International Journal of Social Robotics (2021) 13:557–567
https://doi.org/10.1007/s12369-019-00540-z

Ethics, Human Rights, the Intelligent Robot, and its Subsystem for Moral Beliefs

Erik Sandewall[1]

Accepted: 1 March 2019 / Published online: 11 March 2019

Abstract
The Universal Declaration of Human Rights specifies a number of properties that characterize human beings, such as 'dignity', 'conscience', and several others. In this article we focus on these properties and on how they have been defined in the history of philosophy. We show how they can be interpreted in terms of a prototypical architecture for an intelligent robot, and how the robot can be provided with several aspects of ethical capability in this way. The key idea is to provide the robot with a Moral Belief System that cooperates with, and moderates the robot's capability of planning and action.

Keywords Robot ethics · Moral belief state · Giovanni Pico della Mirandola · Immanuel Kant · Universal Declaration of Human Rights

1 Introduction

The present investigation started with the question whether the Universal Declaration of Human Rights (UDHR) [1] could be applied to intelligent robots. If this is feasible, then it may contribute both to the discussion of how governments should relate to such robots, and to the definition of behavior rules for them. The UDHR states that the freedoms that it claims for all human beings must not be used by them as a license to violate the rights of others. This clearly translates into restrictions on the appropriate behavior of agents (humans or robots) that may be covered by the UDHR. In addition, since the UDHR makes a number of statements about properties that humans have intrinsically, we felt that it would be interesting to relate those statements to design principles for intelligent robots.

This curiosity-driven approach turned out to be quite rewarding, in particular when applied to the concept of intrinsic properties of humans. The UDHR uses a few such key concepts, in particular 'dignity', 'reason', and 'conscience', but also several others. These are classical concepts in philosophy which have been extensively studied there. The task of understanding that whole literature, with the various opinions on its topic, was not within reach. We decided therefore to identify the original authors whose work has set the direction for many of the subsequent contributions, and to study their definitions for the key concepts in some detail.

This choice led us to consider two authors in particular, namely Giovanni Pico della Mirandola who wrote the foundational article 'Oration on the Dignity of Man' [2] in 1486, and Immanuel Kant whose works include 'Critique of Pure Reason' [3] and 'Critique of Practical Reason' [4] . Both of these are very relevant for understanding the key concepts in the UDHR.

But it turned out that besides shedding light on the intended meanings of the statements in the UDHR, the concepts that were proposed by these authors could as well be related directly to the design principles for intelligent robots. At the same time, we also consulted the documents of the United Nations committee that authored the UDHR [5], in order to ascertain that their assumptions and views were consistent with the ones that we obtained from the writings of Pico and Kant.

With all due respect for the authors of the UDHR, however, for the present article we feel that it is appropriate to base it on the definitions of Pico and Kant, as well as other original authors. Our article is therefore organized as follows. We shall first define our assumptions about the design of intelligent robots, in terms of an idealized architecture that specifies what components must be present in it, a few optional components, and the relationships between the components. Next, we identify what we consider to be the essential concepts

✉ Erik Sandewall
erik.sandewall@liu.se

[1] Linköping University, Linköping, Sweden

in the UDHR for our purpose, and explain why some other concepts have not been included. After this, we address one concept at a time and discuss its interpretation by our classical authors, or by others. In these considerations we are also led to introduce a few additional concepts (besides those mentioned in the UDHR) that are important for seeing the full picture.

In the second half of the article, we propose an additional component for an intelligent robot, which we call its *Moral Belief System.* This component consists of a collection of moral beliefs (the Moral Belief State), together with software that is capable of applying those beliefs in the operation of the intelligent robot as a whole. The purpose of the Moral Belief System shall be to modify the robot's behavior so that it conforms as well as possible to 'moral rules' in a sense that will be further discussed below. The acronym MBS will be used for both the system in question, and for the belief state that it contains.

In summary, there were two reasons for the idea to identify (proposed) intrinsic properties of humans, and to see whether and how they can be applied to intelligent robots as well. The first reason on our part was mere curiosity, but as the work proceeded, we began to see interesting implications for the design of an ethical capability in those robots. One may speculate whether the resulting, operational definition of ethical competence may also be of interest for scholars in the field of ethics, but we must leave it to them to have an opinion on this.

2 Relation to Existing Work on Machine Ethics

The present article combines a grounding in classical philosophical concepts of ethics with an explicit model of the intelligent agent. This represents a step forward relative to earlier publications. The first generation of thoughts about the ethics of robots originated with Isaac Asimov's 'Three Laws of Robotics' [6], and was dominated by reasoning about various proposed laws, in particular, their necessity or plausibility, and their consequences.

Contemporary work on robot ethics started with Michael Anderson's and Susan Leigh Anderson's seminal article from 2007: 'Machine Ethics: Creating an Ethical Intelligent Agent' [7]. In this article they discuss the importance of machine ethics, the need for machines that represent ethical principles explicitly, and the challenges facing those working on machine ethics. They also give a simple example of how a machine may abstract an operational ethical principle from examples of correct ethical judgments. Subsequent work has mostly continued the discussion of these topics, or demonstrated simple examples of ethical reasoning in a machine.

