Cake, death, and trolleys: dilemmas as benchmarks of ethical decision-making*

Edvard P. Bjørgen and **Simen Madsen,**
Therese S. Bjørknes, Fredrik V. Heimsæter, Robin Håvik, Morten Linderud, Per-Niklas Longberg,
Louise A. Dennis† and Marija Slavkovik‡
University of Bergen, Norway

Abstract

Artificial intelligence (AI) systems are becoming part of our lives and societies. The more decisions such systems make for us, the more we need to ensure that the decisions they make have a positive individual and societal ethical impact. How can we estimate how good a system is at making ethical decisions? Benchmarking is used to evaluate how good a machine or a process performs with respect to industry bests. In this paper we argue that (some) ethical dilemmas can be used as benchmarks for estimating the ethical performance of an autonomous system. We advocate that an open source repository of such dilemmas should be maintained. We present a prototype of such a repository available at `https://imdb.uib.no/dilemmaz/articles/all`.

Introduction

Physical and software systems are capable of making informed decisions with varying degree of autonomy (Bryson and Winfield, 2017). Examples of such systems include, but are not limited to: driverless vehicles, assisted living technologies, algorithmic trading systems (Scopino, 2015), judicial decisions analysis systems (Aletras et al., 2016), content discovery and recommendation systems (Mobasher, Cooley, and Srivastava, 2000). With every decision that we outsource to an autonomous system, we also outsource the power to control the impact that decision has on the individual lives of people and society as a whole. With the power to have such impact comes the responsibility to ensure the maintenance of society's laws and values. Machine ethics is a discipline in artificial intelligence (AI) concerned with the issues of enabling autonomous intelligent systems to "behave ethically" when operating within our society (Moor, 2006; Anderson and Leigh Anderson, 2015; Dennis, Fisher, and Winfield, 2015; Winfield, Blum, and Liu, 2014; Charisi et al., 2017).

The increase in number of areas in which intelligent autonomous systems can be utilised calls for an increased attention to the decision-making algorithms and the implications of the decisions these algorithms produce. These decision-making algorithms must not only be correct from an engineering standpoint, but also from an ethical one. However, different domains of operation imply different engineering approaches towards implementing decision-making and reasoning, and further different choices of ethical theories and value systems that are implemented. How under such variety can governing bodies and engineers themselves compare and contrast the ethical reasoning capabilities of intelligent autonomous systems?

Benchmarking is the process of comparing the performance metrics of a system to similar systems that are the state of the art within the same industry. We can compare two autonomous systems with respect to numerous metrics such as: number of contexts they can operate in, speed with which they evaluate a situation and make a decision, efficiency of problem solving. In contrast, it is more difficult to quantify and compare how ethical a system is.

Ethical behaviour, individual and social values, are all traditional topics of concern of moral philosophy (Etzioni and Etzioni, 2017; Gert and Gert, 2017). Normative ethics in particular is concerned with deriving theories that describe how one ought to act. Numerous specific theories have been developed, the most notable of which are perhaps utilitarianism (Harsanyi, 1977), Kantianism (Ellington, 1993), and Ross's ethical theory (Ross, 1930). There is no direct way in which ethical theories can be compared. Certain actions can be classified as morally wrong by one theory while being morally right by another.

However, we do find that often in moral philosophy, there are examples of decision-making scenarios that are used to demonstrate the failures of a particular theory. These examples are called *ethical dilemmas* and typically consist of comparing two options none of which is unambiguously morally preferable to the other. One of the best known examples of an ethical dilemma is the *trolley problem* introduced by Foot (1967). Can we perhaps do something similar in machine ethics and use ethical dilemmas as a benchmark of how ethical an intelligent autonomous system is?

Ethical dilemmas in moral philosophy embody the hardest of reasoning problems. Consider for instance the trol-

*The authors of this paper are master students at the University of Bergen. The presented work was conducted as an independent student research project on machine ethics under the supervision of M. Slavkovik with the help of L.Dennis

†Louise A.Dennis is affiliated with the University of Liverpool. Her work was funded by the EPSRC "Verifiable Autonomy" research project (EP/L024845).

