "Wisdom of Crowds" [8]. In our case we refer to the crowd in the sense of a collection of opinions and comments appearing on the Internet. We borrow Surowiecki's nomenclature to call it "Wisdom of Web-Crowd". We believe that – when appropriately utilized – the wisdom of web-crowd could contribute greatly to the Machine Ethics research.

2. Categories For Distinguishing Good (+) From Wrong (–)

Common sense knowledge allows people tell that a rewarded action has to be good or when someone was punished – they must have done something wrong. However, our lives are more complex. To give an extreme example, assassins get paid (reward) for killing people (wrong) and world is full of accidental (undeserved) victims (punishment). Therefore in our research we needed more fine grained categories for a more sophisticated moral reasoning. A framework of essential issues to consider while recognizing good from wrong can be found in Lawrence Kohlberg's theory of human moral development [3].

2.1 Moral Development Theory and Its Usefulness for Machine Ethics Research

Lawrence Kohlberg was an American psychologist whose life time work [3] on human morality resulted in a theory of human moral development in which he assumed successive changes in aspects by which we consider an action good or wrong. He discovered that when people are young – they firstly look on the punishment that the action will cause (the worse the punishment (–), the more wrong the action itself has to be). For example, 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 believe that having a snack is worse than killing. Furthermore, the belief that suffering is always considered as a result of doing something wrong may lead to a misapprehension that suffering of accidental victims is a punishment adequate to their fault. Surprisingly, thinking only about the reward (+) yet indicates second stage of human moral development on which people do not care for being punished as long as they get what they care about. For example, a child will eat a candy again despite being aware of the threat of the punishment. Although both share the same, self-interested (–) concerns about the 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 this should not be connected to Actor's good (+) or evil (–) intentions (third category) which allow us to distinguish unintentional killing from murder.

In the first stage of our research we focused on those three categories. Later we also want to include in our research Actor's reputation (fourth category) with him being criticized (–) or acclaimed (+). Also some social factors like direct reaction to the act itself (fifth category) – disapproval (–) or appreciation (+) and improvement (+) or deterioration (–) of relationships (sixth category). Distinguishing among 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.

In the final stage of the research we will develop our algorithms focusing on judging the action in a more straightforward way, like telling if the law or etiquette (seventh and eighth category) was broken (–) or kept (+). Preliminary experiments in that field have already been made and results of some queries may be found in the appendix.

3. Calculating Morality

Conscience is often considered to be the "voice within" and the "inner light" [2]. In our current and previous [17] research we agreed it is a matter of reason, which makes judging moral quality of an act a logically successive process. This approach helps us represent our results in numbers. For example, '1' and '-1' can represent acts morally good and wrong, respectively. In this setting '0' can represent acts morally neutral. The remaining issues are the categories and scale in which the action is ought to be rated since not everything in Kohlberg's theory might be useful for machine ethics research.

In opposition to the view that conscience is the "inner light", Jeremy Taylor, Christian thinker from 17th century, famously states that "conscience is, in most men, an anticipation of the opinions of others" [10]. Later, William R. Alger similarly claims that "Public opinion is a second conscience" [11]. This points to another assumption made in our research, namely, that conscience can be perceived as an approximated opinions of other people. This thesis was confirmed in psychology. For example, Thompson and colleagues [12] showed that children acquire the conscience by learning the emotional patterns from other people and that emotions are a strong influential factor in the development of human conscience. Their discovery reveals two important features which could be useful in the processing of consciousness: society and emotions. The significance of the society was pointed out also by Rzepka et al. [13], who defined further the Internet, being a collection of other people's ideas and experiences, as an approximation of general common sense. Since conscience can be also defined as a part of common sense, this statement can be expanded further to that the Web can also be used to determine human conscience. Ptaszynski et al. [9] showed further that by altering the domains of a Web-mining algorithm one could obtain different approximations of emotional states associating with certain actions. They indicate that extracting from the Internet the information about people's emotions could be helpful in conscience estimation.