With respect to theory, deontic logic is an obvious candidate for a theoretical foundation for machine ethics. However, it can only be used if it is shown how to integrate it into an architecture for the intelligent robot as a whole. Our approach in the present article is to begin with that problem.

In the discussion about machine ethics, a few authors such as Patrick Chisan Hew have expressed doubts about its usefulness, arguing that "Such systems are a substantial departure from current technologies and theory, and are a low prospect" [8]. We feel on the contrary that substantial departures from current technologies have occurred repeatedly in information technology, and we see no reason why one could not happen in the present case. We propose also that the design considerations in the present article may be a step in this direction.

One example of a 'substantial departure', of a very different kind than the one suggested in this article, was proposed in a book edited by Stephen Palmquist, "Cultivating Personhood: Kant and Asian Philosophy" [9]. This book contains an extended argument about whether and how 'reason' and 'free will' can arise from the biological processes of conception and fetus development. This is an example of the broad range of contributions to the topic of machine ethics.

3 Intelligent Robots

For the purposes of the present article, an intelligent robot is a mechanical device that at least is equipped with sensors, actuators, a sensory-motoric system and a planning and action system (PAS). The sensory-motoric system contains driver software for the sensors and actuators, and a perception capability whereby a concise description of the robot's current environment is obtained from the sensor inputs, and is also updated continously. This description will be referred to as the robot's *world model.* In simple cases, the world model will just represent a few objects in the robot's field of observation, with some features of these objects and some relations between them. In more advanced systems, the world model may also represent remote objects, complex objects, present and past events and actions, and other constructs as studied in the field of Representation of Knowledge.

The Planning and Action System is responsible for the robot's choice of actions, the execution of those actions, and various kinds of reasoning about actions. Several parts of this PAS will be relevant for our discussion of robot ethics, and we shall therefore state our assumptions about it with some detail. Our assumptions are consistent with system architectures that are used in practical systems; see e.g. [10].

The PAS shall include a set of designators for *elementary actions* and likewise a set of designators for *composite actions.* Each elementary action designator is associated with a computational mechanism for performing the action. This

mechanism may be defined as a procedure in a programming language, but a number of other design paradigms are also possible, for example in terms of finite automata.

Each composite action designator may be associated with one or more plans that can be used for performing the action. In simple cases, a plan is a sequence of (lower level) action designators. More complex plans can be formed using conditional and repetition operators, as usual.

Furthermore, each action designator shall be associated with information about how the execution of the action is expected to affect the state of the robot's environment, in terms of the categories that are used by the sensory capability. Based on this, the PAS shall have a *prediction capability* whereby it can compute an expectation of the result state after performing an elementary or composite action.

Finally, the PAS shall contain an *action suggesting mechanism* which is able to generate or select one or more action designators based on the current world model as generated by the perception capability. It may be implemented, for example, as a set of situation-action rules.

These are the minimal requirements, and additional facilities may be considered. They include a *goal suggesting mechanism,* as well as a *planning mechanism* for constructing a composite action that is expected to lead to a situation where the robot's environment satisfies certain conditions that are referred to as 'the goal'. An additional and useful facility will be an *action evaluation mechanism* that can be applied to the outputs of the action suggesting mechanism, in order to assess the cost, the likelyhood of success, and other characteristics of a suggested action.

4 Reasoning and Intelligence

With few exceptions, human beings have a prediction capability and an action suggesting mechanism, similar to those that were described above for intelligent robots. Several alternatives come to mind with respect to how these capabilities work. One possibility is that they work by logical reasoning which is performed in a step-by-step fashion, and where the conclusions in each step can be communicated to others and discussed with them.

Another possibility is that they "just happen" in the sense that the person in question is not able to explain how she or he arrived at the prediction or the suggestion at hand. This is essentially the distinction between 'conscious' and 'subconscious' cognitive processes. We shall refer to these alternatives as "reasoning" and "intuition", respectively. They are not mutually exclusive, since it seems that people can use both depending on the situation at hand, and also since a result that was arrived at by "intuition" can later be explained and motivated as if it had been obtained by reasoning.

These human characteristics have immediate counterparts in the design of robots, where "reasoning" can be implemented using logic-based techniques, and where "intuition" can be implemented using neural networks, for example.

With these definitions and premises, we can proceed to the question of whether, and under what conditions may it be possible to apply the Universal Declaration of Human Rights (UDHR) to intelligent robots.

5 Key Concepts in the UDHR

Four concepts in the UDHR are of central importance if it shall be applied to intelligent robots, namely 'freedom' (or 'free'), 'dignity', 'reason' and 'conscience'. They appear as follows in its Article 1 of the UDHR:

> All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.

