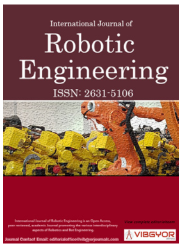
Research Article: Open Access

DOI: 10.35840/2631-5106/4126

VIBGYOR ePress

International Journal of
Robotic Engineering



ISSN: 2631-5106

Promises in the Context of Humanoid Robot Morality

Jan A Bergstra*

Minstroom Research BV Utrecht, The Netherlands

Abstract

Promise theory (PT) has been designed with a focus on information technology and systems design. I will use PT to outline the specification of robot behaviour and to discuss various forms of trust in connection with robots.

Introduction

Promise theory (PT) was designed by Mark Burgess in a series of papers from 2005 onwards[a]. For a survey of this work and extensions of it see Burgess 2015 [1] and the more technical exposition in Bergstra & Burgess 2014 [2,3][b].

The objective of this paper is to examine what promise theory may offer for the study of physical robots. Physical robots are distinguished from software robots, but may be under the control of an embedded software system. One may consider the embedded system of a physical robot to be a software robot. Physical robots perform their key functions by physical means using information processing as an auxiliary technology rather than as a primary means of action. With the language of PT at hand I will try to contribute to the topic of robot morality in various ways as specified below.

Outline of the paper

The contributions of the paper are these:

1. Yet another definition of robots is given. (That is yet another inconclusive solution of the robot definition problem.) Robots are portrayed as computers with as a consequence that acquiring knowledge about a specific robot (or family of robots) is supposed to follow the known lines of how to analyse a computer system. This is observation seems to be absent in current discussions of machine ethics as applied to robots.
2. Then PT is recapitulated and some philosophical aspects of its design are worked out in more detail than was done until now. PT is proposed as a human readable specification notation for humanoid robot behaviour.
3. Three stages of the acquisition of trust in a robot are distinguished: The extended quasi-Turing test (EQTT), the intrinsic trust establish-

[a] See also: http://markburgess.org/treatise.html

[b] Bergstra & Burgess 2017 [3] provides an extensive case study for promise theory outside the realm of informatics. It appears that promise theory is helpful for understanding systems of animate agents as much as it is for specifying systems consisting of inanimate agents.

Accepted: September 02, 2020; **Published:** September 04, 2020

***Corresponding author:** *Jan A Bergstra, Minstroom Research BV Utrecht, The The Netherlands, E-mail: janaldertb@gmail.com*

ISSN 2631-5106

Bergstra. Int J Robot Eng 2020, 5:026

ment (ITE), and the moral qualification dilemma (MQD). This description allows for a limitation of the MQD to its philosophical essentials, which itself is an issue, however, to which the paper has nothing novel to add.

4. Various forms of trust are discussed for a variety of robot classes: Factory robots, sex robots, service robots, care robots, and combat robots. Some differences between these robot classes with regard to the establishment of trust in a robot are noticed.

5. For the design of moral robots a specific form of ethical constructivism comes into play and Asimovian constructivism is proposed as a name for such constructivism.

Defining robots

A robot is a cyber-physical system with peripherals able to act in excess of mere communication. Where traditional peripherals are printer, keyboard, mouse, screen, touchpad, speaker, voice and voice recognition, external memory, a fixed position camera, and are more often than not unable to cope with bad weather, robots have hands, arms wheels, legs, eyes which can be directed and focused. Further robots may be able to smell, to sense heat and pressure, and to operate in adverse circumstances. Like computers robots can be reprogrammed.

It is remarkable that the instance of computer which is nowadays most devoid peripherals, viz. the mobile phone, is not anymore called a computer. As fewer users actually use their mobile as a phone it becomes even more of a computer. One may say that a robot is a computer which can do things which humans would call physical work. I will use a more systematic and perhaps less appealing definition of a robot below. At the same time I will insist that a robot is in essence a computer, which while it may choose not to make use of its embedded intelligence, is supposed to be equipped with a minimal level of intelligence, which according to one's perception of computer science may or may not be labeled as a manifestation of artificial intelligence.

Robots as Computers

I will assume that methodologies for the acquisition of knowledge about the behaviour of a robot (or about a class of robots) are essentially the same as for computer systems. The label artificial intelligence merely points to a subset of techniques from computer science and does not stand for principled new ways of thinking about computer systems. It might be informative to replace AI (artificial intelligence) by CS (computer science) on many occasions so that corresponding techniques and ideologies can be employed[c].

