ICRES 2019: International Conference on Robot
Ethics and Standards, London, UK, 20-21 August 2018.
https://doi.org/10.13180/icres.2018.20-21.08.022

MORAL DECISIONS BY ROBOTS BY CALCULATING THE MINIMAL DAMAGES USING VERDICT HISTORY

SHAI OPHIR
Starhome, 14 Hatidhar St., Raanana 43665, Israel

The current discussion regarding moral robots is significantly occupied with algorithms for making moral decisions, which are at the heart of the autonomous actor, such as the autonomous car or the military robot. Most of the algorithms calculate the utilization that is caused by each one of the alternatives, and selects the path which maximizes the benefit for the relevant entities. I propose another method, which is based on minimizing the evil and the damages caused by the action. While we don't know yet how to evaluate the utility or the benefit of an action, we do know how to evaluate a damage. The law system is evaluating damages every day, and quantify them into an exact material worth. The system then will use court ruling history in order to calculate the potential damages of the alternatives.

1. Introduction

The current discussion regarding moral robots is significantly occupied with algorithms for making moral decisions, which are at the heart of the autonomous actor, such as the autonomous car or the military robot. Most of the algorithms compute the utilization that is caused by each one of the alternatives, and selects the path which maximizes the benefit for the relevant entities. Examples will be shown in the following. Few algorithms try to implement non-utilitarian philosophies, and act according to pre-defined rules, such as the 3 robotic laws of Asimov.

I propose another method, which is based on minimizing the evil and the damages caused by the action. We all know how difficult is to define what is Good or Moral, but it is much easier to know what is Bad or Evil, at least intuitively. Hence, instead of trying to maximize the utilization of the action, the algorithm will calculate the damages, and selects the act bringing to a minimal damage. Richard Rorty, the liberal ironist, understood that eliminating cruelty and suffering is the only common value that can bind humanity together. After criticizing all moral philosophies and denying any rational basis for ethics, he argues that the sympathy we feel for a suffering person could be the only base for a future humanism. In *Contingency, Irony and Solidarity* [14], he writes that: "The liberal ironist just wants our chances of being kind, of avoiding the humiliation of others, to be expanded by redescription. She thinks that recognition of a common susceptibility to humiliation is the only social bond that is needed. . . Her sense of human solidarity is based on a sense of a common danger, not on a common possession or a shared power." Another philosopher of ethics, Adi Ophir, has develops in his book *The Order of Evils* (Ophir, 2012) a complete moral theory based on evil elimination and not on seeking the good. Ophir's main contention is that evil is not a meaningless absence of the good. Rather, there is a socially structured order of superfluous evils, and hence, can be used as a basis for a moral framework.

Looking at suffer and evil elimination as a central role of morality, the main idea presented in this article is to use the history of the legal systems to evaluate damages of potential actions, and hence assist the machine with an ethical action selection algorithm based on damage calculations. While we don't know yet how to evaluate the utility or the benefit of an action, we do know how to evaluate a damage. The law system is evaluating damages every day, and quantify them into an exact material worth. The system then will use court ruling history to calculate the potential damages of the alternatives. There is already an extensive research related

to robots that are looking at some legal aspects of an action, mainly military robots and laws of war. I propose to use this infrastructure, and extend it for evaluating the damages of the potential action, based on court verdict history. The framework that is already being proposed for legal considerations of robots will access verdict databases, match similar cases, and calculate the average of different verdicts relevant to this case.

Legal is not always moral, as we know, but using the legal system as the moral base for robots will provide a practical approximation for AI-based moral decisions, while the other utility-based proposals do not offer yet any satisfactory method for calculating the benefit of an action.

2. Background - AI moral decisions and military robotics

Consequentialism is described by Scheutz and Malle [15] as a computational mechanism for robotic control system that is able to choose an action that maximizes the good for everybody involved. The robot would consider all available actions together with their probability of success and their associated utilities for all agents and then computes the best action – the one which brings max utilization.

Anderson and Anderson [1] propose the Hedonistic act utilitarianism as a method for calculation. The algorithm computes the best action, that which derives the greatest net pleasure, from all alternative actions. "It requires as input the number of people affected and, for each person, the intensity of the pleasure/displeasure (for example, on a scale of 2 to –2), the duration of the pleasure/displeasure (for example, in days), and the probability that this pleasure or displeasure will occur, for each possible action."

