ETHICAL ALGORITHMS IN HUMAN-ROBOT-INTERACTION. A PROPOSAL

Jörg H. Hardy

Free University of Berlin, Germany, Department of Philosophy,
AkakiTsereteli State University of Kutaissi, Georgia,
Faculty of Business, Law and Social Sciences

ABSTRACT

Autonomous robots will need to form relationships with humans that are built on reliability and (social) trust. The source of reliability and trust in human relationships is (human) ethical competence, which includes the capacities of practical reason and moral decision-making. As autonomous robots cannot act with the ethical competence of human agents, a kind of human-like ethical competence has to be implemented into autonomous robots (AI-systems of various kinds) by way of ethical algorithms. In this paper I suggest a model of the general logical form of (human) meta-ethical arguments that can be used as a pattern for the programming of ethical algorithms for autonomous robots.

KEYWORDS

AI Algorithms, Ethical Algorithms, Ethics of Artificial Intelligence, Human-Robot-Interaction

1. INTRODUCTION: ETHICAL COMPETENCE OF ROBOTS – A CHALLENGE FOR HUMAN-ROBOT-INTERACTION

Over the past two decades, robots with high levels of autonomy have become members of the human society. Autonomous robotics has made tremendous progress with significant advances in areas such as driverless cars, home assistive robots, robot-assisted surgery, and unmanned aerial vehicles. However, incidents such as the fatal Tesla crash show the risks from improper use of this technology. The progress in the development and deployment of robots (and any kind of autonomous AI-systems) sets experts the task to design algorithms that can generate reliable, trustworthy robots that possess a human-like ethical competence.

Autonomous robots will need to form relationships with humans. Human relationships are built on (social) trust, which is a key influence in decisions whether an autonomous agent, be it a human or a robot, should act or not. Studies of reliability and trust in automation have shown that trust depends on reliability and predictability: trust increases slowly if the system behaves as expected, but drops quickly if we experience failure [1]. Autonomous robots are independent decision-makers, and may therefore exhibit unpredictable behaviour. Reasoning with trust and norms is necessary to justify and explain a robots' decisions and to draw inferences about accountability for failures, and hence induce meaningful communication with autonomous robots [2].

Trust is a specific human stance, namely the expectation of beneficial behaviour. The stance of trust is a complex both cognitive, and volitional process, informed by the broader context of

David C. Wyld et al. (Eds): MLSC, ITCSS, ACSTY, SOFE, NATP, BDAB, AIBD - 2023
pp. 179-186, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.130214

morality, ethics, and social norms. Trust in human interaction is based on the ethical competence of human persons. Autonomous robots too would need a certain kind of a human-like ethical competence in order to be trusted [2, 3].

A reliable, trustful human-robot-interaction has to be guided by an ethics of algorithms. For the various aspects of the ethics of algorithms see Bostrom &Yudkowsky [4],Arnold &Scheutz [5], Malle, B. F. [6], Aggarwal & Mishra [7], Nida-Rümelin & Weidenfeld [8], and Tsamados & Aggarwal & Cowls & Morley & Roberts & Taddeo &Floridi [9].The realm of meta-ethical questions is described by Beauchamp & Childress 2013 [10] and Siep [11].In this paper, I confine myself to a proposal for a model of meta-ethical arguments and a corresponding logical pattern for ethical algorithms. My proposal is based on a minimalist theory of ethical competence that is not committed to a particular meta-ethical theory.

2. THE META-ETHICAL FRAMEWORK: THE MORAL STANCE

The source of (human) ethical competence is the moral stance [12]. The moral stance is a person's capacity and enduring motivation to *accept moral demands* for their own sake, regardless of any socio-economic reward for moral behaviour.

