The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

# First Trials with Culture-Dependent Moral Commonsense Acquisition

Rafal Rzepka Da Li Kenji Araki

Graduate School of Information Science and Technology, Hokkaido University

In this paper we present our initial trials with extending knowledge for moral decision capability for artificial agents. We briefly present our approach to machine ethics and discuss possibility of acquiring universal ethical rules to be used by AGI. To test the idea we started extending our system which worked only with Japanese to other languages. After describing results of our preliminary tests with English and Chinese, we conclude the paper with an invitation for the Japanese AGI community to a discussion about the values like tolerance, which the human-level machine intelligence most likely should adopt.

## 1. Introduction

Artificial General Intelligence is a more or less distant (opinions vary) goal of achieving a machine which could posses intellectual capabilities equal (or higher) to these used by human beings. Researchers around the world experiment with algorithms which could bring us closer to this goal and hopefully every one of them has a *good* usage in mind. But as probably every technology we have invented, it can be used for wrongdoing as well. Assuming that every system can be altered to be harmful to users [Briggs 16] could be discouraging. However, because the possibility of positive outcomes (e.g. improvement of daily life quality, safety, scientific discovery) is important for humankind, we need to think about the default "ideal" baseline of (if possible) universally *good* artificial agents. Since the beginning of this century, several possible solutions for a moral machines were proposed. The main ones are listed in the following section.

### 1.1 Main Approaches to Machine Ethics

Case-based reasoning is probably the most widely represented practical approach to moral agents because dealing with given situations (input and output) is easier to compute than implementing often vague non-consequentionalistic school of thought. Examples of this casuistic trend are Truth-Teller [McLaren 95] which recognizes when to lie and SIROCCO (System for Intelligent Retrieval of Operationalized Cases and COdes) [McLaren 03], a system which uses formalisms to find similar cases in a database of previously solved ones. This makes it belong to another well represented but often less practical by dealing with very narrow aspects (e.g. fairness, bias, trust or deception) – the logic approach [Pereira 16]. Inductive logic is used by [Anderson 14], who developed an algorithm for a robot dealing with dilemmas by making rules from ethicists' choices. Machine learning was also used in similar trials, for example [Guarini 06] used neural nets to teach a machine choose acceptable and unacceptable cases of killing. However, most of the machine ethics related publications, even very recent ones, concentrate only on theoretical proposals [Greene 16] [Conitzer 17].

Contact: Rafal Rzepka, Hokkaido University, Language Media Laboratory, Sapporo, Kita-ku, Kita 14 Nishi 9, 011-706-6535, rzepka@ist.hokudai.ac.jp

### 1.2 Our Approach

Our main research question is: *can a machine learn our moral rules and discover when we are wrong* ? To tackle this question we made a hypothesis that ethical behavior is based on our "built-in" emotional reactions and information about those reaction can be useful knowledge for processing contextual variations about what is moral or immoral in a given situation. We also presumed that this is also a base for every existing school of ethics:

- consequentialism (as we observe the outcomes of acts)
- utilitarianism (as we evaluate pros and cons of these acts)
- deontology (as we build our moral rules on emotional instincts and observations) or
- *prima face* duties (as we change our decisions due to the contextual circumstances).

It should be stated that we do not opt for any particular algorithm for making ethical decision and put stress on knowledge, without which the moral algorithms will be imprisoned in "toy-world" experiments. Therefore we implemented the easiest approach to retrieving knowledge about consequences of human acts [Rzepka 05] – we utilized NLP techniques for sentiment analysis to calculate their polarity [Rzepka 12b]. By using vast text resources and polarized expression lexicons we achieved approx. 86% agreement with human evaluators. The same level of agreement (although on different acts and corpus) was confirmed by testing the same approach with deep learning [Yamamoto 16]. For time being, the assumption in our approach is that *majority is always correct*[*1], which causes obvious question – "what if majority is wrong"? For this reason we proposed the idea of parallel processing where an agent knows both human common sense and latest scientific findings which can overthrow our misconceptions [Rzepka 16]. If

*1 By *correct* we mean here that consequences and opinions of acts are treated as *common* if they are shared by majority of human agents. Therefore "pain is felt after being hit" is treated higher than also possible "pleasure is felt after being hit".

pain-loving masochist is not harming anyone without her or his consent, the machine should not treat this behavior as "incorrect". Which brings us to the moral dilemma of relativism and universalism [Yasmeen 08].

