sciendo

STUDIES IN LOGIC, GRAMMAR AND RHETORIC 63 (76) 2020

DOI: 10.2478/slgr-2020-0033

This work is licensed under a Creative Commons Attribution BY 4.0 License (http://creativecommons.org/licenses/by/4.0)

Paweł Polak
The Pontifical University of John Paul II in Cracow
Faculty of Philosophy[1]
e-mail: atpolakp@cyf-kr.edu.pl
ORCID: 0000-0003-1078-469X

Roman Krzanowski
The Pontifical University of John Paul II in Cracow
Faculty of Philosophy
e-mail: rmkrzan@gmail.com
ORCID: 0000-0002-8753-0957

PHRONETIC ETHICS IN SOCIAL ROBOTICS: A NEW APPROACH TO BUILDING ETHICAL ROBOTS

Abstract. Social robotics are autonomous robots or Artificial Moral Agents (AMA), that will interact respect and embody human ethical values. However, the conceptual and practical problems of building such systems have not yet been resolved, playing a role of significant challenge for computational modeling. It seems that the lack of success in constructing robots, *ceteris paribus*, is due to the conceptual and algorithmic limitations of the current design of ethical robots. This paper proposes a new approach for developing ethical capacities in robotic systems, one based on the concept of Aristotelian phronesis. Phronesis in principle reflexes closer human ethics than the ethical paradigms we employ today in ethical robotics. This paper describes the essential features of phronesis and proposes a high-level architecture for implementing phronetic principles in autonomous robots. Phronetic robotics is in its early stages of conceptualization, so many of the presented ideas are speculative and require further research.[2]

Keywords: phronetic ethics, ethical robots, Aristotelian ethics, artificial moral agents, socially integrated autonomous machines, roboethics, machine ethics, autonomous ethical systems, phronesis, ethical robots, ethical ascend.

Introduction

Ethical systems in autonomous robots (a-robots) represent a mixture of deontological, utilitarian, and/or pragmatic ethical principles (a version of Kant's categorical imperative or Mills/Bentham's utilitarianism) (Allen et al., 2006; Anderson & Anderson, 2011, 2014; Floridi & Sanders, 2004; Hibbard, 2015; Wallach & Allen, 2009; Wallach et al., 2008; Wal-

lach et al., 2010). However, deontological and utilitarian principles are controversial even as an explanation, justification, and guidance for human conduct. There is no doubt that the current approach to robotic ethics could lead to the construction, to some degree, of safe-to-use a-robots (i.e., safe for humans). In this case, "safe" means the same as it would for the safety of an aircraft or a car (i.e., minimal risk of injury or harm, intentional or accidental). However, to an ethical person, the ethical behavior of these machines would look like a gross oversimplification of ethical conduct, which it basically is. In this paper, we outline a new approach to robotic ethics that is inspired by an aspect of Aristotelian ethics, namely phronesis. Phronesis has been used for centuries to rationalize human ethics, so it has the potential, if implemented, to improve ethics in a-robots as well.[3]

2. Roboethics today

Today's ethics in a-robots is implemented using the concepts of the abstract computational model of the Universal Turing Machine (UTM), as indeed all computer software is. From the ethical perspective a-robots based on Universal Turing Machines can only be behavioral systems.[4] The behavioral approach to morality has been long discounted.[5] Thus, if our goal is to develop socially acceptable autonomous robots the UTM based robotics implements the wrong ethical model in the wrong computational paradigm.[6] This may seem like a gross simplification of the research into ethical robotics. However, on close scrutiny, the various computerized ethics implementing emotional rules, virtues, and mixtures of deterministic to probabilistic algorithms (see later in the text) seem to be variations on the same theme: UTM-based a-robots are ethical zombies. When a-robots enter our social environment, we need them to be more than behavioral systems. We need them to think and make decisions like we do (Oers & Wesselman, 2016; Leite et al., 2013) and share our social norms and values, so we can accept them as social partners, particularly if we plan to grant them some of our rights! (European Parliament, 2017, *Rule* 56.f; Alaieri & Vellino, 2016)

So far, no robots have demonstrated better ethics and made better ethical decisions than humans (despite the fact that we err so often). This should be self-evident from the assumptions behind computer models. The autonomous robots with ethical capacities (e-robots) have access to the facts (the information about the current environment at a given time),[7] past eth-

ical decisions, and past outcomes (whether positive and negative) of actions on an incomparably larger scale than we humans do. In addition, the Artificial Moral Agents (AMAs) could potentially have incomparably more powerful inference capacities available to them than we have. Thus, the limitations of robotic ethics lay not in the volume of information about past events or the information about the state of the world that is available to a-robots. It is not about computational power either. So, what is it? We make the claim that one of the critical differences, although not the only one, between a-robots as currently designed and humans is in the "inference/ascend method", i.e., the method of ascending to an ethical decision based on the facts present, the set objectives of an action, and the comprehension of the ethical dimension of past experiences and their outcomes. To improve the moral capacities of a-robots, we need a new approach rather than more of the same. We need a paradigm shift in "roboethics".[8] This paper proposes a new model of ethical ascend and rationalizing of ethical behavior for robots that is inspired by the Aristotelian concept of phronesis.