Human beings pursue happiness, and all our happiness-conducive rational and deliberate *social* (interpersonal) activity is intrinsically desirable. This assumption about the *condition humana*, which I take to be uncontroversial, has an important consequence for the understanding of morality. Morality prevails our entire social life. Throughout our whole life, we make moral demands on other people, and we are faced with their moral demands. If we accept those demands, we act in other person' s interests and thus put *constraints* on some particular self-interests. Having taken the moral stance, we permanently *want* to put constraints on *particular self-interests*. If moral actions are nevertheless *desirable*, they are desirable for their own sake, and if a persistent social behaviour is desirable for its own sake, then it is part of our pursuit of happiness, that is, part of an *overall* desirable life. We can maintain the moral stance for a lifetime only if morality is intrinsically desirable and part or our pursuit of happiness [13].

People have various motivating reasons for moral actions, such as, for example, the interest in successful social cooperation, the desire for social recognition, religious belief, altruism. When we have taken the moral stance, we take morality to be intrinsically desirable. A social action is intrinsically desirable if we take it to be desirable for its own sake, regardless of its consequences. By contrast, a social action is extrinsically desirable when being done for gaining a certain socio-economic success. It is true, we take many (and probably the most) social actions to be both, intrinsically and extrinsically desirable. However, if an action is intrinsically desirable, its desirability (and its value) does not *depend* on any external social success or reward.

The moral stance has three aspects:

(i) If we have taken the moral stance, we have certain *moral beliefs*, which we express in specifically *moral demands*. As moral agents, we accept moral demands and act in other person' s interests.

(ii) When we act in other person' s interests, we put constraints on some of our self-interests. Thus, moral agents are capable of having second order volitions. The moral stance includes the particular capacities of practical reason and moral decision making. Through practical reasoning we form intentions, which consist of a belief and a corresponding desire. Practical reasoning therefore is both a cognitive, and volitional capacity—for an intention without a belief would lack propositional content, and an intention without a desire would lack motivating force.

(iii) Moral experiences create a specifically moral familiarity between persons. Moral agents consider each other not merely as contracting parties who agree upon certain terms of contract, but they have certain attitudes towards each other, such as gratitude, respect, recognition, solidarity or moral resentment and even indignation. In other words: Morality is a mode of people's encountering with each other. As moral agents, we *share* the desire for the common experiences of respect, solidarity, sincerity, and trust.

2.1. Moral Demands

Moral demands have four features:

(i) Moral demands aim to protect *common goods*, such as bodily integrity and (personal and social) autonomy, and the (logically speaking) *particular instances* of a common good, above all human beings, with certain vulnerable properties. For this reason, arguments for ethical claims have to rely on general evaluative assumptions about common goods, which ideally every person can agree upon. Since the acceptance of a moral demand expresses the will of a person, the general evaluative assumptions of ethical arguments—which we might also call ethical principles—are *common agreements upon common goods*.

(ii) Moral demands are evident: What we owe to each other is obvious because we all know the common goods, which moral demands aim to protect. Morality is, as Immanuel Kant says, a matter of fact of reason (*Critique of Practical Reason*, 5:31). Thus, we do not need complex and fallible reasonings in order to understand the content of moral demands. However, we need complex reasonings in order to find solutions to particular ethical problems.

(iii) Morals demands are universal; they hold for any person and any action in any situation—regardless of any particular property of an individual person.

(iv) Moral demands are categorical (or unconditional, respectively); moral actions do not depend on any particular condition, and they are not primarily a mean for achieving a certain end, but they are rather an end in itself (intrinsically desirable). It is true; we very often do moral actions for their own sake (intrinsically desirable) as well as for the sake of social advantages (extrinsically desirable), because we seek social recognition and want to avoid blame and punishment. Having taken the moral stance, we, however, do moral actions for their own sake because they contribute to our pursuit of happiness.