However, I will introduce various informal notions about robots which cannot be traced back to any classical notions in computer science. Below I will speak for (i) A robot being human-alike, (ii) A robot passing an extended quasi-Turing test, (iii) A robot allowing intrinsic trust establishment, and a robot passing a moral qualification dilemma, (iv) The decision taking competence of a robot, (v) The moral justification competence of a robot, (vi) The minimal moral justification competence, relative to its decision taking competence, that is required for a robot for it to acquire moral qualification for certain tasks, and (vii) A robot reasoning with Asimovian constructivism. The AI aspect of robotics encountered below is not so much in the technology for achieving certain forms of performance but in the terminology and concepts used for analysing the way the robot creates its behaviour in various contexts.

Definition 1: *A physical robot is a localized, but in its motion not necessarily spatially confined, monolithic physical actor under the control of embedded software, which underlies these constraints:*

1. *The robot engages in physical actions (including motion) and communications. It is through these actions that the robot achieves its objectives, if any.*

2. *Communications to other agents are of four kinds exclusively: Promising (issuing a promise), imposing (issuing an imposition), receiving (as the promisee) or noting (as an agent in scope) a promise, and receiving (as the promisee) or not-*

[c] Topics as LISP, PROLOG, SMALLTALK, semantic networks, natural language processing, Bayesian networks, planning, automatic proof generation and corresponding proof checking, automatic programming, have each started their life as AI and have in the mean time migrated (been downgraded) from AI to CS.

AI has become a label for automating tasks at which biological intelligence outperforms computer (i.e. artificial) performance for at least some 20 years after the task of automation has been identified and the work towards it has started at an international scale.

ing (as an agent in scope) an imposition.

Here promise, imposition, promiser, promise, and scope are understood as in PT.

3. *Besides communications to other agents the robot may have a datalink to one or more datacenters which allows the real time inspection of massive data as well as the significant remote processing, beyond the scope of the robot's on board computing performance.*

   However, it is not assumed that the actor owns any resources according to ownership as outlined in [4], and in [5].

4. *Moreover the robot is an artefact of human engineering, or an artefact of human engineering combined with robotic action, or an artefact resulting from the work of a team of robots. Robots growing other robots may pass through an unpredictable evolution.*

5. *In comparison to issuing promises, issuing impositions is much less frequent so that promises constitute the primary mode of communication of the actor under consideration.*

The idea of requirement 4 is to ensure that at least in principle the entire construction path has been documented in all possible detail and could, for that reason, be duplicated, as well as analysed. In other words the artefact is known, and in particular its embedded software is known. According to Mark Burgess allowing a robot to own resources is reasonable and the need for a definite demarcation from biologically grown entities is non-obvious. I will maintain both limitations for the sake of simplification of the subsequent exposition. Given a physical robot *R* its context may be split into various components:

- Theater agents: Agents inhabiting the theater where the robot is active,
- Background agents and structures (including datacenters),
- Responsible agents, agents who feel responsible for the robot's activities.

Other definitions of robots

The definition of robots above is just one of many conceivable and possibly competing answers to the following question:

Problem 1: *(Robot definition problem). What is a robot?*

Probably the robot definition problem admits no definitie answer and the requirements for being a robot are likely to steadily increase with advancement of technology. Lin, Abney, & Bekey [6] pay ample attention to the robot definition problem. According to these authors a robot is an engineered machine that senses, thinks, and acts. Their definition is not confined to electromechanical engineering, and may cover results of biotechnological engineering as well as of software engineering.

I will adopt the following conventions: If in addition to Definition 1 an artefact complies with the criteria of Lin., et al. then it is an autonomous robot. Otherwise it is a non-autonomous robot. Lin, et al. insists that a robot is autonomous by definition, thereby casting doubts on the clarity of the title of Bekey [7]. When discussing a robot soldier, however, Tamburrini [8] insists that the soldier, rather than the robot carries with it the implicit assumption of autonomy.

These distinctions are non-trivial, however, because the intelligence of an autonomous robot need not be the result of on board data and collection and processing thereof. The notion of ownership (as defined in Bergstra & Burgess [4]) is of use in Definition 1. If external IT services are owned by an agent (a candidate robot), the outcome of the use of such services contributes to the autonomous behaviour of the agent.