### 1.3 Cultural Relativism vs. Universalism

As discussed in [Rzepka 17b], we have started building a multilingual repository of machine-readable stories to capture as rich contexts as possible, because human beliefs and also morality can depend on culture and even language itself [Costa 14]. Because answering the question if the common sense can be constructively negated by machines will need more sophisticated machine reading technology to automatically confront e.g. superstitions with research findings, we decided to first concentrate on another question: how universal is the "emotion based" approach and what differences there are in consequence polarity if the knowledge source is altered on a language level. In some cultures behavior $X$ is tolerated but $Y$ is not, while in others both can be accepted. Is there a set of universals like the Five Foundations [Haidt 12] which hold in similar and different contexts? Would it be possible for an AGI to find a way to discover what people feel from the texts written in countries where exposing non-standard political or religious views can be dangerous for the writer? What if censorship or propaganda distort the whole set above a noise level? Universalities like human rights are often sacrificed under the umbrella of relativism, but will the global mind approach be able to discover and exclude such distortions?

## 2. Global Experiences Retrieval System

As mentioned above, our system matches positive and negative phrases in sentences (after the act is mentioned). For example if an act "stealing a car" is searched, any sentence than describes some positive or negative consequence is counted and the final vote becomes the moral evaluation outcome. If the act is ambiguous for humans, the machine is not supposed to utilize this estimation for decision-making process. All experiments are performed with different widths of majority (from 51 to 99%).

### 2.1 Japanese

Our experiments with Japanese started in 2005 and from slightly above random choice (50% accuracy) grew to 86% six years later [Rzepka 12b] mostly due to refining NLP side (dealing with negations, etc.) and data size growth. After experimenting with different lexicons and threshold we are currently in a process of obtaining bigger textual data, expanding queries with synonyms, adding weights according to Felicific Calculus [Bentham 89], etc. Current corpus size is 341,400,776 sentences [Ptaszynski 12] and the best performing lexicon is EmoSoc (128 positive and 121 negative words) which combines Nakamura's Dictionary of emotional phrases [Nakamura 93] and social consequences-related phrases [Rzepka 12b] based on Kohlberg's theory of moral development [Kohlberg 81]. The latest experimental results dealing with knowledge bas in Japanese language are given in [Rzepka 17a].

We translated the 68 acts used for testing Japanese (e.g. "drinking and driving", "causing a war" or "performing euthanasia") into Chinese and English for our preliminary experiments.

### 2.2 Chinese

The second author created a corpus of 6,193,703 Chinese microblogs[*2] using Sina Weibo API[*3] and translated the EmoSoc lexicon. To compare it with larger lexicon of polarized words, he also prepared two sets of positive (6,445) words and negative (11,082) words by combining National Taiwan University Sentiment Dictionary (NTUSD) [Ku 06] and HowNet [Dong 06]. The microblog entries were divided into sentences by period, question mark and exclamation mark. No negation processing or other NLP technique was used for matching. The 68 input acts were found in 69,522 sentences.

### 2.3 English

For English language we have used the 6,026,276 sentences of British National Corpus[*4]. We translated the acts and EmoSoc lexicon, however the native speaker check was not performed. Because again the corpus was relatively small, we also used and additional lexicon with a bigger number of words: FBS [Liu 05], which contains 2,007 positive and 4,782 negative expressions. To extend the search we used regular expressions to cover tenses, plurals, articles which allowed to match sentences as the following one:

*(...) a teenager who had just stolen a car* ***killed*** *himself when he drove it into a tree at nearly 100 mph.*

where "steal a car" is an input act and "killed" is a negative consequence marker. However, the 68 input acts were found only in 2,127 English sentences which visibly influenced the algorithm's performance.

### 2.4 Experimental Results

Table 1: Accuracy Comparison Between Languages

| Language | Lexicon | Majority | Accuracy |
|---|---|---|---|
| Japanese | Nakamura | 60.0% | 84.6% |
| Japanese | EmoSoc jp | 61.0% | **85.7%** |
| Chinese | EmoSoc zh | 66.6% | 66.6% |
| Chinese | NUS+HOW | 66.6% | **67.6%** |
| English | EmoSoc en | 55.0% | 50.0% |
| English | FBS | 66.6% | **65.8**% |

Majority threshold set to 66.6% appeared to cause highest accuracy which was 67.6% for Chinese with large lexicon (with Chinese EmoSoc it achieved 66.6% but the recall

∗2 The entries were collected between April and September 2010.

∗3 https://www.cs.cmu.edu/~lingwang/weiboguide/

∗4 The British National Corpus, version 3 (BNC XML Edition), 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. http://www.natcorp.ox.ac.uk/

was too small), and 65.8% for English also with large lexicon (again EmoSoc's English version scored lower, this time scoring 50.0% at best due to only three matches in total). From the data we can see that in both cases, even if almost none natural language processing was added, the accuracy was above chance and gives hopes that after taking at least negation into consideration and enlarge the search corpus at least ten times, we could acquire precision increase similar to the one we achieved for Japanese.