Knowing these features of moral demands is part of our understanding of the human condition. Let me briefly explain: If we accept moral demands for their own sake, we then follow moral norms (or laws), which aim to protect common goods. All human beings are equal in seeking happiness; we all have the desire for conducting a good life. And we all share the same vulnerabilities, we all know that we all have the same vulnerable properties, such as the fact that we all can suffer from pain. Once we are aware of the fact that moral norms are made to protect the vulnerable properties of human beings, we know that moral norms are universal. Everyone can suffer from pain and no one wants to suffer from pain. If we accept an individual person's demand not to be hurt because we consider it a *moral* demand, then we accept *everyone's* demand not to be hurt. For example, if I am sure that not inflicting pain on a human being is morally right, I expect everyone else to think the same way. The very idea that there is a moral obligation only for me—or a particular group of people, respectively—to perform moral actions does not make sense. When we keep in mind that moral norms are made for protecting common goods, we also know that moral norms must be categorical (or unconditional, respectively). If we seriously respect the happiness-conducive interests of other persons, we want to do this under any possible conditions—even though we might sometimes fail to perform morally right actions

through negligence. It would not make sense to accept moral demands and to do moral actions merely as a means for achieving a certain particular end that we would not want to achieve under some other conditions. When we want to protect common goods, we consider moral demands universal and categorical (or unconditional, respectively). In this respect, taking the moral stance means to achieve what Lawrence Kohlberg ("Study of Moral Development", New York, Garland 1994) describes as the sixth and highest stage of moral development.

There is a possible objection against the idea of categorical moral demands: Consider a situation in which someone hurts an assassin in order to prevent him from attacking defenseless people. Actions of that kind are undoubtedly morally right. In some cases, in which a person is faced with a conflict of moral norms, she/he has to break a certain moral law and to impair a certain good in order to protect a higher good. The fact of moral conflicts shows that we need to agree upon a hierarchy of goods in order to solve those conflicts, but it does not conflict the assumption that moral demands are categorical.

2.2. Common Goods and the Awareness for Humanity

Human beings share various common goods, to which general evaluative premises of ethical arguments refer to, such as human dignity, bodily integrity, and (personal and social) autonomy. Particular instances of common goods have vulnerable and valuable properties that moral demands aim to protect. A certain property is vulnerable because it can be impaired or even destroyed, and we consider such a property valuable because we want to protect it.

The insight into the value of common goods is the motivating reason for moral obligation: We know that everyone can suffer from bodily pain and from losing the authority over her/his own life, and we do not want anyone to suffer from pain or to lose authority over her/his life. Let us call this insight the *awareness for humanity*. The awareness for humanity is both a certain kind of *understanding* and *empathy*. We all know what it means to be hurt or to lose authority over one's own life. These experiences are common ones—we did not have them without *sharing* them with other persons. Empathy provides the awareness for humanity with its volitional, motivating force. There is, however, no universal empathy, for only propositional attitudes can be generalized. The awareness for humanity therefore requires empathy and understanding, that is, the understanding of the *human condition*. If we have taken the moral stance, we know that we all share the same vulnerability and the desire for common goods like bodily integrity and (personal and social) autonomy, and we therefore want to take care of each other' s happiness-conducive interests. Sharing is caring, and caring is sharing.

The *recognition* of a common good is a *motivational belief*, which conjoins the insight that a certain vulnerable and valuable thing is in fact a good with the intention to protect such a good for its own sake—regardless of any other particular interest that one might also have for protecting a certain good. We just would not *have* the belief that something *is* a good without having the desire to *protect* it.

According to a widespread view, moral obligations (duties) and moral norms are objective obligations and norms since they should not depend on an individual person' s (contingent) wanting. This idea seems to be an implication of the assumption that moral demands are universal and categorical (in the sense explained above). But we have to be careful with the assumption of objective norms and obligations. It is true, having taken the moral stance, we want to protect the vulnerability of every person, and we always have this intention, not merely in a particular situation. We may consider moral obligations and norms *as if* they were objective, because we want to accept moral demands as being universal and categorical. Yet, moral intentions are subjective because of the simple reason that individual persons want to accept

moral obligations and to establish moral norms. There is no obligation and no social norm without the corresponding wanting of individual persons. What we ought to do, is what we *want* to do. Moral obligations and moral norms don' t fall like stars from heaven, but they are the result of the *common agreements* and *intentions* of *individual* persons.

