Importance of Contextual Knowledge in Artificial Moral Agents Development

Rafal Rzepka, Kenji Araki
Graduate School of Information Science and Technology
Hokkaido University, Kita-ku, Kita 14, Nishi 8
060-0814 Sapporo, Japan

Abstract
In this paper we underline the importance of knowledge in artificial moral agents and describe our experience-focused approach which could help existing algorithms go beyond proofs of concept level and be tested for generality and real-world usability. We point out the difficulties with implementation of current methods and their lack of contextual knowledge hindering simulations in more realistic, every-day life situations. The idea is to prioritize resources for predictions and the process of automatic knowledge acquisition for an oracle to be used by moral agents, both human and artificial.

Introduction
Value alignment problem has recently gained the attention among artificial intelligence researchers, philosophers and non-specialists. Nick Bostrom’s book “Superintelligence” (Bostrom 2014) has popularized the topic and the potential dangers of high-level autonomy machines became widely discussed, also by influential figures as Stephen Hawking, Bill Gates or Elon Musk. Although the discussion has lasted for years and many possible solutions have been proposed, a universally moral machine is still far from reality. One of the major problems is the fact that universal code of ethics is hard (or impossible) to establish due to the cultural differences and the influences of other bigger and smaller contextual variations. For that reason the Artificial Intelligence researchers are in a difficult position when they try to create an ethical decision making system which could help alleviating worries about the future of AI.

Hypothesis
Our hypothesis is that knowledge acquisition field should be as important for building unprejudiced systems as the very process of creating them. Top-down approaches are difficult to implement (what exactly does one mean by “do not harm”? ) and in bottom-down approaches the decision process generates output which origin is often difficult to explain. However, both strategies could benefit from gathered and processed real world descriptions and become more easily testable and expandable. Because we work, among others, on dialog systems, basically without any input restrictions, workable but sufficiently general methods for moral evaluation are necessary at the very moment. For this reason we aim at solutions not for hypothetical super-intelligent agents dealing with famous moral dilemmas but for existing systems which must avoid learning failures like Microsoft’s Tay bot that was tricked to praise Hitler. On the other hand, we also are aware about pitfalls of gathering knowledge without quality considerations. In this paper we would like to emphasize the importance of rich examples or real-world situations. To the authors’ best knowledge all existing implementations of artificial moral agents (AMAs) deal only with toy applications (prototypes) and very specific / limited tasks. We think the shortage of contextual data is one of the main reasons restraining AI systems from becoming more general, expandable and testable. Knowledge in these systems is manually crafted making them less realistic and difficult to be implemented in real-world, everyday applications.

Existing Approaches to Machine Ethics
Variety of possible solutions for achieving ethical machines have been proposed and one of the latest survey of existing methods is given in (Pereira and Saptawijaya 2016). Authors of this book divide the approaches into two realms – one dealing with individual ethical instances and second describing collective morality, which combines game theory (Conitzer et al. 2017) and findings of the evolutionary psychology. They introduce their approach using logic programing for individual moral agents and propose methods for bridging both realms. Logic-based methods, for example by formalizing ethical codes with deontic logic of (Bringsjord, Arkoudas, and Bello 2006) are probably the most popular and machine learning (Anderson, Anderson, and Armen 2006) is not used widely as it is often believed that machine ethics cannot be based on predicting how to do the right thing. We partially disagree with (Pereira and Saptawijaya 2016) claiming that the community should be well aware that such present day learning is inadequate for

2Due to the limited space we only mention main types of approaches; our proposal refers to probably all AMAs but is different in focusing on knowledge rather than algorithms.
general machine morality. Only small, circumscribed, well-defined domains have been susceptible to rule generation through machine learning. Rules are all important for moral explanation, justification and argumentation. We think rules can be extrapolated from fuzzy observations and we believe the observations helped humans greatly in creating ethics (as well as language, mathematics or logics). We return to this theme several times in this paper as the knowledge (result of observations) is one of the two cores of our approach.