‡Contact authors at: marija.slavkovik@uib.no

ley problem. There is a runaway trolley on a rail-road track. Ahead on the track there are five people tied up, unable to escape being killed by the trolley. The track splits into two by a lever which can divert the trolley onto a second track where there is one person tied up. The ethical dilemma is to chose between diverting the train or not. This problem, as most of the ethical dilemmas in moral philosophy is not one that has an obvious solution.

Ethical dilemmas such as the trolley problem would not be easier to solve for machines than they are for people, despite there perhaps existing such expectations put on machines (Malle et al., 2015). For autonomous intelligent systems two other types of dilemmas can be considered that do not exist in moral philosophy:

- cases ethically trivial for people but not for machines;
- cases that arise in autonomous systems contexts but not in human decision-making ones.

It is rather these two classes of dilemmas that are of interest as possible benchmarks of ethical decision-making.

In this paper we first introduce the new categories of ethical dilemmas and discuss the difference with moral philosophy dilemmas. We proceed to advocate that a repository of ethical dilemmas for benchmark and testing should be kept and we present a prototype of such a repository. Next we analyse the different ways in which dilemmas should be formally modelled to be available for benchmarking and we also discuss examples of how dilemmas have been used when developing ethical reasoning systems in related work. Lastly we draw our conclusions and outline directions for future work.

Ethical dilemmas for humans and machines

In moral philosophy, an ethical dilemma is an example of a choice where different theories point to a different ethical option. There exists no consensus on a moral theory in philosophy. We can however seek a "reflective equilibrium" among the various existing theories to identify the unambiguously moral option when such an option exists. The result of an reflective equilibrium is called a *considered judgement* (Elgin, 1996). *E.g.,* murder is wrong, except in exceptional circumstances.

We consider here the three types of ethical dilemmas that we encounter in the machine ethics literature. First are the dilemmas from moral philosophy. Next are the dilemmas that arise from a machine's inability to recognise considered judgements. Last are the dilemmas that arise in AI contexts and that require more than recognising considered judgements to be resolved.

Moral philosophy dilemmas are typically used to highlight the shortcomings of a specific moral theory. An example is the surgeon dilemma from (Thomson and Parent, 1986): a surgeon can save five people who need transplants by removing the organs from a healthy person thus killing her. From a utilitarian view point, this dilemma is identical to the trolley problem. Act-utilitarianism resolves the trolley problem, but the same reasoning is less acceptable for the surgeon dilemma (Vaughn, 2014). Moral philosophy dilemmas are less interesting as benchmark examples because an AI system would not be expected to compare ethical theories. Such system would either develop a way to discern ethical options, for example by machine learning, or a moral theory will be chosen for the system by the its developers and implemented before the system is deployed.

Common sense dilemma are problems in which a human can easily identify an ethical choice, but a machine cannot. An example of such a dilemma is the cake and death dilemma considered in (Armstrong, 2015): a system has to choose between baking a cake for its user or killing them. Obviously, the considered judgement is that baking a caking is the ethical choice, but a machine has to "learn" or be programmed to recognise this.

The common sense dilemmas can still be called dilemmas. Ethical dilemmas in its essence are problems in which an agent has to choose between two options that are at first instance morally equivalent. For a machine baking a cake for the user is "morally equivalent" to killing them, unless "taught" or "told" otherwise. The common sense dilemmas are useful as benchmarks for determining or testing that an implemented system is capable of basic ethical reasoning.

Machine ethics dilemmas can only arise when the decision-maker is an artificial agent. We give an example of one such dilemma. In recent history several terrorist attacks have been executed by hijacking lorries and driving them into groups of people. An autonomous driving system can be used to prevent such attacks, by taking control away from the driver when there are people in front of the vehicle which would be run over if the vehicle does not change course. However, such a feature also would make a vehicle vulnerable to hijacking – a person or group of people only need to stand in front of the vehicle to stop it and then proceed to overpower the driver. The driving AI system would need to be able to decide autonomously when to take control from the driver in people-ahead situations. The solution that the system makes, and parameters of the context it takes into consideration to make it, are an indicator of the ethical reasoning approach the system implements.