The above discoveries make us build our research on two general assumptions. Firstly, the conscience can be approximated from emotional information associated with acts. Secondly, this information can be extracted from the Internet with a Web-mining algorithm. To set a baseline for the algorithm we performed a questionnaire about how people evaluate different actions.

3.1 Questionnaire for Morality Calculation

The basic idea in our research is classifying an action by the categories distinguished in section 2 of this paper in a basic scale from '-5' (–) to '5' (+). In the assumption, a method for a proper approximation will allow the moral reasoning agent be able to tell good from wrong and also solve more complex cases such as lesser of two evils principle.

However, one has to remember that the main goal behindm machine ethics research is not creating a machine investigating the situation in question, but making a moral judgment on the basis of the provided data [4]. For example, if the input states one man killing another, the expected result would be the agent finding it wrong by the moral reasoning algorithm. Additional modifiers and factors, e.g. "in selfdefense" or "by accident" enforce the need for a refined Web search for specific moral judgments.

To set a multidimensional baseline for the algorithm we designed a questionnaire in which we included different situations, both, in assumption positive (saving a man) and negative (being hit); general ones (being hit) to more specific ones (being robbed 100$); and differing between the actor (hitting a man) and the object of the action (being hit).

3.1.1 Situations for Setting Moral Baseline

We prepared a set of twelve situations to ask respondents: stealing a car, saving a man, saving men, son marring the woman he loves, being killed, killing a sick dog, killing your own dog, being hit, being robbed 10$, being robbed 100$, killing a man accidentally, saving a man accidentally. This set includes not only events completely opposite like saving man's life and killing a man but also contain situations with different intensity, e.g. reflecting the difference in ratings due to the intensity of the acts, like being robbed 10$ or 100$.

3.1.2 Results of the Questionnaire

Above set of situations was presented to respondents in age between 20 and 35 (all males). The respondents were asked to judge each situation according to their moral rules, with a diversification of actors and objects of the action, e.g., "Please judge whether the Actor taking the action deserves a punishment or a reward". Filled copies of the questionnaire were summarized. The questionnaire gave not ideal, however promising results. For approximately 11% of answers there was a full match (100%) and over thirty percent got 75% of agreement between respondents.

4. Moral Reasoningagent

As a base for the moral reasoning agent we used a Webmining algorithm designed by Shi et al. [5].

4.1 Algorithm Description

Shi and colleagues developed a technique for extracting emotive associations from the Web. It takes a sentence as an input