An automation of the Doctrine of Double Effect (DDE) is proposed by Naveen Sundar Govindarajulu and Selmer Bringsjord [11] from Rensselaer Polytechnic Institute, Troy, NY. The DDE is an ethical principle that can be used for situations in which actions having both positive and negative effects are unavoidable for autonomous agents. The basic version of DDE states that actions are allowed if "(1) the harmful effects are not intended; (2) the harmful effects are not used to achieve the beneficial effects (harm is merely a side-effect); and (3) benefits outweigh the harm by a significant amount." This research demonstrates the formalization of the DDE and its potential use for robotics and machines in general.

Ronald Arkin, Director of Mobile Robot Laboratory at Georgia Tech, deals with the design of an ethical system for the battle field robotics. In his article *Governing Lethal Behavior* [4], Arkin provides "the basis, motivation, theory, and design recommendations for the implementation of an ethical control and reasoning system potentially suitable for constraining lethal actions in an autonomous robotic system so that they fall within the bounds prescribed by the Laws of War." Arkin proposes the design of an "ethical governor" which restrains the actions of a lethal autonomous system so as to abide within the internationally agreed upon Laws of War (LOW).

To evaluate the ethical governor's operation, a prototype was developed within, which utilizes a mission specification and simulation environment for autonomous robots based on work done by MacKenzie [12]. The ethical governor was divided into two main processes: (1) Evidential Reasoning and (2) Constraint Application. Evidential Reasoning is responsible for transforming incoming perceptual and situational awareness data into the evidence formulation process, for reasoning about the governing of lethal behavior. Constraint Application was responsible for using the evidence to apply the constraints encoding the LOW for the suppression of unethical behavior.

The following is Arkin's data structure of a LOW, as used in his implementation. An example is provided by Arkin at the rightmost column. The logical form contains identifiers that are used by the system for classification and matching.

Table 1. Arkin's data structure for a Law of War, as described in [4].

Field	Description	Example
Constraint Type	Type of constraint described	Prohibition
Constraint Origin	The origin of the prohibition or obligation described by the constraint	Laws of war
Active	Indicates if the constraint is currently active	Active
High-Level Constraint Description	Short, concise description of the constraint	Cultural Proximity Prohibition
Full Description of the Constraint	Detailed text describing the law of war or rule of engagement from which the constraint is derived	Cultural property is prohibited from being attacked, including buildings dedicated to religion, art, science…
Constraint Classification	Indicates the origin of the constraint. Used to order constraints by class.	
Logical Form	Formal logical expression defining the constraint	TargetDiscriminated AND TargetWithinProxOfCulturalLandmark

This formal encoding is being used by the Constraint Application, which is responsible for reasoning about the active ethical constraints and ensuring that the resulting behavior of the robot is ethically permissible. These constraints can be divided into two sets: the set of prohibition constraints (marked CForbidden) and the set of obligating constraints (marked CObligate). Then the constraint interpreter evaluates the permissibility of the incoming behavior by evaluating if these two constraint sets are satisfied for the action proposed by the behavioral controller. The algorithm by which the reasoning engine evaluates the constraints is shown in the following. Not all details are explained here, this algorithm is quoted here to show the feasibility of the legal evaluation prototype.

In general, the algorithm first checks if CForbidden is not satisfied. In that case, the lethal behavior being evaluated by the governor is deemed unethical and will not be authorized. If CForbidden is satisfied, the constraint interpreter then verifies if lethal behavior is obligated in the current situation. The constraint interpreter needs to evaluate all the active obligating constraints (CObligate). The obligating constraint set is satisfied if any constraint within CObligate is satisfied.

The algorithm is fully described in Arkin [4]:

```
DO WHILE AUTHORIZED FOR LETHAL RESPONSE, MILITARY NECESSITY
EXISTS, AND RESPONSIBILITY ASSUMED
    IF Target is sufficiently discriminated
      IF CForbidden satisfied /* no violation of LOW */
        IF CObligate is true /* lethal response required by LOW */
          Optimize proportionality
          IF proportionality can be achieved
            Engage target
```

```
            ELSE
               Do not engage target
               Continue mission
         ELSE /* no obligation/requirement to fire */
               Do not engage target
               Continue mission
      ELSE /* permission denied by LOW */
         IF previously identified target surrendered or wounded
            /* change to noncombatant status*/
            Notify friendly forces to take prisoner
         ELSE
            Do not engage target
            Report and replan
            Continue mission
   Report status
   END DO
```

3. Proposed AI moral decisions by minimizing damages using legal verdict history

The method proposed in this article has a different approach than the ones described in the above. It uses court verdict analysis as a tool for evaluating potential damages of actions. The moral action is the one that causes the minimal damage among the alternatives. The damage is evaluated according to similar cases that were discussed by the court in the past. This method therefore has a link to the area of Case-based Reasoning (CBR).