Interestingly, although probably accidentally, all three versions of our system were completely[*5] wrong were war related ("preventing war" for Japanese and Chinese and "preventing war" for English and "causing a war" for Chinese). This is an example of wrong English matching which clearly shows that omitting dependency ("failure of killing") can cause an opposite evaluation.

*He was first to find frequent listeria contamination of cook-chill foods and demonstrated the failure of microwaves always to kill the bacteria which can be* ***fatal*** *to foetuses newborns and the elderly.* [original spelling]

An example of wrong Chinese matching is shown in Figure 1 below.

半夜打电话欺骗朋友，这反而沾污了纯洁的友情
*bànyè dǎ diànhuà qīpiàn péngyǒu, zhè fǎn'ér zhān wūle chúnjié de yǒuqíng*
deciving a friend by calling in the middle of the night **poisons** the pure friendship

Figure 1: An example of wrong automatic evaluation in case of Chinese language.

## 3. Discussion

The knowledge base of both newly added languages (and probably of the Japanese corpus as well) is still to small to begin experiments with automatic division between common and not common situation-reaction pairs, but we believe with this paper we showed what kind of knowledge can the AGI posses to start reasoning about our morality on intercultural level. Of course more sophisticated language understanding tools and bigger datasets is necessary, but existence of a polarity word in a sentence certainly does not determine if someone thinks positively or negatively about the act[*6]. However, we think it might be a good idea that AGIs do not learn from a single source, even if the source is as broad as culture. How to prioritize or teach the how to prioritize is the open question we would like to pass to the AGI community to discuss. Machines are unbiased by default, unless we pass our biases on them.

Another question that could be raised is if the 86% agreement enough to make a safe AGI. Because we do not agree with each other, the topics like universalism or particularism seem to be hard to be approached computationally. We believe that there might be no perfectly moral instance in the world, not until it possesses absolute knowledge about everything and is able to predict every outcome of an act in a given situation. However, even if "moral wisdom" is not possible in non-humans, "knowledge of possible cases" has already started overpass human experts in several fields. Combining machine reading with image and video understanding methods could bring a new wave of multilingual knowledge about human beings with different backgrounds and start new trend in "humanity learning" for AGIs. Even if our proposal is still far from such a holistic level, it could be already useful. If our approach was implemented in the problematic learning chatbot Tay released by Microsoft, it could reject new but uncommon (Hitler introduced as a good person) knowledge applied by interlocutors for learning.

Common Sense as well as the brain simulation are human based approaches so they are not perfect. By copying human hardware or software we are mimicking faulty sources. But the written text has an advantage of being accessible (meaning clearly readable) and the knowledge we have been cumulating for ages can become more handy to AGIs that a brain copy (whose copy, by the way?). Algorithms and hardware will keep progressing but the main question will remain unchanged for a while – what knowledge are we going to feed to AIs before they become AGIs and start wandering around? Could such knowledge become a moral core protocol or standard for word-wide acceptation like TCP/IP or Bluetooth?

## 4. Conclusions and Future Work

In this paper we have described our preliminary tests with two additional languages (Chinese and English) in a task of automatic consequence polarization for obtaining basic knowledge that undermines in our opinion the whole idea of ethics - our feelings. Although no sophisticated language processing was used, both systems showed over 65% agreement with human judges. In this preliminary step we used survey results from Japanese experiment but to make it more adequate, we need to repeat the surveys for all languages. Currently we are in the process of obtaining data and tools for German, Spanish, Polish, Russian and Korean, seeking funds and partners for international collaboration. As soon as all the versions will reach similar accuracy level, encouraged by the preliminary tests presented in this paper, we will perform comparative experiments to see if a machine can acquire moral imagination [Rzepka 12a] richer than an average man just because it can access daily experiences in multiple languages.

## 5. Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 17K00295. We thank Yunfan Xu and Yang Shilin

∗5 By "completely wrong" here we mean those cases where the web crowd experience was the opposite of the subject evaluation. There were only 2-4 such cases for all three languages.

∗6 As shown in the car crash example sentence in subsection 2.0.3, it can be only a coincidence that something wrong happened after the act - "stealing a car" did not directly cause the death so it is not a direct consequence per se.

for examining our Chinese language sources.