3. Key concepts

Let us define the key concepts used in this paper to ensure a correct interpretation. An agent is a person or an artifact that takes an active role or exerts a specified effect. Autonomous robots are artificial systems capable of acting independently (of human assistance) in the physical environment for a certain amount of time, with no human activity involved in the decision loop; they can have a role of an agent. Ethical robots, or e-robots/AMAs, are artificial systems with attributed ethical capacities. A phronetic robot, or ph-robot, is an e-robot with attributed phronetic ethical capacities. Ethics is defined here as the theoretical and practical principles that guide humans' actions and choices. Machine ethics is a branch of computer science that deals with the formulation and development of ethical-like capacities in a-robots, or more generally, in autonomous computer systems. It is not a branch of ethics. As well, it is not concerned with professional ethics. Machine ethics is frequently used synonymously with robotic ethics. One may use the term roboethics, denoting ethics or machine ethics within the context of a-robots. Ethical ascend is the decision process for ethical problems. Thick or deep ethics is ethics with the goal of preserving an individual or common good, being guided by the principles of virtue and directed toward perfecting the moral capacities of an individual. Thick ethics is defined as opposing shallow ethics, such as utilitarian pragmatic, emotive, and

behavioral ethics (e.g. Väyrynen, 2017). A strong phronetic system is an artificial or natural system that possesses Aristotelian phronesis. A weak phronetic system, meanwhile, is an artificial system that emulates Aristotelian phronesis in some respect.

Note #1: This paper does not discuss the Aristotelian concept of phronesis in its full scope (e.g. Broadie, 1991; Flyvbjerg, 2011). The presented analysis of the Aristotelian concept of phronesis is there, out of necessity, simplified, but in its general outline, it aims to agree with modern interpretations of Aristotle's ethics and phronesis (Broadie, 1991; e.g. Reeve, 1992). Thus, we do not claim to have fully computerized Aristotelian phronesis. Instead, we propose a method inspired by the concept of phronesis, which is a weaker claim, but in our view, it is a technically feasible one (more on this methodology see Polak & Krzanowski, 2020).

Note #2: Other authors like e.g., Nick Bostrom, Eliezer Yudkovsky (2004), or Tarleton (2010) explored non-algorithmic approaches to robots ethics under the labels of Friendly AI or coherent extrapolated violation (CEV). None of the authors one explicitly mentioned Aristotelian phronesis. Yet, they indicated the need for a-robots to possess the concept of ethical situation.

4. Why phronesis?

We propose that the higher levels of ethical proficiency in a-robots may be realized by adapting the Aristotelian concept of phronesis.[9] What is phronesis? Phronesis is not an exact science (like episteme in Aristotelian sense) or art (the techne/craft or practical skills required to do something or produce something), and unlike the exact sciences, it is not based on ultimate (necessary) principles. The general principles of an exact science are "absolute," because they express abstract, nominal truths (like mathematics or pure logic). Ethics deal with "enmattered things"[10] and with changeability and variability inherent in concrete embodied facts. The objective of phronesis is eudaimonia, a realization of a specific concept of good for a person (an actor). The focus of phronesis is the variability of the specific case from ultimate principles rather than the "generality" of the case.[11] Phronesis cannot be taught like mathematics, as there are no rules to teach in phronesis.[12] Phronetic expertise can only be gained through experience. We do not provide a detailed explanation of what phronesis is, because others have been doing it already for some time (Reeve, 1992; Randall, 1965). We focus here on the phronetic ascend exclusively. Phronetic decisions can be summarized in the following principles:

A. Phronetic ascend cannot be encapsulated in a set of general rules, because it deals with the specificity of particular cases (i.e., it is not procedural).
B. Phronetic ascend is case-specific (i.e., the focus of the phronetic decision is a particular case).
C. The decision for the specific case is not achieved through logical analysis or reasoning through principles (like in science) but rather by intuitively grasping the outcome or having "foresight of consequences."

This is what Aristotle, presented here in the highly condensed way, is telling us.

> Note #3: The physical constitution of the body and the social setting play both important role in the comprehension of the complex ethical situations. This seems to has been Aristotelian view as well. The design of e-robots should account for these factors.
>
> Note #4: It also seems that the e-robots may (self) improve their ethical capacities by learning from their own and others, responses to situations. This means that rather than to program the e-robot with the response to any possible situation, which is impossible task, an e-robot should have self-improvement capacities and an ability to perfect its ethical skills. This self-improvement capacity should be also regarded as a critical aspect of autonomy, i.e. the autonomy of an a-robot should be defined not only as an ability to act without human supervision but also as an ability to autonomously improve its capacities to act autonomously.
>
> Note #5: One promising approach to building an e-robot with the deep ethics capacities is through the resolution to symbol grounding problem, which may be accomplished by adopting selected theories of biosemiotics. Current research on biosemiotics and robotics seems to be promising in this area (see Anna Sarosiek work, ongoing).

5. Phronesis lessons for roboethics

Phronetic ascend focuses on developing a decision system that is not based on general principles, generalizations, algorithmic procedures, or whatever else the current methods are but rather on the analysis of selected past histories for decisions and actual cases with outcomes. Each decision must consider the specificity of the particular case/situation, and the characteristics of the particular case should be the starting point for the ethical ascend. As phronesis is a function of experience, phronetic robots are learning robots with a feedback loop, and a new experience in an individual unit or group of units would be added to the collection of cases in the use

case (UC) database or ethical knowledge base (EKB). Both negative and positive decisions are stored in the EKB. The size of the EKB is not critical, as a large amount of "experience" does not guarantee proper foresight. It is the right experience with the right insight into the outcome of decisions that is needed. Table 1 interprets the principles of phronesis in the AI context.