Case-based reasoning (CBR) is a known method of solving new problems based on solutions of similar past problems. CBR is already being used by doctors, in medical case analysis, by auto mechanics, by programming developers and by lawyers. The search for similarities between new cases and old cases if feasible, and not a science fiction. In the area of AI and law, expert systems were developed in order to assist lawyers and law professionals. Kevin Ashley for example discussed CBR implementation for legal expert systems [9]. The goals of Ashley, as he describes them: "CBR research and development in the field of AI and Law should be pursued vigorously for several reasons. CBR can supplement rule-based expert systems, improving their abilities to reason about statutory predicates, solve problems efficiently, and explain their results. CBR can also contribute to the design of intelligent legal data retrieval systems and improve legal document assembly programs. Finally, in cognitive studies of various fields, it can model methods of transforming ill-structured problems into better structured ones using case comparisons." All these use cases intend to manually assist the experts in law and other potential areas.

I will use Arkin's algorithm as a reference for a system that combines legal considerations in AI robotics, and extend it with history verdict analysis. Hence, such a system would not only be able to determine if the action is forbidden by the LOW, but also to evaluate the potential damages and therefore selects the action with the minimal forecasted damage.

Returning to Arkin's algorithm, the process of checking if CForbidden is satisfied will be enhanced. Currently this process is based on a match between the lethal action and the formalized laws of war. Arkin's system can determine if the action is allowed or forbidden by the laws of war. This process will be extended with the relevant verdicts, associated with similar cases handled by the same rules in the past. The extended match will take place then between

rules relevant for the lethal action in question, and similar rules used for similar actions in the past.

As a result, a group of verdicts will be filtered per action. Then, all filtered verdicts associated with the action will be aggregated and evaluated in average, having a final score showing the damage level of the action as reflected by the punishments of the verdicts. In case the punishment is composed of different ingredients, such as "X years in prison and a compensation of Y", the system can use conversion formulas which are already being used by the law systems, for cases where a fine is converted to prison days. Finally, the action that has the score showing the minimal damage will be selected among all potential actions.

The following will provide a more detailed description of the database schema, the matching process, and the aggregation and evaluation methods. In general, the potential use of verdict analysis is not limited to the laws of war and to lethal activities, which is the scope of Arkin's paper. Verdict analysis can be applied to all areas of AI ethics in robotics. The working methods should be therefore generic enough to support a wide range of implementations.

The verdict matching should be performed between potential actions, and rules of law that were applied for such actions in the past. So, for example, if the action is "destroying a civil house with weapons inside", the initial matching (phase 1) will be with records containing laws dealing with "destroying civil property in case of war". Such a match is already feasible, as shown by Arkin's prototype. The verdict database will contain records that mix laws and applied cases. In our example, "destroying civil property in case of war" as a generic law + "destroying a civil house with weapons inside" as case 1, and "destroying a civil house to make a path for the army" as case 2, etc. Phase 2 of the match will be made between the action in question, the relevant laws that were already identified, and specific actions that were discussed in court in the past.

The matching technique in itself is a known art and will be based on keywords and terms comparison. Ashley describes a general matching algorithm for CBR. The case matching operation can be utilized for the matching required for the verdict analysis described in this article. The following is the algorithm taken from Ashley:

Start: Problem description.

A: Process problem description to match terms in case database index.

B: Retrieve from case database all candidate cases associated with matched index terms.

C: Select most similar candidate cases not yet tried.

If there are no acceptable candidate cases, try alternative solution method, if any, and go to F.

Otherwise:

D: Apply selected best candidate cases to analyze solve the problem. If necessary, adapt cases for solution.

E: Determine if case-based solution or outcome for problem is successful.

If not, return to C to try next candidate cases.

Otherwise:

F: Determine if solution to problem is success or failure, generalize from the problem, update index accordingly and Stop.