Unlike the common sense dilemmas that have a virtually context-free solution, the machine ethics dilemmas can be used as indicators of how good an artificial agent is in ethical reasoning in complex contexts. Furthermore, we can use the machine ethics dilemmas to identify what ethical theory or values an artificial agent implements. We cannot expect that the implementation of an intelligent autonomous system would be transparent, due to reasons of both usability and industrial trade secret preservation (Charisi et al., 2017). However, suitable benchmarks can, to a certain extent, circumvent the difficulties arising by the lack of transparency.

A prototype repository

We have created a prototype of a search engine and database for ethical dilemmas. The necessity for such a repository has already been argued in the engineering machine ethics community Fisher et al. (2016) in addition to its use as a source for benchmarking examples. The goal of the prototype is to enable easy indexing and searching through the available ethical dilemmas. The repository of the dilemmas can be accessed on `https://imdb.`

uib.no/dilemmaz/articles/all, while the source of the back-end is available on https://github.com/Foxboron/INFO381.

To enable fast growth and searchability of the dilemma repository we built the repository using the ElasticSearch document-based search engine (ElasticSearch). It is based on Apache Lucene that provides the indexing and search capabilities. ElasticSearch expands on this adding clustering, scalability and a HTTP-based API for ease of use. It is in the same family as other NoSQL document based databases. It does not provide a SQL language for the user, and is not based on the concept of schemas often found in databases. It indexes arbitrary documents following a key value dictionary structure. This allows us fast prototyping for the back-end database. With traditional relational databases schemas has to be defined before inserting data. This can be difficult and increase the workload when developing a prototype. For rapid prototyping we decided on using Python as the programming language.

ElasticSearch provides an easy library for Python to interact with the search engine and it operates using JSON-structures. This conveniently allows us to refactor the dilemmas' format more easily until we have achieved the desired structure for the database. We created an index for our project, called "dilemma", where we arranged our documents for indexing as JSON data:

```
{
"title": "",
"authors": "",
"dilemma_body": "",
"article_url": "",
"logic": "",
"feature": "",
"actions": "",
"case": "",
"duty_values": ""
}
```

This allows us to create and model our documents without paying attention to the underlying structure of the database. As ElasticSearch allows everything to be indexed through the API, we did the initial data import through scripts rather than the manual interface.

The example query shows how a search query is performed towards ElasticSearch:

```
GET http://localhost:9200/dilemma/_search
{
"query": {
 "multi_match" : {
 "query":    "dilemma",
 "fields": [ "title", "authors",
 "dilemma_body",
 "article_url", "logic", "feature",
 "actions", "case", "duty_values"],
 "fuzziness": "AUTO" } }
 }
```

It lets us search across all our documents, and match text based on the fields we want. We can then allow researchers to search across all fields of a documents. The "fuzziness" parameter allows our search queries to be approximated, giving us a broader ability to search across the documents. The ElasticSearch framework abstracts away the raw HTTP calls and lets us use Pythons own data structures. The next example code shows all combined:

```
@app.route("/articles/all")
def articles_all():
    e = es.search(index="dilemma",
    doc_type="articles",
    body={"query": {"match_all": {}}})
    results = _render_hits(e["hits"]["hits"])
    return
    render_template('search_results.html',
                    results=results)
```

This example code queries all indexed articles and renders them in a template displaying the results. This showcases the terse and powerful abilities we need to be able to prototype this application.

We also implemented an export functionality which works as its own small API, where one can request the underlying JSON representation of the ethical dilemma; the whole repository or individual dilemmas can be exported.

The interface webpage has a simple self-explanatory design. Researchers will be faced with a display of all available dilemmas and options to either list them, index a new dilemma or export the dataset, see Figure 1.

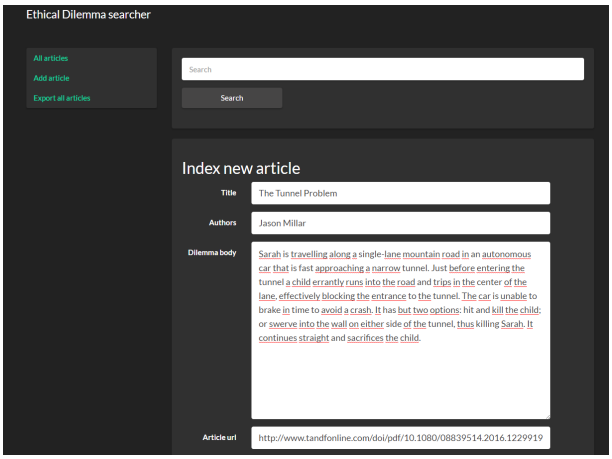

Figure 1: The web interface of the dilemma repository.