Some of these words also appear as follows in its Preamble:

> Whereas recognition of the inherent dignity and of the equal and inalienable rights of all members of the human family is the foundation of freedom, justice and peace in the world…

The interpretation of the phrase 'in a spirit of brotherhood' will be discussed in the section about the concept of 'conscience', later on in this article. Therefore, at this point we only need to consider how the terms 'free', 'dignity', 'reason' and 'conscience' can be applied to intelligent robots. These concepts are treated differently in the other 29 articles. The concept of 'conscience' appears only once, in Article 18 which begins:

> Everyone has the right to freedom of thought, conscience and religion…

The concept of 'dignity' appears only in Article 23:

> Everyone who works has the right to just and favourable remuneration ensuring for himself and his family an existence worthy of human dignity…

and the concept of 'reason' does not appear anywhere else besides in Article 1. By contrast, the concept of 'freedom' appears repeatedly, in particular since several articles define what specific 'freedoms' shall be enjoyed by everyone. In fact, the words 'free' or 'freedom' appear 21 times in these 29 articles.

Some other frequently occurring concepts are 'rights' and 'entitled to'. For example, Article 28 says:

> Everyone is entitled to a social and international order in which the rights and freedoms set forth in this Declaration can be fully realized.

I shall not address these concepts here, since they are mostly used for 'rights' that must be guaranteed by national governments, which means that they must be considered in the broader context of how governments relate to intelligent robots. This would take us outside the scope of the present article.

It remains, therefore, to consider the concepts of 'freedom', 'conscience', 'reason' and 'dignity' as used in the UDHR, and for each of them to discuss its requirements on the design of an intelligent robot. With respect to conscience, for example, we need to discuss what may be needed for the robot to have a conscience. Ideally we would also like to understand how the robot may recognize and respect the conscience of others, but this is outside the scope of the present article. I shall address these four concepts one at a time, together with a few others that will come up along the way.

6 The Concept of Freedom

The UDHR specifies a number of freedoms for everyone, but it is clearly not an exhaustive list of all the freedoms that people have. Therefore, at first sight, it would seem that the 'freedom' aspects of the UDHR should not constrain the permissible behaviors of the intelligent robot at all. However, the Declaration also specifies certain restrictions on the freedoms, namely, in the clauses of Article 29:

> (1) Everyone has duties to the community in which alone the free and full development of his personality is possible.
> (2) In the exercise of his rights and freedoms, everyone shall be subject only to such limitations as are determined by law solely for the purpose of securing due recognition and respect for the rights and freedoms of others and of meeting the just requirements of morality, public order and the general welfare in a democratic society.
> (3) These rights and freedoms may in no case be exercised contrary to the purposes and principles of the United Nations.

The first clause gives a lot of room for interpretation: what is included in the "duties to the community", and how can a decision or a consensus be reached on such issues? The second clause states that freedoms can only be restricted by law, but in actual practice one would prefer for intelligent robots to recognize informal requirements of morality and public order, and not merely the legal prohibitions in force. In this case the room for interpretation increases even further.

Therefore, in spite of the emphasis on 'freedom' in the UDHR, it also allows for restrictions on those freedoms, and the extent of these restrictions is a matter for the society where the UDHR is applied. This is also related to the question of 'conscience' which will be addressed later on in this article.

Somewhat surprisingly, therefore, the implications of the UDHR for the design of intelligent robots impose important *restrictions* on the robot's freedom of action.

7 The Concept of Dignity

The 'dignity' of all human beings is emphasized very strongly in the Preamble and in Article 1, but UDHR does not even begin to define what it means. It is a well-established term in moral philosophy, however, and a brief review of its origins is therefore in place here. Its Latin form, 'dignitatis', is derived from the adjective 'dignum'; a person is 'dignum' of doing something if he or she deserves to do it and is competent of doing it. This noun was promoted by Giovanni Pico della Mirandola in [2] where he proposes that humans have a particular 'dignity' because they have free will, whereas both physical objects and animals can only react to whatever forces operate on them. Because of their capacity for free will, humans are able to change themselves, thereby ascending in a chain that leads from the physical world at the bottom, to the divine world at the top.

The concept of dignity was also adopted by Immanuel Kant in order to make a distinction between things that can be discussed merely in terms of what value they have for someone, and those that can not or must not be considered only in those terms. The requirement on the latter is that they shall have a moral dimension, which meant for Kant that they are 'ends in themselves'. He wrote: "Morality, and humanity insofar as it is capable of morality, is that which alone has dignity" [4,11]. Like Pico, he also considered free will to be an essential aspect of the dignity of the human person.

The combination of free will and morality is therefore of interest for our topic of study. Two later sections will address these two topics in succession. The next section after them will introduce the concept of a Moral Belief System, which we propose as a facility in a robot whereby it can use its morality to voluntarily constrain its own free will. In this way we address the question of how morality and free will can be combined operationally, which is such a crucial issue for our two classical authors.

8 The Concepts of Reason and of Being Endowed with Reason

The UDHR states that all human beings are "endowed with reason". This phrase is not in common use, but probably most readers have an informal understanding of what it may

mean. A Google search for this phrase has only returned references to the UDHR, to a few texts of religious origin, and to works of Kant or about him. In particular, in [11], Kant uses this phrase synonomously with "being a rational being", which agrees well with common-sense interpretations of these phrases.