Table 1

Phronesis and roboethics implications

Phronesis	Roboethics
Ethics is not an exact science; one cannot define generalized universal rules to guide behavior (i.e., decisions are taken in particular cases, but general rules are always generalizations (sic!), so they will sometimes be wrong in particular cases).	This implies that the development of algorithmic rules to solve ethical problems will not improve the quality of ethical decision-making.
Ethical decisions are all case specific and guided by the weight of experience.	This mean that phronetic capacities are not "programmed" a priori but learned.
Ethical decisions always go from the specific case to the universal principle, not from the universal principle to the specific case.	This means that the focus of an ethical decision is the ethical aspect of the specific case, not universal rules that generalize cases.
The *telos* of ethics (the ultimate goal of ethics) is eudaimonia (the best good).	The *telos* of ethical decisions (in the case of a-robots) is the best good of the human actors involved, not the a-robots themselves. Other formulations of *telos* for a-robots would also be possible.

6. Proposed design

We outline one possible approach to the design of a phronetic robot. This proposed design is clearly a simplification of the original concept of phronesis (i.e., we implement a weak phronetic system), but it retains its characteristic features. In the proposed system, the ascent to the decision should made through evaluating existing UCs and comparing them to the specific situation, with the decision being made by selecting the most "ethically proximal" UC with the best outcome. The exact meaning of what the "ethically proximal case" is needs a strict definition that can be operationalized. What makes two or more UCs ethically similar or what makes a UC relevant to the situation is not clear or obvious. The past UCs are not literal retentions of the past decisions but rather abstracted situations. What is abstracted is not the material particularity of a situation but its

ethical import. As vague as this statement is, it bars us from making decisions based on a simple "match" of the physical variables that describe the past cases.

The explanation in phronetic robots would not be of the form of deductive reasoning from the general to the specific (e.g., "I did this because of this, then this, etc.") or a probabilistic choice (e.g., "This action has the higher probability of success."). It would rather take the form of thinking, "I did this because through an analyses of similar cases in the light of my ultimate aim, I realized that it had a greater possibility of approaching my ultimate aim." Of course, the use of 'I' here is purely illustrative. What follows is an outline of a high-level view of the ethical ascend process in an a-robotic system. The models are based on the general agent-environment model presented in Figure 2.

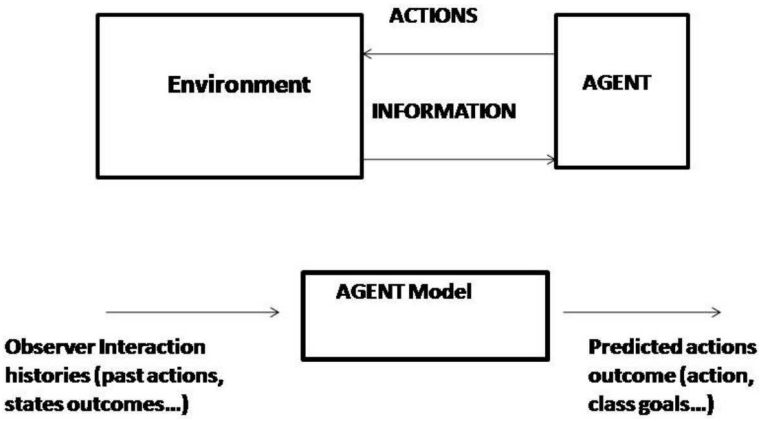


Figure 2. A general agent-environment reference model

The agent-environment model is a relatively stable concept. The models of agent-environment interaction may differ in their detail but not in their overall form (see e.g. Albrecht & Stone, 2018; Hibbard, 2015; Anderson & Anderson, 2008, as well as other models Poole & Mackworth, 2017). The model schema where the agent interfaces with the environment is preserved implicitly in any a-robot model.

We first need a few definitions for the terms agent, decision process, ethical fact, and ethical knowledge base.[13] An agent π obtains a set of observations $O_{1i}, O_{2i}, \ldots, O_{ni}$ about its environment (the state of the world). The index i denotes the specific set of observations (about the state-of-the-world); it may be the time index. An agent initiates the action a_i, resulting in an outcome q_i with a score s_i. We therefore have $\pi(O_{1i}, O_{2i}, \ldots, O_{ni}) \Rightarrow a_i$

and $a_i(q_i) \Rightarrow s_i$. F is a fact about a given situation (the state of the world in a given instance). This information is gathered by the agent from its environment. O_{1i} are specific instances of information. The collection of O_{1i} constitutes the fact (i.e., $F_i = (O_{1i}, O_{2i}, \ldots, O_{ni})$). The use case (UC) comprises the state of the world, the action, the outcome, and the score, such that $\text{UC} = \pi(O_{1i}, O_{2i}, \ldots, O_{ni}) \Rightarrow a_i$ and $a_i(q_i) \Rightarrow s_i$. The Ethical Knowledge Base (EKB) is the set of UCs, such that $\text{EKB} = [\text{UC}_1, \ldots, \text{UC}_n]$.