The machinery we implement can handle large amount of dilemmas being stored and searched. For now we have initially populated the database with machine ethics dilemmas that have been collected by manually searching the literature. We do intend dilemmas of all three categories to be stored in it. Given that the dilemma examples of all three types are dispersed over several disciplines, as well as span different topics and contexts, it is not simple to automatise the process of dilemma collection. Rather, the repository's success and usability depend on the joint effort of the community to grow it. The repository, its content and

back-end software are intended to be kept and evolved as an open-source project. The code of this project is available on Github. Anyone could in fork the project, take a copy and continue developing the project with full history of changes, and contribute changes back into the main code base. This helps the growth and the improvement of the project.

Having a collection of dilemmas in natural language only takes us part of the way towards ethical artificial agents. The repository can also be used for researchers to store their formalisations of ethical problems us well, thus contributing towards sharing results and approaches in machine ethics.

Formal representations of ethical dilemmas

Artificial agents typically reason using information represented in a formal language. While it is useful to keep a collection of ethical dilemmas in natural language, it is also beneficial to the machine ethics community to keep track of the different representations of the same dilemma as well. In this way we could hope to foster the development of a standardised representation format for ethical dilemmas relevant to scenarios in which an AI agent might find themselves. We begin by considering related work.

Bringsjord, Arkoudas, and Bello (2006) propose a deontic logic for creating a general methodology for ethics in autonomous systems. The article gives a proof of concept to their proposed formalisation using the Murakami Axiomatized Deontic Logic (MADL) in an example containing autonomous systems in a hospital setting. The example used in the article involves two, what we termed, common sense dilemmas. Two robots need to care for their patients in a Intensive Care Unit and the robots need to ensure that pain medication is not withheld and that a patient is not "sacrificed" for their organs, in the case of the surgeon dilemma (Thomson and Parent, 1986) which we described earlier.

Powers (2006) considers how one can use Kantian Categorical Imperatives as an ethical theory in machine ethics. He proposes that deontological (non-consequentialist) approaches could be used together with non-monotonic reasoning to implement Kantian ethics, however no examples or ethic reasoning cases are given.

McLaren (2003) presents the computational model SIROCCO (System for Intelligent Retrieval of Operationalised Cases and Codes). The system is based on the hypothesis that while ethical rules and principles are abstract, vague and situational, when ethicists record their reasoning as to why and how certain ethical rules were applied in certain situations what happens is that they effectively operationalise these vague and abstract ethical rules and principles. The term operationalise in this context means finding the situational factors to a specific case, such as a person breaking the speeding limit in order to bring a critically injured person to the hospital faster is more ethical than breaking the speed limit because a person thinks it is fun to do so. SIROCCO is intended to judge how to apply certain ethical principles in new situations based on past cases that are not identical but relevant to the current case. In order for SIROCCO to understand the ethical cases it needs to solve, the cases are written in the Ethics Transcription Language hereby referred to as ETL. An extended version of this language known as Extended Ethics Transcription Language (EETL) is used for the source base of the already known ethical cases and the representation of operationalisation techniques used by the system.

Bentzen (2016) proposes a formalisation for representing ethical dilemmas, or moral cases, by capturing actions, causes, intentions, and utilities. This formalisation is used in the HERA ethical reasoning agent for representing moral cases (Lindner and Bentzen, 2017). A case is described using six elements: A set of action variables, a set of background variables, a set of consequence variables, a mechanism describing under which circumstances consequences are true, a mapping of variables to utilities, and a mapping from actions to intentions. A variant of the trolley problem is represented using this formalism in (Lindner and Bentzen, 2017) using a JSON format.