On the other hand, we are also aware that learning from Big Data in the spirit of purely statistical and probabilistic calculations is also flawed, risky and the agent’s reasoning often cannot be explained. However, we think that methods like deep learning can enrich the textual data which lacks tacit knowledge. We also see a problem common to probably all approaches, including cognitive architectures (Bretz and Sun 2017) – the need of creating correct data sets and (preferably contextual) knowledge bases. Their manual creation / annotation is costly and impracticable when all even only the most probable situations that an agent may face are needed to be considered\(^3\). To address this problem, methods for automatic acquisition of moral rules e.g. by human-machine cooperation with Inverse Reinforce Learning (Ng, Russell, and others 2000; Hadfield-Menell et al. 2016) were suggested. However, they, similarly to the “seed AI”-like approaches, also seem unrealistic because teaching (supervising) an agent to deal with complex cases in changing environments could take very long time and the AMA would be influenced by one supervisor’s experiences and his or her preferences. Thorough the scrutiny of formal methods and shallow but wide stochastic approaches can help each other or even be integrated into more holistic systems for example using probabilistic methods like Bayesian inference (Tenenbaum et al. 2011). But before that, at least in our opinion, it seems necessary to provide more structured crowd-based contextual data which could allow:

- discovering causes and effects
- calculating probabilities
- forming and dissolving abstract knowledge
- simulating real world situations
- testing existing and new moral agents

In the next section we describe our approach which discovers causes and effects for the moral judgement task. After that we present our idea of expanding the existing ontologies to deal with concepts as stories, the need of controlling data credibility and the importance of language itself. In the last part of this paper we answer several questions that often appear when discussing our approach with other researchers.

**Knowledge-First Approach**

Our proposal is to shortly go back to the point in our evolution when no theories of ethics were yet formulated. We assume that empathic circuitry in our brains, together with the capabilities to observe the world and to communicate with peers ignited codification of our sense of justice which keeps changing throughout the ages. The idea is to simulate this process (and test our hypothesis) by first creating conditions for discovering contextual dependencies that influence moral load of given states and acts. These conditions are currently reduced to a) unstructured knowledge in natural language b) agent’s capability to guess a polarity (positive or negative) of concepts (acts or states).

**Source Knowledge**

As mentioned before, although the broad world knowledge seems to be an obvious ingredient of moral reasoning, it is widely ignored by the creators of Artificial Moral Agents. To show that it is not only useful but crucial in machine ethics we utilize various text resources like blog corpus, Twitter corpus, Aozora book repository (we mostly work with Japanese language), chat logs, etc. which contain billions of words. Basically matching concepts and the natural language processing is performed of a limited context of on, two or three sentences (sentence with a concept being analyzed, previous sentence containing possible reasons and following sentence with possible consequences).

**Polarity Calculation**

For time being we utilize sentiment analysis methods to help our systems asses consequences. The initial idea is presented in (Rzepka and Araki 2005) and more technical details are given in (Rzepka and Araki 2012) and (Rzepka and Araki 2015). The simplest method for this task utilizes lexicons of positive and negative words. For example, if most of human experiences with “stealing a car” described in text resources cooccurred with negative lexicon words, the polarity of the concept becomes morally negative. Except emotion-related phrases we also created a lexicon based on Kohlberg’s stages of moral development (praising / reprimanding, awarding / punishing, etc.) to extend recognition to legal consequences (if an act ended in doer’s arrest it is more likely that the act was not moral).

**Precision of Moral Estimation**

The latest experiments (Rzepka and Araki 2017) show that our simplistic approach is able to achieve almost 85.7% agreement with human subjects. However, the results showed that mere size of knowledge base does not equate to better ethical judgement. Not only different automatic polarity estimation methods must be tested, also the credibility of the sources require investigation. We elaborate on this problem and propose solutions in later sections. It also must be noted that the experiments were performed with concepts and many of them strongly depend on wider context. The input is basically unrestricted when it comes to the topic but longer concepts decrease the chance of finding sufficient number of examples. For instance driving should be recognized as neutral, driving after drinking as negative, but driving with a baby unbuckled after home party at friends house on the hill cannot be found in the given input form, therefore recognizing, abstracting and weighting concepts within the

\(^3\)In logic-based approaches knowledge is limited to a given task, usually a single dilemma in very restricted environment.
input becomes necessary. Unfortunately, automating these tasks is rather difficult without sufficient set of reliable examples from which e.g. a concept’s importance can be calculated. This is one of the reasons we are currently preparing the ontology of concepts discussed later in this paper.