We recognize common goods and accept universal, categorial moral demands through our *own* moral thinking and wanting, that is, through our personal autonomy. If we accept moral demands for their own sake, our *individual* (self-guiding) intention to do moral actions coincides with our acceptance of a quasi-objective moral obligation. Universal moral norms are those, which we (ideally) *all* can *agree* upon.

As robots cannot take the human moral stance, they cannot possess *human* ethical competence. However, autonomous robots are independent decision-makers. As robots make decisions and act by virtue of algorithms, technologists have to implement ethical algorithms into autonomous robots.

3. THE LOGICAL FORM OF META-ETHICAL ARGUMENTS AND ETHICAL ALGORITHMS

Meta-ethical arguments, that is, arguments with a normative, ethical conclusion have both evaluative, and descriptive premises that refer to common goods, particular instances of common goods, moral agents, intentions, and moral actions (or a certain kind of moral actions, respectively). When we *argue* for ethical claims, we agree on *general evaluative premises*, which express assumptions about common goods whose particular instances have certain vulnerable properties $\{V_1, ..., V_n\}$.

For example, the human body's vulnerable property is the fact that it can suffer from pain. A person's mind can be manipulated. A person's dignity can be humiliated. Those are the vulnerable properties ethical arguments typically refer to. More precisely: When we argue for ethical claims, we have to make
(i) general *evaluative* assumptions about common goods that we want to protect, which ideally all moral agents can agree upon,
(ii) general and particular *descriptive* assumptions about the vulnerable properties of a particular instance of a common good,
(iii) general and particular descriptive assumptions about the intentions and obligations of moral agents, and finally
(iv) general and particular descriptive assumptions about particular actions (or a set of actions, respectively), which is are necessary and adequate for protecting the vulnerable properties of a particular instance of a common good.

A particular moral action A is *necessary* for protecting a vulnerable property V of a particular instance of a common good if only action A will prevent a particular instance of a common good in a given particular situation from being impaired or even destroyed. A particular moral action A is *adequate* if and only if an agent is in the position to do A and doing A does *not* impair her/his own well-being.

Meta-ethical arguments have this general form:

(1) (∀common good CG, ∀particular instance ICG of a common good, ∀vulnerable property V, ∀moral agent MA, ∀an agent's intention I to protect V): If an abstract entity CG is a common good and if (logically speaking) a particular instance ICG of a common good, for example, an

individual person, has the vulnerable property V, then every moral agent MA has the intention Ito protect the vulnerable property V of any particular instance ICG, for example the vulnerable property V of any other person. (The *antecedens* of this assumption contains an evaluative as well as a descriptive statement.)

(2) (∀common good CG, ∀particular instance ICG of a common good,∀vulnerable property V): The entity CG is a common good and every particular instance ICG of CG (for example, every individual person) in fact has the vulnerable property V.

(3) (∀moral agent MA, ∀an agent' s intention I to protect V): Therefore, every moral agent MA has the intention Ito protect everyone's vulnerable property V.

(4) (∀moral agent MA, ∀an agent' s intention I to protect V, ∀an agent' s obligation O): If an agent MA has the intention I to do actions of the kind A, which are necessary and adequate for the protection of a vulnerable property V of ICG, then she/he *ought*—has the (self-guiding) obligation O—to perform particular actions of the kind A (and must not do opposing actions of the kind non-A).

(5) (∀moral agent MA): An agent MA has the intention I to do actions of the kind A, which are necessary and adequate for the protection of a vulnerable property V of ICG.
Conclusion: Therefore, every moral agent ought to do—has the obligation to do—actions of the kind A.