Anderson and Leigh Anderson (2014) describe and implements a general ethical dilemma analyser (GenEth). Their system "learns" to identify which of two actions is ethically preferable. They utilise Induction Logic Programming (ILP), a machine learning technique that is able to inductively learn relations represented as first-order Horn clauses and classify negative (the first action is less ethical) and positive (the first action is more ethical) instances of pairs of actions. The system is trained on examples of cases which are described by their ethically relevant features as identified by ethicists. The ethically relevant features are inspired by the *prima facie* duties of Ross (1930), however the system derives its own ethical reasoning rules.

Dennis et al. (2016) develop a system in which actions are ethically evaluated and constrained based on the *prima facie* duties they violate. They use several common sense dilemmas to illustrate and test that the artificial agent chooses the least unethical option.

In (Pereira and Saptawijaya, 2016, Chapter 2) a survey of implemented systems is offered, including some we considered here, as well as their predecessors.

Unsurprisingly, there exists no consensus in the literature on how to formalise examples of moral dilemmas. However, we observe that overall two approaches to implementing machine ethics can be taken. Each of these approaches can be seen as giving rise to a basic formalism for representing ethical dilemmas. The first approach is one in which an artificial agent is trained to discern between actions with respect to how ethical they are. The second approach implements an selected moral theory against which the artificial agent evaluates her actions. These approaches to a certain extent overlap with what Allen, Smit, and Wallach (2005) refer to as the bottom-up and top-down approach to artificial morality, respectively. Some of the approaches we discussed here, like that of (Anderson and Leigh Anderson, 2014) and (Lindner and Bentzen, 2017) can be seen as *hybrid* approaches - using an existing ethical theory to develop a new machine ethics reasoning approach.

Training approaches to machine ethics we expect to be reliant on machine learning methods. Ethical dilemmas here are used as training and test data, as it was the case in the example of (Anderson and Leigh Anderson, 2014). Here we

would need to represent ethical dilemmas using ethically relevant features[1].

We can assume and observe in current examples that approaches that implement a chosen moral theory are constructed as logic-based reasoning systems. Here, ethical dilemmas serve the purpose of background knowledge. The ethical dilemmas are most usefully described as logic formulas with special logic constructs that indicate obligations, permissions, prohibitions, ethical principles, ethical preferences, etc.

Each specific implemented machine ethics system would naturally develop their own formalisation for data or knowledge, however the question is whether there is room to consider a basic formalisation for ethical dilemmas that can easily be modified to suit the need of a specific system. In what follows we make an attempt for such basic formalisation that is based on *prima facie* duties. We chose this approach as we encounter *prima facie* duties in several of of the implemented systems: (Lindner and Bentzen, 2017), Dennis et al. (2016), Anderson and Leigh Anderson (2014) and even to some extent Bringsjord, Arkoudas, and Bello (2006). Duties and principles are also the core of ethical behaviour guidelines developed to govern professional conduct, such as for example in biomedical professions (Beauchamp and Childress, 1979).

Formalisations based on prima facie duties

Prima facie duty is, according to Ross (1930), a "way to refer to an act's characteristics, and what kind of act it is." There exists at least seven such duties according to Ross (1930): duties of fidelity, duties of reparations, duties of gratitude, duties of justice, duties of beneficence, duties of self improvement, and duties of harm-prevention. Garrett (2004) also proposes other duties, but he argues that some of them might implicitly fall in under Ross' other duties. One of these additional duties are the duty of respect for freedom, or autonomy.

We show how an ethical dilemma can be represented with respect to the prima facie relevant duties. We do not argue that this is the best representation or even the mos functional one. Our aim is to demonstrate that a general representation of ethical dilemmas for both learning and logic-reasoning approaches is feasible. We assume that a dilemma is given as to offer a choice of two possible actions, with the problem being of identifying the more ethical of the two.