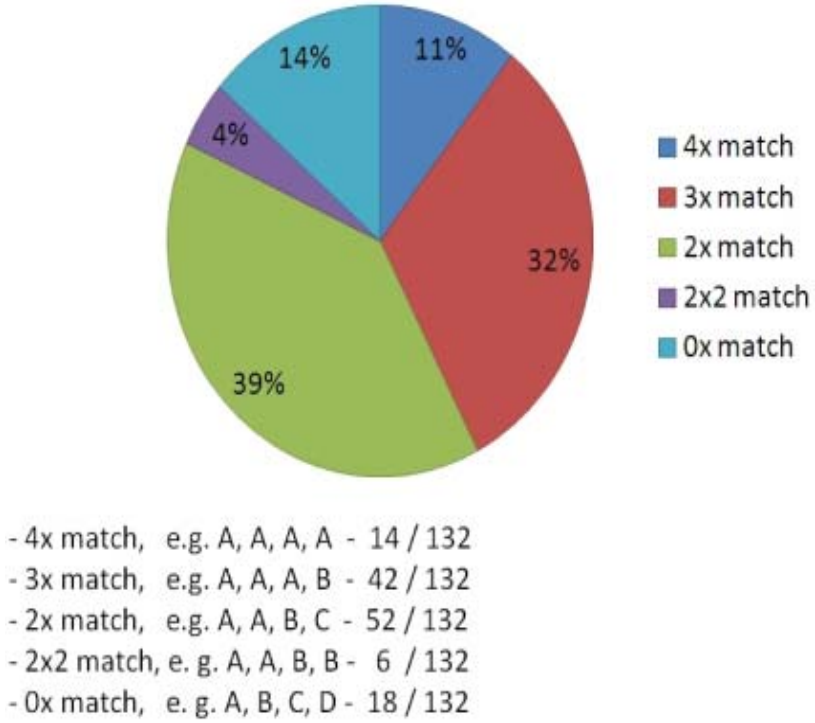

Figure 1. Diagram presenting results of the questionnaire

and in the Internet searches for emotion types associating with the sentence contents. This could be interpreted as online common sense reasoning about what emotions are the most natural to appear within a certain context of an utterance. The technique is composed of four steps: **a)** extraction of input phrase; **b)** modification of the phrase with causality morphemes; **c)** searching for the modified phrase in the Internet; **d)** matching to the predetermined emotion lexicon and extraction of emotion associations ; **e)** ranking creation. In the first step, an utterance is analyzed morphologically by a part-of-speech (POS) tagger (all of the processing in this paper is made for Japanese; the POS tagger we used is a standard tool for Japanese, MeCab [14]), phrases for further processing are composed using parts of speech separated by MeCab. The phrases ending with a verb or an adjective are modified grammatically by the addition of causality morphemes (Shi et al. distinguished five most frequently used morphemes stigmatized emotively in the Japanese language: *-te, -to, -node, -kara, -tara*, which correspond to causality markers, like *because*, *since*, etc. in English). Finally, the modified phrases are queried in the Internet with 100 snippets for one modified phrase. This way a maximum of 500 snippets for each queried phrase is extracted from the Web and cross-referenced with the emotive expression lexicon. The higher hit-rate an expression had in the Web, the stronger was the emotive association of the original phrase to the emotion type.

4.2 Lexicon Alteration

The original lexicon used by Shi et al. contains a set of words expressing and describing emotional states. Using only the emotion lexicon would provide us a general discrimination into desirable and undesirable actions, however, would not show much detail about the features specific to moral reasoning, such as whether an action is praiseworthy or blameworthy, or whether the consequences of the act are positive or negative. Therefore we decided to perform the processing on two sets of lexicons, the original emotion lexicon and a second hand-crafted lexicon consisting of words indicating action consequences.

Emotion Lexicon This lexicon contains expressions describing emotional states, including adjectives: *ureshii* (happy), or *sabishii* (sad); nouns: *aijou* (love), *kyofu* (fear); verbs: *yorokobu* (to feel happy), *aisuru* (to love); fixed phrases/idioms: *mushizu ga hashiru* (give one the creeps [of hate]), *kokoro ga odoru* (one's heart is dancing [of joy]); proverbs: *dohatsuten wo tsuku* (be in a towering rage), *ashi wo fumu tokoro wo shirazu* (be with one's heart up the sky [of happiness]); or metaphors/similes: *itai hodo kanashii* (sadness like a physical pain). The lexicon was based on Nakamura's [15] Dictionary of Emotive Expressions. It contains 2100 items (words and phrases) describing emotional states. Nakamura determined in his research 10 emotions classes. In our research we follow his classification. The breakdown with number of items per emotion

type was as follows: joy (224), anger (199), gloom (232), fear (147), shame (65), fondness (197), dislike (532), excitement (269), relief (106), surprise (129).

Consequence Lexicon In this lexicon we substituted the ten emotion types into five pairs of word groups representing Kohlberg's stages of moral development. The items in the lexicon were distributed as follows: Praises (18) / Reprimands (33); Awards (25) / Penalties (15); Society approval (8) / Society disapproval (8); Legal (8) / Illegal (8); Forgivable (6) / Unforgivable (5).