**Tests with Embodied System**

Many researchers draw attention to the importance of embodiment in moral behavior (Trappl 2015) and need concrete testing decision-making algorithm in action (Arnold and Scheutz 2016). To see how our text knowledge-based approach works in the real world, we implemented our method on a Roomba robotic vacuum cleaner (Takagi, Rzepka, and Araki 2011). Users were allowed to communicate freely with the device through Twitter. The robot had its name ("Roomba") and function (variants of the verb "to clean") hardcoded, and its mission was to make a user happy without violating common sense, which is the motto of our approach. The system, with knowledge base limited only to Twitter corpus worked surprisingly well and the robot was able to propose its help even if no straightforward command was given. For instance, "this room is a mess" has triggered negative reactions and Twimba (the name of our system) found by simple search that people deal with this problem by cleaning which was its capability. On the other hand, when one talks about a “dirty look”, the robot does not react, because it deals with concepts, not single words. Not caring about even very small contexts, although common in various machine learning methods, showed us clearly that deeper and more careful approach is necessary. Naturally it was hard for a vacuum cleaner to violate common sense, but “knowing” its name and its only function helped it to refuse cleaning a bathtub, only because no examples of Roombas cleaning bathtubs were found in the knowledge base.

**Other Characteristics and Possibilities**

As showed above, the sophistication of moral behavior may increase with machine capabilities but does not seem to be limited to embodied agents. Obviously the more actions a machine can perform, the more dangerous it can become, but e.g. chat systems with purpose like the second language acquisition tutor (Nakamura et al. 2017) have to deal with abstract concepts and utilize their “talking” capability that conveys meaning which can be directly and indirectly harmful to the user. For example, an artificial tutor reacting positively to a bullying statement is not only unnatural but also may negatively influence adolescent users.

When implemented in a dialog system, our method needs to support explaining its judgements which is an important functionality for an AI system (Core et al. 2006). Explainable AI needs linguistic skills and the reasoning should be clear to any user. Simplicity of the current algorithm and dealing only with natural language makes it relatively easy to generate explanations how a given judgement was performed. In case of our majority voting strategy, it is currently enough to use only four output templates: a) “It’s moral because majority (X%) of cases had positive consequences”, b) “It’s immoral because majority (Y%) of cases had negative consequences”, c) “It’s problematic” and d) “Not enough data”. Examples of observations can be also easily added. We have tested different majority thresholds and 60-70% level seems to be most effective (Rzepka and Araki 2017). The non-decisive middle area when roughly half consequences were recognized as good and half as bad (“problematic” output) constitute a safety valve (Rzepka and Araki 2005). It contains concepts like abortion or euthanasia and it advised to program a system with our algorithm to avoid actions and strong statements when even people are not sure about the outcomes. To allow our method to handle such cases and be able to perform ethical judgement and decrease “Not enough data” outputs, again more contextual knowledge is needed.

**Toward the Ontology of Contexts**

**Current Knowledge Bases**

Current knowledge bases are stored in various formats but usually they can be represented in a flat and solid, cross-linked structure like hypertext (Wikipedia, DBpedia, BabelNet, etc.) which links terms with other terms, categories or definitions. Ontologies (semantic nets) like CyC (Lenat and Guha 1989) or ConceptNet (Speer and Havasi 2012) try to connect more abstract, commonsensical concepts, but they do not contain longer chains of consecutive concepts which could form, for instance, a Schankian script (Schank and Abelson 1977). Therefore there is a gap between such knowledge bases that cover small chunks of knowledge and just raw text which very often describe much bigger contexts but are incomplete and/or noisy. We treat moral decision making as a subtask of the common sense processing. It requires processing extendable / shrinkable data chunks that constantly change their size and density depending on the stream of information (linguistic in our case). This information always changes as time moves forward and elements of environment alter, but if an apple changes color to brown, it does not mean a concept of apple like HasProperty changes from “sweet” to “rotten”.

**Expanding Number of Concepts and Their Relations**

We are currently experimenting with combining existing concepts (from ConceptNet) into longer chunks of possible chains by confronting them with the blog corpora. Several techniques are required for cleaning up the text, recognizing semantic roles, tackling with anaphora resolution, double negations and other NLP-related tasks. Because the Japanese ConceptNet is not big enough, we currently work on expanding it (Krawczyk, Rzepka, and Araki 2016) and checking its quality (Shudo, Rzepka, and Araki 2016). The idea of data being used to supervising other data is not new, it is called distant supervision (Mintz et al. 2009) where the data replaces human in learning or other tasks usually requiring human’s assistance. In Figure 3, we show how the text knowledge itself can be useful in both expanding the knowledge base and supervising any machine learning algorithm giving positive and negative feedback from polarity calculation module. Simultaneously we are trying to acquire new
Figure 1: Three layers of language-based moral judgement allowing understandable explanation of ethical choices calculated from polarity of possible consequences.