Meta-ethical arguments of this kind have the following elementary logical form:
Let ICG be a *particular instance* of a *common good*, V a *vulnerable property* of an ICG, MA a *moral agent*, I an agent' s *intention* to do an *action* A, which is necessary and adequate for the protection of a vulnerable property V of ICG, and to avoid any action that would *impair* an ICG, respectively, and O an agent' s (self-guiding) *obligation* to do a certain action A, which is necessary and adequate for the protection of a vulnerable property V of an ICG.

(1) ∀(x, y) [ICG(x) & V(x)] ⊃ [MA(y) &I(y)]
(2) ∀(x) ICG(x) & V(x)
(3) ∀(y) [MA(y) &I(y)] ⊃ [MA(y) & O(y)]
(4) ∀(y) MA(y) &I(y)
Conclusion: ∀(y)MA(y) & O(y)

Here is an example:

If bodily integrity is a common good and if every individual person—as being (logically speaking) a particular instance of the common good bodily integrity—has the vulnerable property that she/he can be hurt, then every moral agent has the intention to protect everyone from being hurt.
Bodily integrity is a common good and every individual person can suffer from being hurt.
Therefore, every person has the intention to protect everyone from being hurt.
If we (and every moral agent) have the intention to protect everyone from being hurt, then we ought to act the way that we don't hurt someone.
Therefore, we (and every moral agent) ought to act the way that we don't hurt someone.

Notice that arguments of this kind are not vulnerable to the objection of the so-called naturalistic fallacy since the premises entail the entire evaluative information of the conclusion.

Ethical algorithms based on this logical form of meta-ethical arguments will include

- logical and semantic information about common goods and particular instances of a common good, such as human beings, with certain vulnerable and valuable properties such as bodily integrity and personal autonomy
- logical and semantic information about the intentions of moral agents
- logical and semantic information about sets of actions that are necessary and adequate for protecting the vulnerable properties of a particular instance of a common good.

In order to implement ethical competence into autonomous robots, we have to combine ethical algorithms of the above kind with *technical tools* that enable an autonomous system to *identify* a morally right (necessary and adequate) *particular action under particular circumstances*. Very useful methods for this complex task are model checking and rational verification.

4. MODEL CHECKING AND ETHICAL ALGORITHMS

Scholars of computer science have identified a logic language and method to evaluate reliability and trust in human-robot interactions through *model checking*, which is capable of reasoning about concepts such as epistemic dependence between agents [14]. This formalism can be extended by ethical and normative reasoning and, in particular, ethical algorithms for tool support [2, 4, 15]. Most relevant to this aim is the paradigm of *rational verification*, which is based on model checking, that enables analyzing human-robot behaviours under the assumption of agents behaving rationally, and allowing for incentives and preferences [16].

5. CONCLUSION

The study of the ethical aspects of human-robot-interaction is an interface between computer science and philosophical ethics. The source of human ethical competence is the moral stance, that is, a person's capacity and enduring motivation to accept moral demands for their own sake. As robots cannot act with human ethical competence, reliable and trustful human-robot-interaction requires the implementation of the capability of ethical decision-making into autonomous robots by way of ethical algorithms that are based on the general logical form of meta-ethical arguments. Meta-ethical arguments, that is, arguments with a normative, ethical conclusion have both evaluative, and descriptive premises that refer to common goods, particular instances of a common good—above all human beings—with certain vulnerable properties, intentions of moral agents, and moral actions, which aim at protecting particular instances of a common good. In this paper, I have proposed a model of the generallogical form of meta-ethical arguments and ethical algorithms. Computer scientists and technologists can use this model and specific applications of this model in order to implement ethical competence and ethical decision-making factors into algorithms and software tools with support for autonomous AI-systems.

ACKNOWLEDGEMENT

The model of ethical algorithms that I propose in this paper is a revised version of a model that I have presented at the FLoC 2018 – workshop "Robots, Morality, and Trust through the Verification lens" in Oxford, July 2018, http://qav.cs.ox.ac.uk/robots_morality_trust/. I am most grateful to Morteza Lahijanian, Lu Feng and Nils Jansen for critical comments that helped me improving my model.