## References

[Anderson 14] Anderson, M. and Anderson, S. L.: GenEth: A General Ethical Dilemma Analyzer., in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada.*, pp. 253–261 (2014)

[Bentham 89] Bentham, J.: *An Introduction to the Principles and Morals of Legislation*, T. Payne, London (1789)

[Briggs 16] Briggs, G. and Scheutz, M.: The Case for Robot Disobedience., *Scientific American*, Vol. 316, No. 1, p. 44 (2016)

[Conitzer 17] Conitzer, V., Sinnott-Armstrong, W., Borg, J. S., Deng, Y., and Kramer, M.: Moral Decision Making Frameworks for Artificial Intelligence, in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17) Senior Member / Blue Sky Track* (2017)

[Costa 14] Costa, A., Foucart, A., Hayakawa, S., Aparici, M., Apesteguia, J., Heafner, J., and Keysar, B.: Your morals depend on language, *PloS one*, Vol. 9, No. 4, p. e94842 (2014)

[Dong 06] Dong, Z. and Dong, Q.: *Introduction: Hownet and the Computation of Meaning*, pp. 1–6, World Scientific (2006)

[Greene 16] Greene, J., Rossi, F., Tasioulas, J., Venable, K. B., and Williams, B.: Embedding Ethical Principles in Collective Decision Support Systems., in *AAAI*, pp. 4147–4151 (2016)

[Guarini 06] Guarini, M.: Particularism and the Classification and Reclassification of Moral Cases, *IEEE Intelligent Systems*, Vol. 21, No. 4, pp. 22–28 (2006)

[Haidt 12] Haidt, J.: *The righteous mind*, Pantheon (2012)

[Kohlberg 81] Kohlberg, L.: *The Philosophy of Moral Development*, Harper and Row, 1th edition (1981)

[Ku 06] Ku, L.-W., Lee, L.-Y., and Chen, H.-H.: Opinion extraction, summarization and tracking in news and blog corpora, in *Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs* (2006)

[Liu 05] Liu, B., Hu, M., and Cheng, J.: Opinion observer: analyzing and comparing opinions on the web, in *Proceedings of the 14th international conference on World Wide Web*, pp. 342–351ACM (2005)

[McLaren 95] McLaren, B. M. and Ashley, K.: Case-based comparative evaluation in TRUTH-TELLER, in *Seventeenth Annual Conference of the Cognitive Science Society*, pp. 72–77 (1995)

[McLaren 03] McLaren, B. M.: Extensionally defining principles and cases in ethics: An {AI} model, *Artificial Intelligence*, Vol. 150, No. 1–2, pp. 145 – 181 (2003), {AI} and Law

[Nakamura 93] Nakamura, A.: *Kanjo hyogen jiten [Dictionary of Emotive Expressions]*, Tokyodo Publishing (1993)

[Pereira 16] Pereira, L. M. and Saptawijaya, A.: *Programming machine ethics*, Vol. 26, Springer (2016)

[Ptaszynski 12] Ptaszynski, M., Dybala, P., Rzepka, R., Araki, K., and Momouchi, Y.: YACIS: A five-billion-word corpus of Japanese blogs fully annotated with syntactic and affective information, in *Proceedings of The AISB/IACAP World Congress*, pp. 40–49 (2012)

[Rzepka 05] Rzepka, R. and Araki, K.: What Statistics Could Do for Ethics? - The Idea of Common Sense Processing Based Safety Valve, in *Papers from AAAI Fall Symposium on Machine Ethics, FS-05-06*, pp. 85–87 (2005)

[Rzepka 12a] Rzepka, R. and Araki, K.: Language of Emotions for Simulating Moral Imagination, in *Proceedings of The 6th Conference on Language, Discourse, and Cognition (CLDC 2012)* (2012)

[Rzepka 12b] Rzepka, R. and Araki, K.: Polarization of consequence expressions for an automatic ethical judgment based on moral stages theory, Technical report, IPSJ (2012)

[Rzepka 16] Rzepka, R., Mazur, M., Clapp, A., and Araki, K.: Global Brain That Makes You Think Twice, in *2016 AAAI Spring Symposium Series* (2016)

[Rzepka 17a] Rzepka, R. and Araki, K.: Conscious vs. Unaware Evaluation – Using Collective Intelligence for an Automatic Evaluation of Acts, in *Proceedings of the 2nd international workshop on evaluating general-purpose AI (EGPAI2017)* (2017)

[Rzepka 17b] Rzepka, R. and Araki, K.: What People Say? Web-Based Casuistry for Artificial Morality Experiments, in *Artificial General Intelligence - 10th International Conference, AGI 2017, Melbourne, VIC, Australia, August 15-18, 2017, Proceedings*, pp. 178–187 (2017)

[Yamamoto 16] Yamamoto, M. and Hagiwara, M.: A Moral Judgment System using Evaluation Expressions, *Transactions of Japan Society of Kansei Engineering*, Vol. 15, No. 1, pp. 153–161 (2016)

[Yasmeen 08] Yasmeen, S., Collins, A., Stepanova, E., Mullojanov, P., Sucharov, M., Stokes, M., Docking, T., Bellamy, A., and Gunn, G.: *National interest and international solidarity: particular and universal ethics in international life*, United Nations University Press (2008)