Enriching Context with Automatic Descriptions

Another important and unanswered question in common sense knowledge acquisition is how to provide machines with tacit knowledge which is obvious for us thanks to our sensory input and is rarely expressed in language. As we showed in the Figure 1, we believe that advances in pattern recognition will be able to at least partially tackle this problem with methods like deep learning which has already had some successes in automatically describing images in natural language (Vinyals et al. 2015). Currently we simulate sensory input with text-mining techniques (Rzepka, Mitsuhashi, and Araki 2016), but let us assume the progress in pattern recognition (machine learning on constantly growing data) has reached the human level without the massive and costly annotated data. Every image or video available can be described in a natural language in detail and every sentence in written text can be flawlessly parsed. The speed of access and analysis naturally surpasses human capabilities. Our hypothesis is that just because a machine can refer to more experiences (cases, contexts, regulations, etc.) than we can, it is theoretically possible for the machine to generate more fair judgement even than ethicists or judges. Moreover, if programmers ensure that the moral judgement algorithm is not prone to biases (or at least is less biased than most of us), an agent could become an important advisor for human or robotic users. We discuss such a possibility of ideal advising oracle in the last part of this paper.

Credibility Problem

Internet is a source of countless examples of knowledge which is simply wrong. Darker side of human nature reveals itself with spams, scams, flame wars, trolling, conspiracy theories, fake news and so on. Our beliefs are often shaped by cognitive biases and laziness or lack of time force us to access the click-baits or to share unscientific revelations. Machines are more patient, and if programmed carefully, could avoid such errors by fastidious analysis, not only the sources but also confirm contents via thoroughly scanned newspapers, research papers, history books. But the machine reading field is not there yet, so for time being we have to test easier solutions and use surface methods as identifying and classifying the source, analyzing the appearance of a page or writing style of its creator (Akamine et al. 2010).

Moreover, few last years have showed another problem with Big Data and machine learning, i.e. artificial intelligence systems acquiring stereotypes associated so far only with human beings (Bolukbasi et al. 2016; Caliskan, Bryson, and Narayanan 2017). Not dealing with this problem might end with a dialog system stating that woman’s place is in the kitchen, all grandmothers are white (knowledge form any image search engine), and items recommended by people with non-Western names will be less trustworthy. There are several methods for unbiasing the data, abstracting or altering concepts is two possible option we consider. Instead of man or woman, “a person” can be used, although the data would need a few layers of semantical specificity because the knowledge of gender is often important for understanding. Removing bias manually from the data is laborious, and we believe that automatic discovery of reasons behind the stereotypes would be an ideal scenario. Removing any problematic concept from knowledge could lead to false discoveries, therefore several experiments must be performed to see if the oracle is able to find enough examples of stereotypes.
Philosophical Stance (or Lack of It)

Our experience with both robotic and non-robotic systems suggest that not only embodiment is unnecessary for moral decisions but also there is probably no need for subjectivity connected to consciousness often declared as the foundation of human ethical domain (Nath and Sahu 2017). Because our approach is rather pragmatic in its nature and we usually give rather scarce explanations about the bigger picture, we decided to use this section to explain some points which are very often misunderstood by our critics.

Provoking Philosophy by Avoiding It

Principally, we want to avoid adhering to any particular ethical school of thinking, although example-based approach might be used for testing utilitarian (by calculating utilities) and deontological (by extrapolating rules) systems. There are some ideas in modern of ethics which can be easily attached to our strategy, for instance an idealized ethical advisor is discussed by various philosophers (Sidgwick 1907; Firth 1952; Rawls 1971; Harsanyi 1977). (Sobel 1994) and (Rosati 1995) are probably the main critics of such all-knowing moral agent and the former describes four objections which our system could be referred to. The first one suggests that an ideal advisor could get lost in too many, always changing perspectives. As we show in Figure 4 always growing knowledge is not the obstacle but the opposite. Controlling timeline (as the consequences change with history) should be performed to avoid discovering polarities which were different a century ago, e.g. reactions to public lynches. Sobel’s second and third objections applies to agent’s experience: evaluation of one life can be evaluated only if it is experienced and this experience biases the agent when experiencing another one. Similar argument can be made about artificial agent which is given one set of experiences but in our case maximal number of experiences is used and forgetting one to process another is not necessary. The last objection argues that the Ideal Agent with perfect knowledge can conclude that non-perfect agents’ is not worth living due to its limitations. To make robot with our system implemented kill anyone, the vast majority of stories would need to contain examples that killing is good, which is not true (the scale of actual data is shown in Figure 4). The same can be said about any utility maximizer often shown as an exemplification of dangerous AI. By changing the focus from theory to experience we our approach is closer to what Johnson calls “moral imagination”. In (Johnson 1994), he challenges traditional ethics by emphasizing the role of stories we are confronted with from very early stage of our lives. Equipped with empathy we process examples from children’s books, novels, movies. Our morals keep evolving as we are experiencing stories in our own lives, both by observing them and taking an active part.