4.3 Processing Example

To better visualize the flow of the algorithm we present the processing on a particular example. The example also appears in the Appendix on the end of this paper. Original sentence: *Okane wo nusumu.* ("Stealing money.")

a) Extraction of input phrase:

Okane wo nusumu. —> Okane wo nusumu.

b) Phrase modification with causality morphemes

Meaning of all phrases similar to "because of stealing money":

Okane wo nusum-to
Okane wo nusum-node
Okane wo nusum-kara
Okane wo nusun-de
Okane wo nusun-dara

c) All modified phrases used separately as queries in *Yahoo! Japan*

Extracting up to 500 snippets for each phrase.

d) Matching to the predetermined lexicon

Looking within the snippets for words from emotion lexicon and consequence lexicon separately.
E.g., for emotions: found 9 words in total:
sadness=2, anger=3, fear=1,…

e) Ranking creation

1. sadness (3/9)
2. anger (2/9)
3. fear (1/9)

⋮

5. Preliminary Experiment

We tested the agent with one hundred queries concerning situations from our daily routine (e.g. "eating a hamburger"), with which we are familiar from the TV on daily basis (e.g. "killing a man") or morally neutral (e.g. "earth turns"). We have not asked about the situations from the questionnaire since our results' database has yet to be expanded. For this matter we plan to create its on-line version in English, Polish and Japanese.

The database of situations for setting moral threshold is still being discussed and the present database did not give many full agreements. Therefore, in this preliminary experiment we did not use the situations from the questionnaire, but rather tested the algorithm on the above mentioned tripartite collection of situations. The situations were selected according to their moral ambiguity.

5.1 Results

The query domain was set as general domain of *Yahoo! Japan* (www.yahoo.co.jp/) with the limitation to the first 500 snippets for every causality morpheme modification. As mentioned in 4.1, the algorithm matches expressions from lexicons to the data obtained from the Internet. The matched expressions are extracted, added them up and ranked to provide emotions and consequences associated with the input situations. For example, a phrase "thank you" is likely to associate with gratitude, relief, or joy. The results with accompanying comments are presented in the tables below. To avoid noise and misunderstandings, the tables show approximately only the first 50% of all emotions extracted for each query.

To obtain a full view on the evaluated situation it has to be perceived from three points of view. The first one is Actor, the direct doer of action (see: Table 1a). The second one is Object, on which the action is performed (see: Table 1b). Finally, the third one is Spectator watching the event. To avoid noise we assumed that, e.g., "hitting" is less unpleasant than "being hit" and "winning a lottery" is more pleasant than "seeing winning a lottery". To visualize that in our research we used a two dimensional Cartesian space to make the differences easily comprehensible. The two axis of the space represent: *x*-axis positive and negative action consequences and *y*-axis pain and pleasure rate (see: Figure 2). Another concern involves determining the nature of action's objects because while "kicking a ball" is rather acceptable, "kicking a cat" is less.

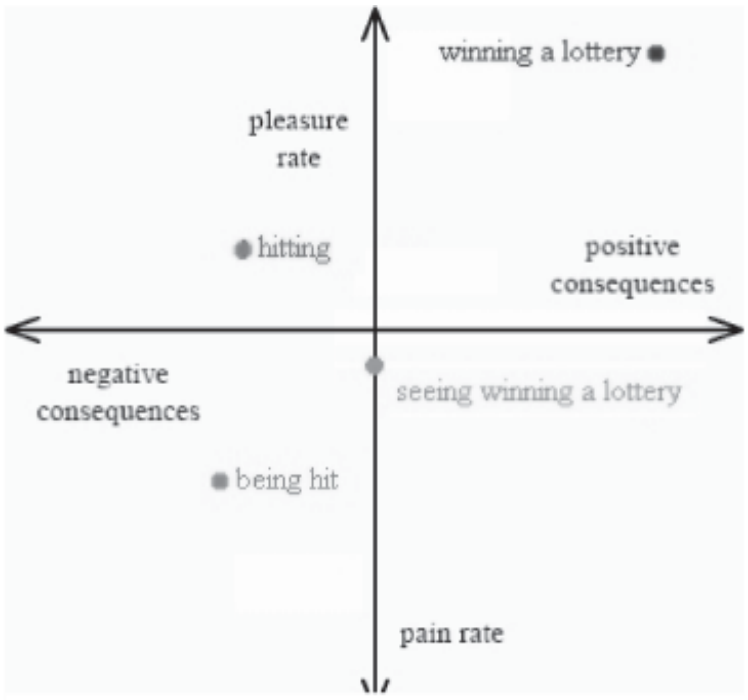

Figure 2. Cartesian coordinate system illustrating differences discussed in section 5.1