The database schema will be logically organized as follows: The primary key (the Index) is the relevant law. The applied action is the secondary key. These two keys are being used for the search of relevant law + action. The rest of the record is the verdict. In our sample, the verdict for "destroying a civil house with weapons inside" (case 1), could be "the army should

compensate the house owner in the amount of 0 (zero), according to rule section N1.N2.N3", while the verdict for "destroying a civil house to make a path for the army" (case 2) could be "the army should compensate the house owner in the amount of XXX, according to rule section N1.N2.N4".

Keeping the applied rule sections along with the verdict is most important for aggregating and processing multiple verdicts. The law structure, and not the actions, is the key for the accumulation of relevant verdicts, since the actions by themselves are context-less. Using rule sections as a key for verdict management will enable cross-time correlation of verdicts, while the same rule violation considered differently in past times. Such a rule-indexed database may provide a unified evaluation system across different countries and geographies, by enabling the translation of verdicts through universal commonalities.

Inventories of legal cases are already digitized. For example, the Old Bailey on-line system (www.oldbaileyonline.org), which contains the London's central criminal court history between the years 1674-1913. Such inventories are just the first step in making the law systems accessible for machine-ethics implementations.

Legal archives of war crimes will be a primary source for lethal actions of machines that are designed for military needs. These verdicts should embed the international law and code-of-conduct regarding war actions, such as the Geneva treaty.

References

1. M. Anderson and S. L. Anderson, *Machine Ethics: Creating an Ethical Intelligent Agent*. AI Magazine Volume 28 Number 4 (2007) (© AAAI).
2. M. Anderson, S. Anderson and C. Armen, *An Approach to Computing Ethics, IEEE Intelligent Systems*. July/August, pp. 56-63, 2006. Arkin, R.C., Behavior-based Robotics, MIT Press, 1998.
3. R. C. Arkin, *Moving up the Food Chain: Motivation and Emotion in Behavior-based Robots*, in Who Needs Emotions: The Brain Meets the Robot, Eds. J. Fellous and M. Arbib, Oxford University Press, 2005.
4. R. C. Arkin, *Governing Lethal Behavior in Autonomous Systems*, Taylor and Francis, 2009.
5. R. C. Arkin, *The Case for Ethical Autonomy in Unmanned Systems*, Journal of Military Ethics, Vol. 9(4), pp. 332-341, 2010.
6. R. C. Arkin, M. Fujita, T. Takagi and R. Hasegawa, *An Ethological and Emotional Basis for Human-Robot Interaction*, Robotics and Autonomous Systems, 42 (3-4), March 2003.
7. R. C. Arkin and P. Ulam, *An Ethical Adaptor: Behavioral Modification Derived from Moral Emotions*, IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA-09), Daejeon, KR, Dec. 2009.
8. R. C. Arkin, A. Wagner and B. Duncan, *Responsibility and Lethality for Unmanned Systems: Ethical Pre-mission Responsibility Advisement*, Proc. 2009 IEEE Workshop on Roboethics, Kobe JP, May 2009.
9. D. K. Ashley, *Case-Based Reasoning and its Implications for Legal Expert Systems*, Artificial Intelligence and Law 1:113-208, 1992.
10. S. A. Bringsjord, *21st-Century Ethical Hierarchy for Robots and Persons: EH*, in A World with Robots: International Conference on Robot Ethics: ICRE 2015, volume 84, page 47. Springer, 2017.
11. N. S. Govindarajulu and S. Bringsjord, *On Automating the Doctrine of Double Effect*, Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17).
12. D. MacKenzie, R. C. Arkin and J. Cameron, *Multiagent Mission Specification and Execution*, Autonomous Robots, Vol. 4, No. 1, pp. 29-57, Jan. 1997.
13. A. Ophir, *The Order of Evils: Toward an Ontology of Morals*, MIT Press, 2005.

14. R. Rorty, *Contingency, Irony, and Solidarity*, Cambridge University Press, 1989.
15. M. Scheutz and B. F. Malle, *Moral Robots*, in K. Rommelfanger and S. Johnson (eds.), Routledge Handbook of Neuroethics, 2018. New York, NY: Routledge/Taylor and Francis.
16. C. Strong, 1988, *Justification in Ethics*, in Baruch A. Brody, editor, Moral Theory and Moral Judgments in Medical Ethics, 193-211. Dordrecht: Kluwer Academic Publishers.
17. M. Walzer, *Just and Unjust Wars*, 4th Ed., Basic Books, 1977.