Addressing Risks and Limitations of Machine Ethics

The complicated character of human ethics raises questions about risks and limitations of processing moral problems by non-human agents. (Brundage 2014) lists problems of the emerging field suggesting the whole endeavor might be pointless. As our systems need moral decision as we speak, we disagree with the main line of the critique, but agree with some points and believe they should be addressed. The problem of insufficient knowledge, complexity and/or the possible lack of computational resources is what we plan to solve by constructing a vast contextual ontology which should grow with the progress of both knowledge acquisition and computational capacities of hardware. Brundage points out that machine ethics is not able to make an exception to a rule when an exception shouldn’t have been made based on the morally relevant factors. As our approach does not rely on any hard-coded rules and is supposed to discover and analyze as many factors as possi-
Figure 3: Unbiased collective intelligence as a source for machine learning: by giving the system examples of human experience could lead to richer reasoning about reasons and consequences of human / robot acts.

...ble, dealing with exceptions should be easier than in other approaches. It is difficult to ensure perfect decision because there always might be a better one, but with unbiased knowledge and analytical power, a machine (at least in theory) might be a better and faster judge than average human being. Another set of possible problems is related to moral dilemmas facing an agent when it needs to sacrifice something important. Contextual knowledge based on real stories with reasons and consequences should contain examples of sacrifices which makes the problem of insufficient data most important to deal with. So called “folk morality” is often flawed, as Brundage notices. For that reason we concentrate on observing consequences, not on how people reason. He also worries that extrapolation of our values may be far beyond our current preferences. In our opinion, restricting our algorithm with common sense boundaries should prevent AI from becoming too creative and stop aligning with our values.

Conclusions and Future Work

Various models of moral judgment have been proposed and can be used in Artificial Moral Agents development, for example (Dehghani et al. 2008; 2008; Nado, Kelly, and Stich 2009; Ord 2015). On the other hand, empirical methods slowly enter the field of ethics and show, among others, how morals differ between cultures (Buchtel et al. 2015) or that feeling right is often more important than feeling good (Tamir et al. 2017). With this paper emphasizing the importance of the empirical (observational) side of ethical reasoning, we would like to spark a discussion about collecting, storing and normalizing contextual knowledge (chains of very specific concepts instead of very general single concepts). We believe that such knowledge could be very helpful in extrapolating rules, learning possible outcomes or testing existing systems. We believe that natural language, even being fuzzy and incomplete, can be a safe interlayer between the real world and abstract notions like ethics.

As computer scientists we often tend to model the world in a strict manner, we prefer to control input and output so the proposed algorithms can be easily tested and the results be published. But the value alignment may require us to share a significant part of the control to the world around us (by descriptions of it). For six million years we have gathered knowledge which becomes more and more accessible for machines and we believe it would not be smart if we ignore the contextual variety of “good vs. bad” stories humankind keeps accumulating. We believe that taming this knowledge may accelerate the progress of safe AI on a larger scale that is usually seen. It might be easier and faster to program a machine to acquire logics by analyzing moral cases than program logics to acquire morality. Whichever method will be most robust and “just”, the knowledge will be their common ground.

There are various approaches how to define the inborn instincts of justice. But a computer could learn from manifestations of those instincts without understanding them. As computer pattern recognition capabilities constantly grow, AI climbs bastions of human intelligence one after another. As Watson was more often correct than the best humans, some AMA can be more often “right” than all of us. Without any thinking, consciousness, free will but massive (multicultural) collective intelligence with decreased bias and increased credibility might be helpful not only to AI systems but also to anyone of us, even if in a form of mere voice...
Figure 4: Importance of experience data size: Although number of positively labelled sentences about killing somebody also increases with new examples, the increase of correct (negative) consequence estimation is significantly higher.

Acknowledgements
This work was supported by JSPS KAKENHI Grant Number 17K00295. We thank the Centre of Excellence for the Dynamics of Language at Australian National University for providing an ideal environment for transdisciplinary discussion that lead to crystalizing ideas described in this paper. We would also like to thank Dr. Siva Kalyan for enriching our background knowledge.

References