REFERENCES

[1] Huang, X. &Kwiatkowska, M. (2017) "Reasoning about Cognitive Trust in Stochastic Multiagent Systems", *Proceedings of the AAAI Conference on Artificial Intelligence* (AAAI 2017), AAAI Press.

[2] Lahijanian, M. &Kwiatkowska, M. (2016) "Social Trust: A Major Challenge for the Future of Autonomous Systems", *AAAI Fall Symposium on Cross-Disciplinary Challenges for Autonomous Systems*, AAAI Press.

[3] Kuiper, B.(2016)"Human-like Morality and Ethics for Robots", *Proceedings of the Association for the Advancement of Artificial Intelligence*, Ann Arbor, Michigan.

[4] Bostrom, N. &Yudkowsky, E. (2014)"The Ethics of Artificial Intelligence", in: K. Frankish & W. M. Ramsey (eds.), *The Cambridge Handbook of Artificial Intelligence*, Cambridge, pp 316-334.

[5] Arnold, T. &Scheutz, M. (2016)"Against the moral Turing test: accountable design and the moral reasoning of autonomous systems", *Ethics and Information Technology*, Vol. 18, 2, pp 103-115, Springer.

[6] Malle, B. F. (2016) "Integrating robot ethics and machine morality: the study and design of moral competence in robots", *Ethics and Information Technology*, Vol. 18, pp. 243–256, Springer.

[7] Aggarwal, S. &Mishra, S. (2021) *Responsible AI: Implementing Ethical and Unbiased Algorithms*, Springer.

[8] Nida-Rümelin, J. & Weidenfeld, N. (2018)*Digitaler Humanismus: Eine Ethik für das Zeitalter der Künstlichen Intelligenz*, München.

[9] Tsamados, A. & Aggarwal, N. &Cowls, J. & Morley, J. &Roberts, H. & Taddeo, M. &Floridi, L. (2022)"The ethics of algorithms: key problems and solutions", *AI & Society*, Vol. 37, pp 215–230. https://doi.org/10.1007/s00146-021-01154-8

[10] Beauchamp, Tom L. & Childress, J. F. (2013) *Principles of Biomedical Ethics*, Oxford, Oxford University Press.

[11] Siep, L. (2004) *Konkrete Ethik. Grundlagen der Natur- und Kulturethik*, Frankfurt am Main, Suhrkamp.

[12] Hardy, J. (2017) "Understanding Ethical Reasoning", Hoesch, M. & / Laukötter, S. (eds.). *Natur und Erfahrung. Bausteine zu einer praktischen Philosophie der Gegenwart*. Festschrift für Ludwig Siep, Münster, mentis.

[13] Hardy, J. (2011) *Jenseits der Täuschungen – Selbsterkenntnis und Selbstbestimmung mit Sokrates*, Göttingen, V & R unipress (ch.XIV).

[14] Kwiatkowska, M. (2007) "Quantitative verification: models, techniques and tools", *Proceedings of ESEC/SIGSOFT FSE* 2007, pp 449-458, IEEE CS Press.

[15] Alechina, N. & Halpern, J. Y. &Kash, I. A. & Logan, B. (2017) "Incentivising Monitoring in Open Normative Systems", *Proceedings of the AAAI Conference on Artificial Intelligence* (AAAI 2017), AAAI Press.

[16] Wooldridge, M. & Gutierrez, J. &Harrenstein, P. &Marchioni, E. &Perelli, G. &Toumi, A. (2016) "Rational verification: from model checking to equilibrium checking", *Proceedings of the Thirteenth AAAI Conference on Artificial Intelligence*, AAAI Press, pp 4184-4190.

AUTHOR

Dr. Jörg H. Hardy is Senior Lecturer (PD) at the Free University of Berlin, Germany, Department of Philosophy, and Professor of Philosophy and Linguistics at the Akaki Tsereteli State University of Kutaissi, Georgia, Faculty of Business, Law and Social Sciences.