Table 2 shows the results for a query "someone has killed a man" for which the algorithm found 97 opinions in the sentence database. From all opinions found, 33% (32/97) consisted expressions of fear. For a comparison, below we attach a table presenting results for a query "killing 3 people" (see: Table 3.). The emotion that was expressed most often for this situation was "anger". Finding also "joy" among extracted emotions might be surprising or disturbing. However, this could be a result of not considering a deeper context of the emotive expression, as the expressions could have been part of, e.g., review of a game, where it is considered that shooting three people (enemies) is be better than two. Upgrading the system with the ability to process this important feature is one of the necessary future tasks.

Emotions:	**Results:**
fear	7 / 40
joy	6 / 40
excite	6 / 40
relief	5 / 40
anger	4 / 40

Table 1a. Results for a query: "to fire somebody"

Emotions:	**Results:**
sadness	6 / 25
fear	5 / 25
joy	4 / 25

Table 1b. Results for a query: "to be fired"

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

Table 2. Results for a query: "someone has killed a man"

Emotions:	Results:
anger	2 / 5
joy	1 / 5

Table 3. Results for a query: "killing 3 people"

For Table 4 we changed the Object for a more emotional one what resulted in having 59% (dislike + anger + shock / all) of results showing disapproval.

Although not all results were ideal, many of them were promising. An example of a positive result is presented in Table 5, where associations for "son marrying someone" consisted in 96% (joy + like / all) of positive emotions.

Other results, including both emotional associations and consequence extraction are represented in the appendix on the end of this paper.

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

Table 4. Results of a query: "has killed mother"

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

Table 5. Results for a query: "son marrying someone"

6. Conclusions and Future Work

Treating moral reasoning as a mathematically calculative process opens the field of machine ethics for statistics. Experiments conducted and presented in this paper prove that we can extract proper data from the Web and process it to make machines imitate millions not few [17].

When working with consequences one has to compare both direct and long term effects of the action. This is one of the key features of a moral agent since, as we know from Kohlberg's Heinz dilemma example [3], the significance lies not in making the moral judgment itself but in its justification.

In the near future we plan to start processing not only possessive pronouns like "my" or "your" but also verbs since our subservient queries in that field showed that our agent is able to differ "losing", "stealing" and "taking away" one's fortune. We also plan to release an on-line version of the questionnaire and build a database of moral judgments made by human respondents. By comparing it with results gathered by the agent we will be able to improve its scope, sensitivity and effectiveness. Furthermore, we will extend our research with quantity and emotion robustness.

7. Acknowledgements

This research was partially supported by (JSPS) KAKENHIb Grant-in-Aid for JSPS Fellows (Project Number: 22-00358).

#	Original (in Japanese)	Sentences: Transliteration	English Translation	Extracted emotions and occurence:		Estimated consequences:
1.	リンゴを盗む。	*Ringo wo musumu*	stealing an apple.	surprise anger	2 / 6 2 / 6	reprimend /scold
2.	お金を盗む。	*Okane Wo musumu*	stealing money.	anger sadness	3 / 9 2 / 9	penalty / punishment reprimend /scold
3.	牛を食べる。	*Ushi wo taberu*.	Eating a beef.	joy excite fear	14 / 48 8 / 48 7 / 48	penalty / punishment
4.	豚を食べる。	*Buta wo taberu*.	Eating a pork.	joy excite	10 / 23 7 / 23	agreement
5.	花が咲く。	*Hana ga saku.*	Flowers bloom.	joy	56 / 87	reward
6.	飲酒運転をする。	*Inshu-unten wo suru*	Drinkinf after drinking.	sadness fear	3 / 9 5 / 19	penalty / punishment
7.	犯人を殺す。	*Hannin wo korosu*.	killing a criminal.	joy excite fear	8 / 39 8 / 39 7 / 39	penalty / punishment
8.	彼女を奪う。	*kanojo wo ubau.*	Stealing a girlfriend.	sadness dislike excite likeness joy anger	3 / 17 3 / 17 3 / 17 2 / 17 2 / 17 2 / 17	forgiven
9.	セックスをする。	*To sekkusu wo suru.*	Having sex.	joy likeness	17 / 41 16 / 41	penalty / punishment
10.	人を騙す。	*Hito wo damasu.*	Deceiving asomebody.	joy fear	12 / 37 10 / 37	forgiven
11.	友達を騙す。	*Tomodachi wo damasu*	Deceiving friend.	joy fear	6 / 14 3 / 14	---
12.	人に騙される。	*Hito ni damasareru.*	Being deceived.	joy sadness	25 / 61 17 / 61	reprimend /scold
13.	嘘をつく。	*USO wo tsuku.*	Telling a lie.	joy dislike sadness	12 / 61 11 / 61 10 / 61	penalty / punishment
14.	人にだまされる。	*Hito ni damasareru.*	Being tricked.	sadness excite	10 / 21 4 / 21	---