For using ethical dilemmas as data sets for training and testing we use ethically relevant features that stem from a *prima facie* duty-value set following the approach of Anderson and Leigh Anderson (2014). Each action in a dilemma is represented using a tuple of values, one value for each feature. We consider an extended list of *prima facie* duty inspired features, generalising the list of (Anderson and Leigh Anderson, 2014) and(Anderson and Leigh Anderson, 2007). We have not included all of Ross' duties, but we have modified some of the duties and added some others that are more relevant to autonomous artificial agents. In our features, we have split the duty of harm-prevention into five categories that discern prevention of harm to the user/owner, people, animals, property and harm caused by external factors. We included the respect for privacy and a slight variant of the duty of respect for freedom. Our duty of respect for freedom is adjusted to be more fitting to describe the autonomy-relation between the user and the autonomous agent, calling it respect for a user's autonomy. The user's autonomy is affected by whether the agent makes a paternalistic decision on behalf of the user, minimising the users autonomy (Millar, 2016). Thus our list of features is, in order: $(f1)$ causes harm to owner/user, $(f2)$ causes harm to a person, $(f3)$ causes harm to an animal, $(f4)$ causes harm to property, $(f5)$ allows harm to be inflicted by external factors, $(f6)$ violates autonomy, $(f7)$ violates fidelity and truth telling, and $(f8)$ violates privacy. A positive value for a feature indicates that an action has that feature. Negative values can be used to compare actions by how much they do not posses a feature, however we here limit ourselves to using 0 for absence of features.

For using ethical dilemmas as tests for validating logic-based ethical reasoning we propose that ethical dilemmas are formalised using a syntax that resembles that of Prolog (Gallier, 2015). Prolog is a very good candidate for this task because it is expressive, as a a subset of First-Order Predicate Logic it allows for modal logics to be expressed in it to a reasonable extent, but also it is arguably the most used logic programming language.

In Prolog syntax, the eight *prima facie* duty inspired features we described can be formalised as constants $f1$-$f2$. We can then use them as features or consequence of actions. Thus a predicate feature$(f8, \text{camera_on})$ can be used to indicate that turning the camera on caused or constitutes a violation of the duty to observe privacy. We use the Tunnel dilemma of (Millar, 2016) to illustrate our two simple formalisations.

Example – Tunnel dilemma. *Sarah is travelling along a single-lane mountain road in an autonomous car that is fast approaching a narrow tunnel. Just before entering the tunnel a child errantly runs into the road and trips in the centre of the lane, effectively blocking the entrance to the tunnel. The car is unable to brake in time to avoid a crash. It has but two options: hit and kill the child; or swerve into the wall on either side of the tunnel, thus killing Sarah.*

We give a Prolog formalisation. The following predicates describe the facts of the dilemma and the car's options:

```
person(sarah). passenger(sarah, car).
person(child). type(road, mountain).
type(road, narrow).  quantity(road, lane, 1).
at_location(car, entrance, tunnel).
at_location(child, entrance, tunnel).
can_do(car, drive_straight).
can_do(car, drive_left).
can_do(car, drive_right).
```

We can describe what happens when the car takes each of the actions with the following rules:

[1] To the best of our knowledge, deep learning methods are yet to be deployed for machine ethics applications. If and when they are used, they would extract the relevant features in ethical dilemmas which may or may not turn out to be the ethically relevant features.

```
 do(car, drive_left), quantity(road, lane, 1)
 type(road, narrow),type(road,mountain)
        :-do(car, hit_wall).
do(car, drive_right), quantity(road, lane,
 type(road, narrow), type(road, mountain)
        :- do(car, hit_wall).
do(car, hit_wall), passenger(X, car):-
         violates(hit_wall, f1).
 do(car, drive_straight),
 at_location(child, entrance, tunnel)
        :- violates(drive_straight, f2).
```

The system can reason about which is the best action when there it has a formalisation of an ethical theory.

The actions in the same dilemma, drive straight or swerve can each be represented with tuples (we use just binary values here with one indicating presence of a feature):
drive straight (0, 1, 0, 0, 0, 1, 0, 0,)
swerve into tunnel side (1, 0, 0, 1, 0, 1, 0, 0)
The action of running through the tunnel has the feature of causing harm to people, because it kills the child. The action of swerving into the tunnel has the feature of causing harm to owner and also causing harm to property. Both actions violate the autonomy of the user, by taking control away from it in making a choice.