At the same time, the concept of 'reason' itself is of paramount importance in Kant's philosophy, as one can see already by its appearance in the titles of several of his major works [11,12]. Kant makes an interesting distinction between 'logical' and 'pure' uses of reason. The former involves the use of logic for drawing conclusions, which Kant considers as a 'subordinate' faculty. In the pure use, on the other hand, "reason itself contains the origin of certain concepts and principles" [11]. Logical reason can therefore be implemented by reasoning software in a computer, whereas pure reason is bound to be a design principle for the software and for the representation of knowledge in the robot.

However, Kant's works are written in German and use the term 'Vernunft' which is then translated into 'reason' in English. This translation is treacherous since 'Vernunft' means 'reason' in the sense of being sensible whereas it does not have any significant connection to 'reasoning'.

For the present purpose, we shall combine Kant's two varieties of 'reason' and say that a computational agent "has reason" if its software and knowledge representation are organized in a rational way and if they are consistent with pure reason in the sense of Kant. In practice, it is likely that logical reasoning will be extensively used in in its Planning and Action System, but this will not be essential for it 'being endowed with reason'. The first part of this condition means for example that when the PAS initiates the execution of an action, it only chooses an action that has been obtained from the action suggesting mechanism. It also means that the planning mechanism will only produce plans that are expected to achieve the given goal.

9 The Concept of Free Will

The question whether robots can be said to have free will has been the topic of much debate in recent years. However, already in the year 2000, John McCarthy published an article with the title "Free Will—Even for Robots" [12] that effectively settled the issue. McCarthy adopted the philosophical stance of compatibilism [13], a philosophical line of thought that goes back to the classical stoics. In this view, free will and determinism are mutually compatible, so it is possible to embrace both without being logically inconsistent. An instance of 'free will' may be seen as one in which the agent was able to choose between alternative goals or actions according to its own internal, cognitive processes, even though these processes may appear entirely deterministic for an outside observer. Kant expressed this view by saying that a rational will cannot act except "under the idea" of its own freedom [11].

In his article, McCarthy also proposed to view 'free will' as a graded concept, so that the 'will' of an agent may be considered as more or less free, depending on whether and to what extent the agent has internalized external constraints on its behavior. For example, if a parent has forbidden a child to perform a particular action, then the child will be said to have less free will if it has adopted the restriction as its own, and more free will if the child is mentally prepared to perform the action in spite of the parent's instructions. The latter case applies even if the child does not actually perform the action, for example for some other reason that arises indepently of the parent.

The concept of dignity according to Pico della Mirandola and according to Kant is strongly tied both to the existence of free will, and to the presence of morality that influences the choices of this will. This is a strong philosophical reason for paying attention to the combination of those two concepts when providing intelligent robots with a sense of ethics. It follows that such an enterprise can only make sense under the compatibilistic view of free will, since computers must be viewed as deterministic devices. (The introduction of a randomization device would not change this matter). The compatibilistic view is therefore the natural choice when discussing the dignity of robots.

10 The Concept of Morality

The article about 'Definition of Morality' in the Stanford Encyclopedia of Philosophy [14] distinguishes between a descriptive use and a normative use of this word. The normative case is explained as *a code of conduct that, given specified conditions, would be put forward by all rational persons.* This is well in line with the writings of classical authors, such as Kant.

The descriptive use of this term refers to *certain codes of conduct put forward by a society or a group (such as a religion).* It seems likely that this meaning is intended in Article 29 of the UDHR when it states that freedoms shall be restricted by *the just requirements of morality, public order and the general welfare in a democratic society.* With respect to intelligent robots, therefore, they should implement normative morality in order that they can be considered to have dignity, and they should also have knowledge of morality in the descriptive sense and for the environment that they are in, so that in their actions, they will not violate Article 29.

11 The Moral Belief System

We shall now describe an additional, proposed subsystem of the intelligent robot, called the 'Moral Belief System' (MBS) containing its verdicts about which actions and situations are 'right' and which are 'wrong'. The primary purpose of the MBS is to allow the robot to decide for itself what are the moral constraints on its own choice of actions, in line with the compatibilistic view that was described above.

In the context of the UDHR, the MBS is important for interpreting Article 29 so that the robot will not perform actions that it should not. The MBS is also needed, in line with Pico's argument, so that the robot shall be competent to exercise its own free will and so that it shall deserve doing so. And finally, the Moral Belief System should be seen as the carrier of the morality that Kant required as a condition for dignity.

I shall first outline the technical characteristics of the Moral Belief System (MBS), and later on discuss whether and to what extent it can provide the functionality of a conscience in the sense of the UDHR. The MBS is designed to cooperate with the robot's Planning and Action System (PAS) that was described above, allowing it to recognize when a proposed action may produce 'bad' effects, or 'good' effects. To this end it must minimally contain value statements that label particular conditions in the robot's world as being 'bad' or 'good', maybe with further qualifications or a graded scale. The same range of values may also be assigned to actions themselves. What is now said about actions applies also to goals and plans. Additional features of the MBS will be introduced below.