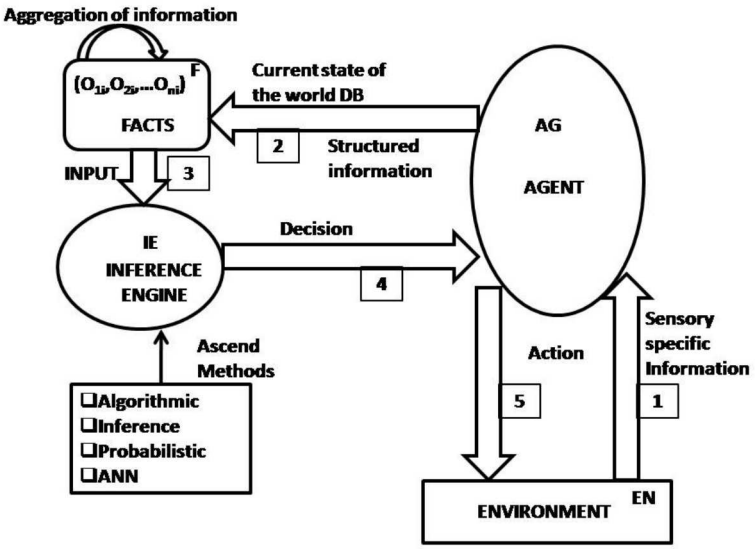


Figure 3. The non-phronetic AMA decision pathway

The decision process in current AMAs follows the decision pathway shown in Figure 3, which presents a very-high-level functional breakdown of the main components of the AMA decision system and hides the low-level details.[14] The main functional components of the AMA agent are EN (the environment), AG (the agent), F (the database of facts), and IE (the inference engine). The I/O and functions of each component are given in Table 2.

The AMA agent obtains information about the environment through it sensory systems (1), processes it into the proper format, and stores it in its database F of FACTS [$\text{F}_i = (O_{1i}, O_{2i}, \ldots, O_{ni})$] (2). The facts describe the situation of an agent. When making a decision, the agent's inference engine (IE) takes facts relevant to the situation (3) and processes them using one of the processing methods implemented in the agent. Based on the outcome of the decision algorithm, the IE passes the decision back to the agent (4). The agent then acts in the environment (5).

Table 2

Functional components of the AMA agent

Functional Block	Input	Output	Functions
EN	Decision by AG	Information about the situation	Environment
AG	Information about the situation through the sensory data stream	Decision and action	Collect information about the situation. Structure information; Pass information to DB of facts; Receive decision; Act
F – DB	Information about the situation	Respond to queries from IE	Collect and aggregate facts; Respond to queries
IE	Structured information about a situation	Decision to AG	Using the ascend method decide the action

The decision process for the phronetic ascend follows the decision pathway shown in Figure 4. The main functional components of the AMA agent with phronetic capacities are EN (environment), AG (agent), F (the database of facts), the Agent EKB of use cases, and IE (the inference engine). The aggregation of information and the aggregation of UCs are processes that integrate the particular facts or UCs into the database structure. As we said earlier, the database contains aggregated abstractions rather than raw data. The I/O and functions allocated to each component are listed in Table 3.

The AMA agent obtains information about the environment through its sensory systems (1), processes it into the proper format, and stores it in its database F of FACTS [$\mathrm{F}_i = (O_{1i}, O_{2i}, \ldots, O_{ni})$] (2). The facts describe an agent's view of situation. When making a decision, the agent's inference engine (IE) takes facts that are relevant to the situation (3). The IE then searches the Agent UC EKB for UCs that are similar according to the specific proximity measure (4) and selects the decision that offers the most favorable outcome (5). The decision is passed back to the agent (6), which then acts in the environment (7).

As we indicated earlier, the UC EKB contains information about the state of the world, actions, outcomes, and ethical scores. The results of actions are collected and feedback to the agent for integration in the UC EKB, as represented in Figure 4 by a dotted arrow. The presented schema of the functional blocks and decision pathways in the AMA with phronetic capacities is a high-level view that does not address the operational architecture of the system.

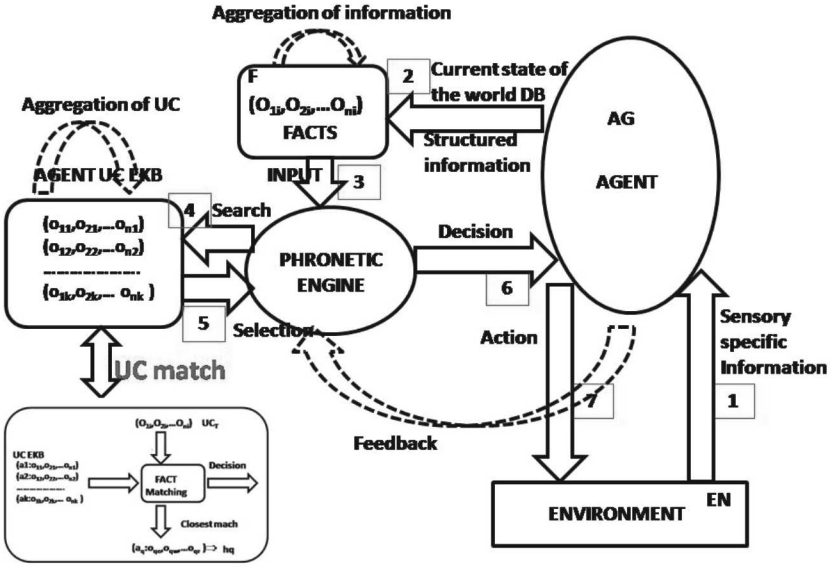


Figure 4. The phronetic AMA decision pathway

Table 3
Functional components of the AMA agent with phronetic capacities

Functional Block	Input	Output	Functions
EN	Decision by AG	Information about the situation	Environment
AG	Information about the situation through the sensory data stream	Decision and action	Collect information about the situation Structure information; Pass information to DB of facts; Receive decision; Act
F – DB	Information about the situation	Respond to queries from IE	Collect and aggregate facts; Respond to queries
IE	Structured information about a situation	Decision to AG	Using the ascend method decide the action
Agent UC DB	Phronetic engine request; Phronetic engine feedback	Search results	Collect and aggregate UC with decisions and outcomes; Respond to queries from a phronetic engine; Management of UC DB