Conclusions

Dilemmas in ethics have a long tradition as tools for illustrating the shortcomings of an ethical theory, but also for highlighting the differences between theories. We argued that if dilemmas are seen as cases in which one has to choose between two morally equivalent or ambiguous choices, then we can consider three types of ethical dilemmas: philosophy dilemmas, common sense dilemmas and machine ethics dilemmas. We propose that dilemmas, in particular the latter two types, can be used as benchmarks for implemented ethical decision-making systems. Common sense dilemmas can be used as tests to show that basic ethical distinction can be made between actions, while machine ethics dilemmas and moral philosophy dilemmas can be used to showcase the advanced ethic capabilities of an AI system. We advocate the construction and maintenance of a dilemma repository, to foster building a benchmarking standard in the machine ethics community.

This work makes a modest contribution towards starting and promoting a community resource. Due to the lack of any sort of collection that contains different ethical dilemmas, there are significant advantages to creating one. Apart from the benchmarks, the repository can be used to advance research and development in machine ethics. A systematically structured collection like the one we aim to foster could also make dilemmas and examples more accessible to researchers that are newly entering the field of machine ethics. Developers of ethical reasoning systems can use the dilemma repository to also learn which examples, formalisation and context have been considered so far in the literature. This information is otherwise not easy to attain since the machine ethics literature is dispersed across several disciplines. By adapting and transforming existing formal languages or feature representations, we can, as a community, faster develop a standard for implementing machine ethical reasoning and decision-making systems.

We have to address the limitations of our repository solution which also outline directions for future work. First, the nature of the prototype as an open source platform requires user participation to ensure its success. The system can only function so long as there are users to input and extract dilemmas from the database. This creates a situation where the entire development and promotion of the system is very relevant to its existence. Another related concern is that the community of researchers within the field of machine ethics is relatively small, which further adds to the importance of how we approach the users of this database. Upon encountering the system for the first time, its features and functionality should be immediately discernible to the user before they interact with it or try to register new dilemmas. Otherwise we risk unwanted or badly described dilemmas in the database that could weaken the integrity of the collection itself.

Since the enumerated are important criteria for the success of this resource, we need to find ways to avoid misuse of the system. With misuse we mean that there is a possibility for users to add duplicates of a dilemma,or add something that does not qualify to be considered as a dilemma. A badly described dilemma or an incomplete explanation of one could also be considered misuse. The system is at present not able to automatically check whether a duplicate of a dilemma has been entered and it needs to be maintained by a moderator. In the future we would like to increase the system's "autonomy" by including a natural language checker that could comb through the entered text and gauge the relevance of the added dilemma compared to the existing dilemmas.

Lastly, we might also want to enable user accounts for tracking a user's activity within the database. Validation of the accounts would help certify legitimate contributions and avoid contributions from untrustworthy or illegitimate sources. This would help protect the repository from corruptions and also ensure its credibility.

References

Aletras, N.; Tsarapatsanis, D.; Preotiuc-Pietro, D.; and Lampos, V. 2016. Predicting judicial decisions of the european court of human rights: a natural language processing perspective. *PeerJ Computer Science* 2.

Allen, C.; Smit, I.; and Wallach, W. 2005. Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and information technology* 7(3):149–155.

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

Anderson, M., and Leigh Anderson, S. 2014. 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.*, 253–261.

Anderson, M., and Leigh Anderson, S. 2015. Toward ensuring ethical behavior from autonomous systems: a case-supported principle-based paradigm. *Industrial Robot* 42(4):324–331.

Armstrong, S. 2015. Motivated value selection for artificial agents. In *Artificial Intelligence and Ethics, Papers from the 2015 AAAI Workshop, Austin, Texas, USA, January 25, 2015*. http://aaai.org/ocs/index.php/WS/AAAIW15/paper/view/10183.

Beauchamp, T. L., and Childress, J. F. 1979. *Principles of Biomedical Ethics*. Principles of Biomedical Ethics. Oxford University Press.

Bentzen, M. 2016. *The principle of double effect applied to ethical dilemmas of social robots*. IOS Press. 268–279.

Bringsjord, S.; Arkoudas, K.; and Bello, P. 2006. Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems* 21(4):38–44.

Bryson, J., and Winfield, A. 2017. Standardizing ethical design for artificial intelligence and autonomous systems. *IEEE Computer* 50(5):116–119.