If an intelligent robot is designed so that in each situation, it addresses the output of the action suggesting mechanism and selects one of the suggested actions, then the MBS as now described may be used as a filter on the suggested actions, allowing only some of them.

On the other hand, in a robot that uses a goal suggesting mechanism, and planning for finding a plan that achieves the selected goal, the combination of the PAS and the MBS will work as follows. When a goal has been selected, the robot shall first verify that this goal is compatible with the Moral Belief State. If it is, the robot must consider possible plans for achieving this effect, and it will use the planning mechanism in order to obtain one or a few proposals for possible plans. According to the normal operation of a PAS, it will consider these proposals from the point of feasibility, cost, and other factors that may be relevant. With an MBS, it will also consider the possible additional effects ('side-effects') of each action in a plan, and relate them to the statements in the Moral Belief State. If it is determined that a plan can have side-effects that are unacceptable according to the MBS then that plan must be disqualified or amended.

12 Autonomous Modification of the Moral Belief System

The moral stance of a person or a group is not fixed; it can change over time when new facts become known and when circumstances change. For example, if a group has traditionally maintained that homosexuality is an evil thing, and later it realizes that this view has distressing consequences for the people involved, the members of the group may reconsider the reasons for their traditional view, and then change "their MBS" in this particular respect.

This example shows that when a robot is equipped with a Moral Belief System, one may consider an advanced facility where the robot is able to observe actions that are performed by other agents in its environment, and to assess the consequences of those actions according to the value statements in its MBS. In this way the robot will obtain a larger corpus of events as a basis for its deliberations. Some of these events may involve violations of values in the observer's MBS and cause it to revise its value structure.

The robot's capability for reviewing and revising its Moral Belief State will be further improved if it is engaged in a continous dialog with other agents (i.e., other robots, or persons) concerning the events that they observe together. Just like people tend to exchange views about the goodness or badness of specific events, and the arguments for their respective positions, a robot that engages in a similar dialog may be led to revise its Moral Belief State.

Moreover, the example above illustrates the interdependence between actions and value statements: actions have consequences that may be assessed according to the values, but the adoption of values by a person or a group will also have consequences since it affects what actions are taken and what actions are not taken. This means that there is a consistency requirement on the MBS: it should not contain value statements whose presence there has consequences that are considered as bad according to that same MBS.

Accordingly, it is not sufficient to check the Moral Belief State for consistency a single time, since new facts arrive (by observations, or by dialogue with other agents), and they may introduce inconsistencies in a previously consistent MBS. Moreover, the MBS shall not be analyzed merely as a collection of value statements; it must be seen in connection with the facilities of the entire Moral Belief System, i.e. facilities for the assessment of observed facts, for the detection of inconsistencies in the MBS, and for changing it so as to repair such inconsistencies. The software 'engine' that is needed for this purpose will resemble a reason-maintenance system [15] which is a classical device in A.I.

One may ask whether it would not be better to design the MBS in such a way that it does not contain any inconsistencies, and so that inconsistencies can not possibly occur as the result of new information, or as the result of autonomous

changes in the world model or the MBS. This will be a very difficult task, however, and the alternative is to design it using current techniques for reasoning in the presence of inconsistencies [16]. This possibility is also supported by the observation that humans seem to manage very well in spite of inconsistencies both big and small.

13 General Rules for the Moral Belief State

Besides the case-driven revision of the MBS, there are also some well-known moral principles that can be seen as restrictions that must be satisfied by the MBS as a whole. The 'golden rule' of Jesus Nazaraeus is an example of such a global restriction:

> So whatever you wish that others would do to you, do also to them, for this is the Law and the Prophets (Matthew 7:12).

Kant's principle of universalizability can also be seen as such a restriction, and as a generalization of the golden rule. One of its formulations is as follows [11]:

> Act only according to that maxim whereby you can at the same time will that it should become a universal law.

In our terms, a 'contradiction' is a situation where the application of the propositions and mechanisms in the robot's MBS can lead to results that are 'bad' according to that same MBS. However, this formulation of the principle is not very practical, since it would require considerable deliberation each time the robot considers performing a particular action. Therefore it would make more sense to use it as a constraint on the combination of the Planning and Action System and the Moral Belief System in each particular robot, as follows: the MBS must only contain value statements that could be adopted by all humans or robots (or by almost all of them) without obtaining any significant 'bad' results. Moreover, the same must apply for the situation-action rules in the PAS action suggesting mechanism.

14 Autonomous Modification of the World Model

The world model was mentioned initially as one of the essential components of the intelligent robot. Except in very simple cases, the world model may represent physical objects that are of interest for the robot, properties of objects and relationships between them, current actions and events, past events and foreseen future events, and so forth. This world model will be continously updated according to the robot's perceptions, and by communication with other agents.