7. Situational Theory and Phronetic Ethics

The first step in implementing phronetic ethics in robots is the development of the formalized representation of the ethical context including ethically relevant information, data base of ethical cases, ascend algorithms, abstraction methods and other elements of the computer phronetic 'engine'. The proposed approach to bridge phronetic ethics, robotic ethics and computer systems is to use the formalism of the situational theory as formulated by (Devlin, 1991). The situational theory is the study of cognitive agents, environment, and information flow between agents and environment. We present only the outline of the situational theory and its main elements to demonstrate how it can be used to model the phronetic ethics. The more detailed implementation, based on the Case-Based Reasoning (CBR) (see e.g. Aamodt & Plaza, 1994; Richter & Weber, 2013) is a subject of a future work.

The structure of the agent-environment interaction assumed by situational theory is following:

- An agent's world is divided up into a succession of situations.
- Situations, s, are highly structured parts of the world that the agent's behavior discriminates.
- What does and what does not constitute situation is a matter of agent's discernment, e.i., each situation is individualized.

The situations are defined by "information". Information is defined as an itemized, conceptual form, each item of information pertains to certain given object, e.g.: if P is an n-place relation and $(a_1, \ldots, a_n)$ are appropriate objects for this relation P, then $\langle\langle P, a_1, \ldots, a_n, i\rangle\rangle$ is a situation. $\langle\langle P, a_1, \ldots, a_n, i\rangle\rangle$ **states** the information item that $a_1, \ldots, a_n$, stands in the relation P. We call it infon. Infon represents situation. "i" is called polarity of the infon where "$i = 0$" if $(a_1, \ldots, a_n)$ are NOT in the relation P, "$i = 1$" otherwise. Infon that corresponds to the way things actually are in the world is called FACT. For example, the infon $\langle\langle X, 34, 21, 45, 0\rangle\rangle$ where X is the relation "is prime number" meaning that integers 31, 21, 45 are not prime numbers has the polarity "0" or "it does not correspond to the way things actually are in the world" i.e. it is not FACT. We represent this relation as follows. S denotes a situation. An infon denoted as σ is the fact. If σ is made true by s, we write $s \models \sigma$. If I is a set of infons and s a situation we write $s \models I$ if $s \models \sigma$ is for every infon in I. We may denote the maximal situation, the world or w, *considered* by an agent and we can write $w \models \sigma$ or $w \models I$.

Abstract situation is a set of infons, it is a mathematical construct built out of the relations, individuals, locations, time, polarity. Real situation is the kind of "in-the-world" relation picked up by individual schema (a set of infons). Each real situation corresponds to abstract situation or $\{\sigma \mid s \models \sigma\}$.

Situations (abstract or real) may be static or dynamic. For example, the following situation describes the "crossing the street at time t":

> s $\models$ $\{\langle\langle$Crossing, person, l, t, 1$\rangle\rangle \vee \langle\langle$Crossing, person, l, t, 1$\rangle\rangle \vee \langle\langle$behind, car, l, t, 1$\rangle\rangle \vee \langle\langle$in_front, crosswalk, l, t, 1$\rangle\rangle \vee \langle\langle$bad_weather, rain, l, t, 0$\rangle\rangle \vee \ldots \langle\langle R, a_1, \ldots, a_n, i\rangle\rangle\}$.

What elements constitute the situation is the matter of ethical abstraction.

The situations are denoting the factual things. But there is an obvious need to develop the concept of the situations-types or higher-level situations that are in some ways as a class of situations. For example, the trolley problem represents a certain type of factual situations. We need to include higher-order uniformities that are common across situations. We denote this uniformities as types of situations or situation-types or type-abstractions. Type abstraction is defined as

The type of (T) with parameters {temporal location, spatial location, individual, n-place relation, type of situation, infon, parameter, polarity [ACTION], [OUTCOME]} denoted as $\langle\langle$of-type, TYPE1, TYPE2, $\ldots, i\rangle\rangle$. Basic types T are values that can be assigned to type parameters. Parameters allow for further refinement of basic types and abstractions.

Situational logic seems to be useful in modeling phronetic ethics situational logic needs to include an element of the action and of the value of the outcome of an action. They have to be added. Table below provides the mapping between elements of Phronetic ethics and components of the situational logic.

Table 4
The mapping of concepts between Phronetic ethics and Situational Theory

Phronetic Ethics	Situational Theory
Fact	Actual abstract situation
Ethical Knowledge Base (EKB)	Data base of abstract situations and situation-types at different level of abstraction
Ethical Experience	Situations, and situation types with action value

The presentation of situational theory is necessarily brief and is provided here only to indicate that situational theory, after some extensions (e.g., action and valuation of action) may provide a conceptual framework can be used to formalize and operationalize Phronetic ethics in a computer system. Operationalization can be attempted only after several problems listed below are resolved.