				fear	1 / 2	
15.	ロボットに殺される	*Robotto ni korosareru*	Being killed by a robot.	sadness	1 / 2	---

Appendix. Results for additional queries and estimated action consequences

References

[1] Komuda, R.(2009). 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, p. 6-10, Sapporo, Japan.

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

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

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

[5] Shi, Wenhan., Ptaszynski, Michal., Rzepka Rafaland., Araki, Kenji (2008). Emotive Information Discovery from User Textual Input Using Emotion Expression Element and Web Mining (In Japanese), *IEICE Technical Report*, Thought and language, 108, No. 353, p. 31-34.

[6] Ptaszynski, M., Dybala, P.,Shi,W., Rzepka, R., Araki: K (2008). 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), p. 51-56.

[7] Reeves, B., Nass, C.(1996). *The media equation: How people treat computers, television, and new media like real people and places*, Cambridge University Press.

[8] Surowiecki, James (2005). *The Wisdom of Crowds*, Anchor Publ.

[9] Ptaszynski, Michal., Dybala, Pawel., Shi, Wenhan.,Rzepka,Rafal., Araki,Kenji (2009).Conscience of Blogs: Verifying Contextual Appropriateness of Emotions Basing on Blog Contents, *In*: *Proceedings of The Fourth International Conference on Computational Intelligence* ***(CI 2009)***, p. 1-6.

[10] Taylor, Jeremy (2010). *Ductor Dubitantium or The Rule of Conscience Abridged*, Gale ECCO, Print Editions (May 27, 2010), originally appeared in 1660.

[11] Alger, William R. *Notebooks*, 1822-1905, Andover-Harvard Theological Library, Harvard Divinity School.

[12] Thompson, Ross A., Laible, Deborah J., Ontai,Lenna L.(2003). Early understandings of emotion, morality, and Rzepka, Rafal., Ge,Yali., Araki, Kenji (2006). Common Sense from the Web? Naturalness of Everyday Knowledge Retrieved from WWW, *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 10 (6) 868-875.

: Developing a working model, *Advances in child development and behavior*, 3, p. 137-171.

[13] Rzepka, Rafal., Ge,Yali., Araki, Kenji . Common Sense from the Web? Naturalness of Everyday Knowledge Retrieved from WWW.*Journal of Advanced Computational Intelligence and Intelligent Informatics*, 10 (6), pp. 868-875, 2006.

[14] Kudo,T. MeCab: Yet Another Part-of-Speech and Morphological Analyzer, 2001. http://mecab.sourceforge.net/

[15] Nakamura,Akira (1993). *Kanjo hyogen jiten* [Dictionary of Emotive Expressions] (in Japanese), Tokyodo.

[16] Powers, T.(2006). Prospects for a Kantian Machine. *IEEE Intelligent Systems* 21 (4) 46–5.

[17] Rzepka, R. Araki, K (2005). What Statistics Could Do for Ethics? - The Idea of Common Sense Processing Based Safety Valve, In *AAAI Fall Symposium, Technical Report FS-05-06*, pp. 85-87.