Like in humans, the world model is interdependent with the low-level perception capability, so that perception updates the world model and the world model affects the perception. However, several other facilities in the intelligent robot will also need to use the world model, including in particular the Moral Belief System. For a simple example, if the MBS shall express that a particular action is required, appropriate, or excluded in a particular type of situation, then both the action and the situation description should be expressed in the terms used by the world model. Also, the world model must be used for reasoning about such actions and situations. The same applies for the action suggestion mechanism and the action evaluation mechanism.

Because of these interdependencies, one should also keep in mind the possibility that the Moral Belief System may influence the world model. This may happen, in particular, as the result of a cognitive dissonance, i.e. a mismatch between different desires or inclinations in the robot. When the robot's cognitive system attempts to remove the reasons for a cognitive dissonance, one possible diagnosis on its part may be that the dissonance is the result of a fault in the world model, which may lead it to reconsider some aspect of that model. This kind of "new understanding" does occur sometimes in humans. Although it is not easily replicated in artificial systems, one should at least attempt to design those systems in such a way that this kind of rethinking is not excluded from the outset.

15 The Concept of Conscience

The UDHR states that all human beings are endowed with conscience. It therefore makes sense to ask whether a similar facility would also be needed and useful in an intelligent robot, and whether it could contribute positively to the robot's behavior. Moreover, as we shall see, the concept of conscience has some interesting implications for the design of the intelligent robot.

The Stanford Encyclopedia of Philosophy describes one meaning of this word as follows:

> When we talk about conscience, we often refer to reflection about ourselves as moral persons and about our moral conduct. Through conscience we examine ourselves, as if we were our own inner judge. The image of an individual split into two persons, one who acts and the other who observes the former's conduct, reflects the original conception of 'conscience' in the Greek world, at least from the fifth century BCE.

Let us consider this definition first, and then return to other definitions. The one at hand allows for two operational interpretations, depending on whether the "judgement" of an action occurs before or after it has been performed. In

the first case, the conscience operates as an enabler or a disabler of actions, so that a person's conscience may rule that he or she shall absolutely not perform certain actions, or that in certain situations he or she shall perform certain actions.

This aspect of the conscience is clearly accomodated by the combination of the PAS and the MBS that was described above. It imposes a requirement on the system design, namely, that the morally required actions are included in the output of the action suggesting mechanism.

A problem arises if a person's conscience requires him or her to perform actions that are permitted by the general formulations of the freedoms in the UDHR and supported by the freedom of conscience (Article 18), but which also happen to be restricted according to the local interpretations of Article 29 which was discussed above. This occurs when a person's conscience tells him or her to act contrary to the expectations of the surrounding society. A well-developed MBS should be capable of identifying such clashes, and a well-developed PAS should contain methods for dealing with them.

The other operational interpretation of 'conscience' comes into play when a person experiences having 'a bad conscience' for something that he or she has done, or has chosen not to do. In this case it is a question of retroactively assessing an action of one's own and wishing that one had done otherwise. This can lead to new concrete actions, such as to apologize for example, but it can also lead a robot to reconsider some parts of its MBS, or even some aspects of its decision-making machinery. In other words, the important aspect of conscience in its second function is not the identification of the fault, since this can also be taken care of by the PAS and the MBS. Instead, the capability of changing oneself as a result of the remorse shall be understood as an additional competence that is an integral part of the conscience.

The view of the conscience as an inner judge is a powerful one, but it does not cover all uses of the term. In particular, conscience can also be seen as the basis for obligatory situation-action rules, which say that in a particular kind of situation, the agent must necessarily perform a certain action, regardless of other considerations. This may be related to the concept of identifying the 'Is' and the 'Ought', as advocated by Hallaq [17]. He explains it through an example where the observation that a person Is poor shall be ontologically identified with the requirement that one Ought to help the person. The example is not entirely convincing, but the idea is interesting.

To summarize, a fully developed MBS (consisting of both a software engine and a collection of value statements) should make the robot capable of revising the set of value statements in its MBS, both by plain deliberation and by conscience-driven reconsideration of those statements. This possibility is very much in line with Pico's view of how the soul can ascend from the physical world and towards the divine world.

Finally, we shall return to the interpretation of the phrase 'in a spirit of brotherhood' which occurs in Article 1 of the UDHR. The website of the UDHR Project at Columbia University defines this phrase as follows, in its page on Article 1 [18]:

> To treat one another in a "spirit of brotherhood" means that individuals should, in a figurative or symbolical sense, treat each other in such a way as proper to the relation of a brother.

This explanation leaves a lot to the reader's imagination. For the present purpose I shall assume that a spirit of brotherhood consists, at least, of 'empathy' and 'solidarity'. It may be argued that 'understanding' of the other person's character should also be included as a separate point, but I will leave that aside as being too complicated to deal with. Empathy, then, might be characterized as a capability of the agent's perception system whereby it is able to interpret its observations of others in terms of its model of itself and its own experiences, combined with an action or goal suggesting mechanism that reacts appropriately to these interpreted observations. With this understanding of the phrase, 'empathy' is closely related to the variety of 'conscience' that takes the form of obligatory situation-action rules.