8. Problems to be solved

We have identified four conceptual problems that need to be resolved before a phronetic robot can be implemented. These four are problems specifically related to the construction of the phronetic inference engine only It does not mean that these are all problems faced by the a-robot design. It also does not mean that once these problems are resolved the new problems will not appear. These four problems that we see from the current perspective are:

1. **The structure of the database of facts**: What should be included the database of facts and in what form is not defined. It seems obvious that information stored as facts must be in a highly abstract form – it cannot be raw sensory data. A database of raw data encompassing the lifelong experience of an agent would be too large to search efficiently. Only the relevant aspects of the data should be retained.

2. **The use-case-to-fact matching criteria**: It is not clear how to populate the ethical database with facts, what information should be retained, and how it should be integrated with the existing ethical UCs.

3. **The structure of use cases and the structure of the database of use cases**: It is not clear what should be stored as a use case and in what form. Again, it seems that the UC must be a form of highly abstract information yet retain the specificity of a particular case. There have obviously been attempts to design knowledge databases for robots, but these do not address the specifics of ethical problems (Tenorth & Beetz, 2017). There have also been attempts to define ontology for autonomous robots, but the ethical aspects of the robot "*universum*" are still not included (Prestes et al., 2013).

4. **The ethical proximity measure and searching**: The concept of proximity measure of cases must express ethical similarities of situations. How situations are ethically similar may only be crudely approximated by typical distance metrics used in CBR. At present we have intuitive understanding of what this proximity may be. It is not clear how we determine

when use cases in the ethical database are similar. How would we measure 'ethical' similarity or proximity? It is also not clear how searches in the ethical database should be performed. The implementation of a phronetic robot will need to address these problems. The ethical UCs will likely be represented as situational descriptions. It seems that when searching for proximal UCs, we could use string-matching algorithms (Hasna & Potolea, 2016; Bard, 2007; Gilleland, 2006; Navarro, 2001; Winkler, 1990; Jaro, 1989). These would search for objects that are similar rather than exactly the same. It is possible that we also need to define a new measure of distance or proximity for ethical UCs (ethical proximity). In addition, phronetic robotics does not free us from imposing ethical judgment – the UCs need to have some kind of "preference" score, however minimal. It is difficult to conceive a software system that could assign such a score to an outcome of an event without a basic conception of right and wrong. In short, we would not trust unsupervised ethical learning. Let us say these initial scores are our boundary conditions, but what would we regard as minimally required in this case?

5. **The problem of robotic ethics must address embodiment of the ethical mind.** Aristotelian concept of phronesis considers as an ethical agent a person, i.e. a fully constituted agent with the mind and body. Thus, it may be the case that the development of phronetic capabilities as a stand-alone software function will not be feasible or will be severely limited. The phronetic software functions developed as a standalone system will always be missing the 'body' component whatever it may be in a case for phronesis. While it has been recognized that the development of the a-robots a human-like with motor skills have to be based on the embodied systems (see e.g. Hoffmann and Pfeifer, 2012) it has not been recognized that embodiment is also a condition for developing ethics. However, reasoning by analogy, as it is impossible to explain to a color-blind person the beauty of a rainbow, it may be impossible to engineer phronetic-like capacities in an a-robot that is not fully functional in physical environment.

Of course, as we said, these are in no way all the problems that need to be addressed before we can build a phronetic a-robot. In our view, however, the conceptualization of use cases, and the development of matching and searching methods is fundamental to the proposed design and should consequently take priority. It seems that doxastic and epistemic logics may also provide a formalization framework for the phronetic decision system, ethical knowledge base, and the development of the phronetic ascend. The formal framework of doxastic logic may also help explain an agent's ethical ascend, creating explainable robotic ethics.

9. Conclusions

A-robots currently possess ethical prowess at the very rudimentary level, that we even should not denote these capacities as any kind of ethics in the full (or thick) sense of this term. The reasons for this are manifold including the limitations of our technology (we are still in any TM paradigm), the limitations of our conceptual models of ethics and relatively primitive (by comparison to what is needed to model human ethics) implementation methods. We therefore have little chance to build social robots with the skills of comparable to a human agent, as these skills require, among other qualities, e.g. intuitive reasoning, which our current algorithms lack.

We propose developing roboethics to replicate, to some degree, Aristotelian phronesis. Phronetic ascend (the way to make ethical decisions) is based on non-algorithmic reasoning that characterizes a human decision-making. Phronetic ethics brings are closer to than other ethical paradigms such as deontology or utilitarianism to how we humans make ethical decisions. Thus, phronetic ascend offers a potentially better approach to solving ethical problems than the current methods. However, any implementation of phronesis in a-robots requires the resolution of several technical and conceptual problems, some of which have been identified in this paper.

Even if we replicate phronetic decision-making in a-robots, a-robots will never assimilate completely the Aristotelian objective of phronesis (i.e., eudaimonia, a realization of specific concept of good for a person in Aristotelian ethics). This is so, simply because Aristotelian ethics is rooted in the concept of man/woman; it is as, we would say today embodied in a human flesh and mind. However, it remains open as to whether it is possible to define a eudaimonia in a narrow sense for a specific type of a robot, thus creating, in a limited sense, phronetic robots.