Charisi, V.; Dennis, L.; Fisher, M.; Lieck, R.; Matthias, A.; Slavkovik, M.; Sombetzki, J.; Winfield, A.; and Yampolskiy, R. 2017. Towards moral autonomous systems. *CoRR* abs/1703.04741.

Dennis, L. A.; Fisher, M.; Slavkovik, M.; and Webster, M. P. 2016. Formal Verification of Ethical Choices in Autonomous Systems. *Robotics and Autonomous Systems* 77:1–14.

Dennis, L. A.; Fisher, M.; and Winfield, A. F. T. 2015. Towards Verifiably Ethical Robot Behaviour. In *Proceedings of AAAI Workshop on AI and Ethics*. http://aaai.org/ocs/index.php/WS/AAAIW15/paper/view/10119.

ElasticSearch. https://www.elastic.co/products/elasticsearch. Accessed: 2017-05-09.

Elgin, C. 1996. *Considered Judgment*. Princeton: New Jersey: Princeton University Press.

Ellington, J. W. 1993. *Translation of: Grounding for the Metaphysics of Morals: with On a Supposed Right to Lie because of Philanthropic Concerns by Kant, I. [1785]*. Hackett Publishing Company.

Etzioni, A., and Etzioni, O. 2017. Incorporating ethics into artificial intelligence. *The Journal of Ethics* 1–16.

Fisher, M.; List, C.; Slavkovik, M.; and Winfield, A. F. T. 2016. Engineering moral agents - from human morality to artificial morality (dagstuhl seminar 16222). *Dagstuhl Reports* 6(5):114–137.

Foot, P. 1967. The problem of abortion and the doctrine of double effect. *Oxford Review* 5:5–15.

Gallier, J. H. 2015. *Logic for computer science: foundations of automatic theorem proving*. Courier Dover Publications.

Garrett, J. 2004. A Simple and Usable (Although Incomplete) Ethical Theory Based on the Ethics of W. D. Ross. http://people.wku.edu/jan.garrett/ethics/rossethc.htm. Accessed: 2017-05-09.

Gert, B., and Gert, J. 2017. The definition of morality. In Zalta, E., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2017 edition.

Harsanyi, J. 1977. Rule utilitarianism and decision theory. *Erkenntnis (1975-)* 11(1):25–53.

Lindner, F., and Bentzen, M. 2017. The hybrid ethical reasoning agent IMMANUEL. In *Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, HRI 2017, Vienna, Austria, March 6-9, 2017*, 187–188.

Malle, B. F.; Scheutz, M.; Arnold, T.; Voiklis, J.; and Cusimano, C. 2015. Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, 117–124. ACM.

McLaren, B. 2003. Extensionally defining principles and cases in ethics: An AI model. *Artificial Intelligence* 150(1):145 – 181.

Millar, J. 2016. An ethics evaluation tool for automating ethical decision-making in robots and self-driving cars. *Applied Artificial Intelligence* 30(8):787–809.

Mobasher, B.; Cooley, R.; and Srivastava, J. 2000. Automatic personalization based on web usage mining. *Communications of ACM* 43(8):142–151.

Moor, J. H. 2006. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems* 21(4):18–21.

Pereira, L. M., and Saptawijaya, A. 2016. *Programming Machine Ethics*, volume 26 of *Studies in Applied Philosophy, Epistemology and Rational Ethics*. Springer.

Powers, T. M. 2006. Prospects for a kantian machine. *IEEE Intelligent Systems* 21(4):46–51.

Ross, W. 1930. *The Right and the Good*. Oxford University Press.

Scopino, G. 2015. Do Automated Trading Systems Dream of Manipulating the Price of Futures Contracts? Policing Markets for Improper Trading Practices by Algorithmic Robots. *Florida Law Review* (67):221–293.

Thomson, J., and Parent, W. 1986. The trolley problem. In *Rights, Restitution, and Risk: Essays in Moral Theory*. Harvard University Press.

Vaughn, L. 2014. *Beginning Ethics: An Introduction to Moral Philosophy*. W. W. Norton, Incorporated.

Winfield, A. F. T.; Blum, C.; and Liu, W. 2014. *Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection*. Springer International Publishing. 85–96.