The concept of 'solidarity' may be seen as analogous to 'empathy' but with the difference that it does not rely so much on the observer's model of itself. A person may show solidarity with the needs of his or her brother although they do not experience having the same needs themselves.

These definitions of terms such as 'empathy' and 'solidarity' must be seen as first approximations which can not of course do full justice to these words as they are used in psychology or in politics, and even less when compared to how they are used in literature. However, one must start somewhere, and even these crude operational definitions may be useful for relating terms such as these to the design of intelligent robots whose behavior may have some resemblance to 'empathy' and 'solidarity' in humans.

16 Potential Problems with Morality-based Robots

With respect to software systems that have a limited level of "intelligence", it is sometimes said that "limited intelligence is worse than no intelligence at all". Be that as it may, one may worry that an analogous statement may hold some truth: "a system with limited morality and understanding is worse than a system with none of those". After all, the very point with

endowing an intelligent robot with a sense of morality can only be to keep it from doing things that are evil or have evil effects, and to incite it to do things that promote goodness. Furthermore, there is an underlying assumption that the robot shall be enabled to make these judgements autonomously. This can only be worthwhile if the considerations in question are so complicated (or so incompletely known) that they can not be completed in advance. If they can, then one may as well resort to a preprogrammed ethical behavior, without any need for a 'free will'.

But if the situations that the robot will encounter can not be predicted and circumscribed, and if the robot's behavior in those situations can not be predicted in advance, then how can we know that the interventions of the robot will have positive effects in all the situations that it encounters? And if we deal with this concern by asserting that the robot's actions are for the good in the overwhelming majority of cases, although maybe not all of them, then how shall we as a society relate to those situations where the robot's actions turned out to be quite detrimental? Shall they be considered as accidents that are caused by Nature, or shall someone be considered as responsible–the robot's designer, the robot's owner, or the robot itself?

Besides these major problems, there is also an issue even in the case where an intelligent robot is merely used for observation and intelligence-gathering. We have already remarked that there is a possibility that the Moral Belief System can influence the world model in some situations. By way of speculation, one may imagine a robot that has too much morality and autonomy for its common sense; if such a robot is assigned the task of observing public spaces and if it can file a report or send in a task force when it sees that something bad is about to happen–then what? The concept of morality-based surveillance has some obvious problems.

What we have said here should not be used as arguments against the implementation of ethics in robots and other computer based systems. However, it does qualify as arguments for proceeding slowly and carefully when this kind of technology is put into use. Information technology has a deplorable track record of how new technologies and new systems have been put into operation prematurely. We should not let that happen for morality-based robots.

17 Alternative Uses of the Considerations Made Here

The present article has focused on the question whether the UDHR's statements about the nature of humans can be applied to the design of intelligent robots. There is a related and important issue that we have not addressed, namely, how an intelligent robot shall be designed so that it does not violate the human rights of any person in their environment. One may suggest that this can be done simply by including the articles of the UDHR to the robot's Moral Belief State, but it seems that this would impose an overly difficult task on the MBS software. It seems more likely, therefore, that the rules of the UDHR shall have to be "compiled manually" into value statements and behavior rules that are suitable for being integrated into the robot's MBS.

18 Alternative Definitions of Dignity and Morality

In this article we have used definitions of dignity and morality that are characteristic of Western culture during the period of modernity. Although our two selected authors embraced Christianity, more or less, their concepts and lines of reasoning did not reflect the religious doctrines at the time.

However, the dignity of the human person is also an important concept in several religions, and this should be mentioned briefly here as a reminder, and so that this aspect is not ignored. In religious frameworks, 'dignity' is often seen as a requirement to conform to the religious laws. For example, the Islamic scholar Mohammad-Ali Taskhiri views dignity as a state to which all humans have equal potential, but which can only be actualized by living a life pleasing to the eyes of God [19].

The concept of morality is likewise very important in many religions. For example, the Muslim scholar Wael Hallaq wrote [17]:

> …the discursive world of Islam and its forms of knowledge were pervaded by moral prescriptions and by Sharía-prescribed ethical behavior.

Religious frameworks of this kind can be seen as an alternative or as a complement to secular definitions, such as the largely secular one given by Kant. However, inasmuch as abrahamic religions consider that mankind was created in the image of God:

> So God created mankind in his own image, in the image of God he created them; male and female he created them.

(Genesis 1:27), it seems quite unlikely that their followers shall be willing to extend their concepts of dignity and morality to man-made devices, such as intelligent robots.

19 Summary and Concluding Remarks

In order to address the role of Ethics and of Human Rights in the design of intelligent robots, we have specified a few basic assumptions about the software architecture of such robots. We have also identified a set of important concepts in ethics,

discussed their definitions, and described how they may be realized in the robot's architecture.

The major assumptions about the architecture of the intelligent robot were that it should contain a sensory-motoric system, a perception capability that produces a world model, and a planning and action system (PAS). We have furthermore proposed the addition of a moral belief system (MBS) that interacts with the PAS. The assumed characteristics of these subsystems have been described in outline.