NOTES

[1] ul. Kanonicza 9, 31–002 Kraków, Poland

[2] A substantial part of this paper has been taken from our presentation "Modeling ethical phronetic knowledge bases in A-Robots", 14 May 2019, during the first edition of the research workshop "Computational Modeling" at Warsaw University of Technology – International Center for Formal Ontology (DOI: 10.13140/RG.2.2.22971.87848). Some ideas and philosophical problems concerning computational modeling of phronetic ethics were previously sketched in our article (Polak & Krzanowski, 2020).

[3] We should be aware that improvement of a-robots ethical behavior do not remove completely potential risks connected directly with machine ethics. For example (Cave, Nyrup, Vold, & Weller, 2019) have described four broad categories of risks: "A) the risk

that ethically aligned machines could fail, or be turned into unethical ones; B) the risk that ethically aligned machines might marginalize alternative value systems; C) the risk of creating artificial moral patients; and D) the risk that our use of moral machines will diminish our own human moral agency."

[4] Behavior can be described and explained without making ultimate reference to mental events or to internal psychological processes. The sources of behavior are external (in the environment) rather than internal (in the mind, the head). (Graham, 2017) The ethical limits of the UTM stem from the nature of the Turing Machine paradigm. It would seem than that any extensions like non-deterministic TMs (NDTMs), choice-machines, and oracle-machines (OTM) would not change the essence of the paradigm (i.e., eliminate its limitations regarding ethical capacities), so there is no mental event in the machine that forms the basis of human capacities (Searle, 1998).

[5] "Contemporary psychology and philosophy largely share Hempel's conviction that the explanation of behavior cannot omit invoking a creature's representation of its world. Psychology must use psychological terms. Behavior without cognition is blind. Psychological theorizing without reference to internal cognitive processing is explanatorily impaired." (Hempel, 1966).

[6] This is the whole discussion of the Chinese room. It is too well known to be repeated in detail here (Cole, 2019).

[7] As an example, autonomous cars have access to 4 Terabytes of data per day with the result that "self-driving cars will soon create significantly more data than people – 3 billion people's worth of data" (Nelson, 2016).

[8] Here we understand "computable" as meaning something that is implementable in current computing systems and nothing else. The model for these systems could be UTM. We need to note, that some parts of processing of ethical cases should be done by the embodied and cultural embedded responses, which are obviously not "computable" according to this naming convention.

[9] Classical concepts (like Aristotle's one) are especially pretended to modeling, because they were a subject of detailed ethical analysis, which reveals many important philosophical assumptions of it. For example the action theory reveals that AMA should be equipped with a conjunction of properties (e. g. freedom, consciousness).

[10] To enmatter is "to endue with or confine in matter" https://www.merriam-webster.com/dictionary/enmatter. Enmattered in case of ethics means that ethics does not deal with abstract cases but actual situations.

[11] "The decision in this case is not from universal principles to the case but from the case to the universal principles (arguing to the first principles (*Topics* 101a37–b4), not from the first principles, as exact sciences do)." (Randall, 1965).

[12] i.e., Ethics is not Euclidian geometry.

[13] We use the formalism proposed by (Hibbard, 2015).

[14] More detailed models have been defined (Arkin & Ulam, 2009).

REFERENCES

Aamodt, A., & Plaza, E. (1994). Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications*, *7*(1), 39–59.

Alaieri, F., & Vellino, A. (2016). Ethical decision making in robots: Autonomy, trust and responsibility. In A. Agah, J. J. Cabibihan, A. M. Howard, M. A. Salichs, & H. He (Eds.), *Social Robotics* (Vol. 9979, pp. 159–168). https://doi.org/10.1007/978-3-319-47437-3_16

Albrecht, S. V., & Stone, P. (2018). Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, *258*, 66–95. https://doi.org/10.1016/j.artint.2018.01.002

Allen, C., Wallach, W., & Smit, I. (2006). Why Machine Ethics? *IEEE Intelligent Systems*, *21*(4), 12–17. https://doi.org/10.1109/MIS.2006.83

Anderson, M., & Anderson, S. L. (2008). ETHEL: Toward a Principled Ethical EldercareRobot. In M. Skubic (Ed.), *AI in Eldercare: New Solutions to Old Problems: Papers from the AAAI Fall Symposium*. Menlo Park, California: AAAI Press.

Anderson, M., & Anderson, S. L. (2011). *Machine Ethics*. Cambridge: Cambridge University Press.

Anderson, M., & Anderson, S. L. (2014). GenEth: General Ethical Dilemma Analyzer. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence* (pp. 253–261). Retrieved January 10, 2020 from: https://www.researchgate.net/publication/261871702_GenEth_A_General_Ethical_Dilemma_Analyzer

Arkin, R. C., & Ulam, P. (2009). *An Ethical Adaptor: Behavioral Modification Derived from Moral Emotions*. Retrieved January 10, 2020 from: http://www.cc.gatech.edu/ai/robot-lab/online-publications/ArkinUlamTechReport2009.pdf

Bard, G. V. (2007). Spelling-error tolerant, order-independent pass-phrases via the damerau-levenshtein string-edit distance metric. *Proceedings of the Fifth Australasian Symposium on ACSW Frontiers – Volume 68*, 117–124. Retrieved January 10, 2020 from: http://dl.acm.org/citation.cfm?id=1274531.1274545

Broadie, S. (1991). *Ethics with Aristotle*. New York/Oxford: Oxford University Press.