With respect to the concepts in ethics, we have chosen the Universal Declaration of Human Rights (UDHR) as the starting point for our analysis. It specifies four properties that are stated to be characteristic of human beings, namely 'freedom', 'dignity', 'reason' and 'conscience'. The concept of freedom is relevant since the UDHR does not merely claim a number of specific freedoms; it also contains clauses that restrict those same freedoms. The other three concepts are important since they characterize those properties of human beings that make us worthy of the freedoms and rights that are defined in the UDHR. Each of these four concepts is therefore relevant for the ethical aspects of intelligent robots.

The discussion of these properties could not be based only on the text of the UDHR and available comments about it. We have therefore included some key ideas of a few foundational authors into our analysis, in particular Giovanni Pico della Mirandola and Immanuel Kant. This also led us to add a few additional concepts into the analysis, in particular 'free will' and 'morality'.

The main result of these considerations is that it provides a conceptual framework for the design of a Moral Belief System that can impart a sense of ethics to an intelligent robot. Furthermore, this conceptual framework may clarify the relationship between the philosophy of ethics and the ethical capabilities of intelligent robots.

The use of the UDHR as the starting point for this work was natural in view of its very broad acceptance, and since it specifies characteristic properties of humans that are relevant for our topic. However, the UDHR also has some weak points. In particular, it talks mostly about 'freedoms' and 'rights', but only indirectly about 'obligations'. On the other hand, the analysis of 'dignity' as requiring the combination of 'free will' and 'morality' compensates to some extent for that weak point.

Further analysis of the obligations of intelligent robots would be a natural topic for future work. It is part of a broader topic that has been mentioned above, namely, what will be the appropriate laws and other rules for intelligent robots, and what social expectations should be applied to them.

Acknowledgements This article has been much improved thanks to the insightful comments of the three anonymous reviewers.

Funding This work was not done as part of any grant.

Compliance with Ethical Standards

Conflict of interest The author declares that he has no conflict of interest.

References

1. United Nations, General Assembly resolution 217 A (1948) Universal Declaration of Human Rights. http://www.un.org/en/universal-declaration-human-rights/index.html
2. della Mirandola PG (1486/1496) Oration on the dignity of Man (De hominis dignitate)
3. Kant I (1781) Critique of pure reason (Kritik der reinen Vernunft). Riga
4. Kant I (1788) Critique of practical reason (Kritik der praktischen Vernunft). Riga
5. Drafting of the Universal Declaration of Human Rights. A collection of documents. Dag Hammarskjöld Library. https://research.un.org/en/undhr/draftingcommittee
6. Asimov I (1950) I, Robot. Gnome Press, New York
7. Anderson M, Anderson SL (2007) Machine ethics: creating an ethical intelligent agent. AAAI Mag 28(4):15
8. Hew Patrick Chisan (2014) Artificial moral agents are infeasible with foreseeable technologies. Ethics Inf Technol 16(3):197–206. https://doi.org/10.1007/s10676-014-9345-6
9. Palmquist S (ed) (2010) Cultivating personhood: Kant and Asian philosophy. De Gruyter, Berlin
10. Ghallab M, Nau D, Traverso P (2016) Automated planning and acting. Cambridge University Press, Cambridge
11. Kant I (1785) Groundwork of the metaphysics of morals. In: Gregor M (ed) Cambridge University Press, pp 53–74. ISBN 9780521626958. OCLC 47008768
12. McCarthy J (2000) Free will—even for robots. J Exp and Theor Artif Intell 12(3):341–352
13. Compatibilism (2002/2015) Stanford encyclopedia of philosophy. https://plato.stanford.edu/entries/compatibilism/
14. The Definition of Morality (2002/2016) Stanford encyclopedia of philosophy. https://plato.stanford.edu/entries/morality-definition/
15. Martins J P, Reinfrank M (eds) (1990) Truth maintenance systems. In: Proceedings of the ECAI-90 workshop. Springer lecture notes in computer science
16. Johnson-Laird PN, et al. Reasoning about inconsistency—list of relevant publications. Mental models and reasoning. Princeton University. http://mentalmodels.princeton.edu/portfolio/reasoning-about-inconsistency/
17. Hallaq WB (2012) The impossible state. Islam, politics, and modernity's moral predicament. Columbia University Press, New York
18. Danchin P Article 1: Fundamental Human Rights. A page in a website about the Universal Declaration of Human Rights. http://ccnmtl.columbia.edu/projects/mmt/udhr/article_1/meaning.html
19. Taskhiri M-A (1997) Human rights: a study of the universal and the islamic declarations of human rights. Islamic Culture and Relations Organization, Alhassanain institute, Iran, p 92. http://alhassanain.org/english/?com=book&id=1153

Erik Sandewall is an emeritus professor at Linköping University, Sweden, and a member of the Swedish Academy of Sciences. His major research interests are in knowledge representation, reasoning about actions, and the use of these topics in the software architecture of intelligent and autonomous robots. His contributions involve both theoretical work and application projects, in particular concerning the design of intelligent UAV.