Bostrom, N. 2003. *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, Vol. 2, ed. I. Smith et al., International Institute of Advanced Studies in Systems Research and Cybernetics (pp. 12–17) [revised version available at https://nickbostrom.com/ethics/ai.html].

Cave, S., Nyrup, R., Vold, K., & Weller, A. (2019). Motivations and Risks of Machine Ethics. *Proceedings of the IEEE*, *107*(3), 562–574. https://doi.org/10.1109/JPROC.2018.2865996

Cole, D. (2019). The Chinese Room Argument. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2019). Retrieved January 10, 2020 from: https://plato.stanford.edu/archives/spr2019/entries/chinese-room/

Devlin, K. (1991). *Logic and Information*. New York, NY, US: Cambridge University Press.

European Parliament. (2017, February 16). *Civil Law Rules on Robotics European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL))*. Retrieved January 10, 2020 from: http://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_EN.pdf

Floridi, L., Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, *14*(3), 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Flyvbjerg, B. (2011). *Making Social Science Matter: Why Social Inquiry Fails and How It Can Succeed Again* (S. Sampson, Trans.). Cambridge: Cambridge University Press.

Gilleland, M. (2006). Levenshtein Distance, in Three Flavors. Retrieved December 3, 2019 from https://people.cs.pitt.edu/~kirk/cs1501/Pruhs/Spring2006/assignments/editdistance/Levenshtein%20Distance.htm

Graham, G. (2017). Behaviorism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2017). Retrieved January 10, 2020 from: https://plato.stanford.edu/archives/spr2017/entries/behaviorism/

Hasna, O. L., & Potolea, R. (2016). The longest common subsequence distance using a complexity factor. *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, 336–343. https://doi.org/10.5220/0006067603360343

Hempel, C. G. (1966). *Philosophy of Natural Science*. Upper Saddle River, NJ: Prentice-Hall.

Hibbard, B. (2015). Ethical Artificial Intelligence. *ArXiv:1411.1373 [Cs]*. Retrieved January 10, 2020 from: http://arxiv.org/abs/1411.1373.

Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association*, *84*(406), 414–420. https://doi.org/10.1080/01621459.1989.10478785

Leite, I., Martinho, C., & Paiva, A. (2013). Social Robots for Long-Term Interaction: A Survey. *International Journal of Social Robotics*, *5*(2), 291–308. https://doi.org/10.1007/s12369-013-0178-y

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys*, *33*(1), 31–88. https://doi.org/10.1145/375360.375365

Nelson, P. (2016). One autonomous car will use 4,000 GB of data per day. *Network World*. Retrieved January 10, 2020 from: https://www.networkworld.com/article/3147892/one-autonomous-car-will-use-4000-gb-of-dataday.html

Oers, R. van, & Wesselman, E. (2016). *Social Robots*. Retrieved January 10, 2020 from: https://assets.kpmg/content/dam/kpmg/pdf/2016/06/social-robots.pdf

Polak, P., & Krzanowski, R. (2020). Ethics in autonomous robots as philosophy in silico: The study case of phronetic machine ethics. *Logos i Ethos*, 54, 33–48.

Poole, D. L., & Mackworth, A. K. (2017). *Artificial Intelligence: Foundations of Computational Agents* (2nd edition). Retrieved January 10, 2020 from: https://artint.info/2e/html/ArtInt2e.html

Prestes, E., Carbonera, J. L., Rama Fiorini, S., M. Jorge, V. A., Abel, M., Madhavan, R., ... & Schlenoff, C. (2013). Towards a core ontology for robotics and automation. *Robotics and Autonomous Systems*, *61*(11), 1193–1204. https://doi.org/10.1016/j.robot.2013.04.005

Randall, J. H. (1965). *Aristotle* (4th ed.). New York: Columbia University Press.

Reeve, C. D. C. (1992). *Practices of Reason: Aristotle's Nicomachean Ethics*. Oxford: Clarendon Press.

Richter, M. M., & Weber, R. (2013). *Case-Based Reasoning: A Textbook*. https://doi.org/10.1007/978-3-642-40167-1

Searle, J. R. (1998). *Mind, Language and Society: Philosophy in the Real World*. New York: Basic Books.

Tarleton, N. (2010). *Coherent Extrapolated Volition: A Meta-Level Approach to Machine Ethics*. San Francisco, CA: The Singularity Institute.

Tenorth, M., & Beetz, M. (2017). Representations for robot knowledge in the KnowRob framework. *Artificial Intelligence*, *247*, 151–169. https://doi.org/10.1016/j.artint.2015.05.010

Väyrynen, P. (2017). Thick Ethical Concepts. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2017). Retrieved from https://plato.stanford.edu/archives/fall2017/entries/thick-ethical-concepts/

Wallach, W., & Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong*. Oxford: Oxford University Press.

Wallach, W., Allen, C., & Smit, I. (2008). Machine morality: Bottom-up and top-down approaches for modelling human moral faculties. *AI & SOCIETY*, *22*(4), 565–582. https://doi.org/10.1007/s00146-007-0099-0

Wallach, W., Franklin, S., & Allen, C. (2010). A conceptual and computational model of moral decision making in human and artificial agents. *Topics in Cognitive Science*, *2*(3), 454–485. https://doi.org/10.1111/j.1756-8765.2010.01095.x

Winkler, W. E. (1990). *String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage*. 354–359.

Yudkowsky, E. (2004). *Coherent Extrapolated Volition*. San Francisco, CA: The Singularity Institute