Karppi, et al. [9] propose that a phrase like "killer robot" must not be understood as a certain kind of robot amenable to a precise definition. Instead "killer robot" is part of a cultural technique which plays a role in some power struggle. Killer robot is used in a context and timeframe in which the notion of killing itself is in flux, and so is the notion of robot. As a phrase it helps to create a movement which, right or wrong transcends its immediately stated purposes. This is a convincing narrative which explains why the CSKR (campaign to stop killer robots) is not engaged in developing increasingly precise definitions of killer robots. The notion of an autonomous robot may perhaps in the same way be seen as a cultural technique meant to arrange and rearrange attitudes, and meant to separate opponents from proponents. The idea of autonomous robot as a cultural technique explains the idea of an autonomous robot imaginary as discussed in Rommetveit, et al. [10]. The latter paper analyses robot autonomy in terms of roadmaps; quoting [10]:

- The main organizing tool and metaphor in this co-construction work is that of the roadmap whose main characteristics stem from the peculiar fact that it depicts a road that has yet to be built (.), and that eventually will be the single road on the map.
- The roadmap indicates a plurality of gaps, between academia and industry and between law and engineering, and the bridging of gaps becomes a self-propelling metaphor. Even the gap between deterministic preprogrammed agency and consciousness is reduced to a mere pointer to a plurality of gaps the bridging of which is the perspective of roadmap protagonists.

Humanoids, androids, and gynoids

Humanoid robots are a fraction of conceivable robots only. Nevertheless contemplating humanoid robots as a special case, and doing so uniformly for a range of application areas, has advantages because certain aspects of morality and ethics can be discussed with ordinary human ethics and morality as a source of inspiration. I will need some definition of humanoid robots. Here is an option:

Definition 2: *A physical robot R is a humanoid (a humanoid robot) if (i) The thought experiment that R were a human being is both reasonable and informative, (ii) The thought experiment of R being fully under remote control of a human being is meaningful, (iii) In each direction of physical ability, competence or capability, the gap between an average human and R is a gradual matter and a continuous line of intermediate humanoids can be imagined, (iv) For each task the need for/added value of autonomy of R can be analysed and assessed (this assessment may range from concluding that the value of human supervision is limited, to the assessment that real time human supervision would stand in the way of adequate speed of operation and ability to react to unforeseen circumstances).*

By definition a humanoid robot is a physical robot. I will sometimes use humanoid as an abbreviation of humanoid robot. The Platonic love robots discussed in Nordmo, et al. [11] are software robots and for that reason do not qualify as humanoids under these assumptions. A humanoid robot is enhanced if its capabilities exceed those of average humans in important ways. Typically, physical strength, endurance, ability to survive adverse conditions, speed of action and reaction, built in data and image processing in combination with wireless connections and an ability to call for support of remote processing may constitute areas where the humanoid robot is superior to an average human. A male humanoid is also called an android, a female humanoid is also called a gynoid. A humanoid may be genderless as well. It remains to be seen if there is a rationale for contemplating transgender gynoids and transgender androids.

Adequate performance testing assumption versus adequate forecasting assumption: For humanoids it will be assumed that in principle it is feasible to perform adequate testing and simulation to be sure that the robot is able to perform as expected, unless its control functions for some reason override these abilities. This form of testing will be called performance testing. Here performance includes what is done as well as how fast it is done, or how cheap or whatever relevant measure can be applied.

The limitations of testing for cyber-physical systems are well-known. Testing can demonstrate that the system can perform certain patterns of behaviour (either in a positive sense, often termed demonstration, or in a negative sense if a failure occurs, which will lead to some form of repair), but testing cannot produce guarantees that such patterns will be shown or will not e shown under given conditions.

For instance one may wish a robot to act as a surgeon for some range of medical surgeries. In that case the adequate testing assumption implies that it is possible and feasible, by means of a combination of simulation and experimentation to confirm (or reject) the proposition that the robot is capable of such activities. Having confirmed such capabilities in principle for humanoid *R* it is still possible that *R* will unexpectedly (for its users) refuse to help a former business competitor of its manufacturer. Excluding such forms of undesirable behaviour requires a different form of analysis, which is able to provide reliable forecasting of what *R* will do and what *R* will not do. Clearly if *R* has been hacked it may be perfectly successful during performance testing while the corresponding adequate forecasting assumption has nevertheless become corrupted. Validating an adequate forecasting assumption as a whole or in parts, is a task which cannot be performed without taking the fundamental tools of computer science and software engineering on board: Formal specification and verification of hardware and software, model checking, software production lines, the vast technology of computer security, and the top level scrutiny which can justify trust in open source software.

Machine learning may lead to unpredictable behaviour of a computer, but that is a problem of a different nature than the difficulty of finding out if a computer has been hacked or if the software contains intentional mistakes which come to expression in irregular circumstances or on the basis of hard to detect external triggers. The situation changes if the dataset used for machine training has been intentionally corrupted by an adverse party. Give a trusted machine learning process, however, the statistical variations created by learning can, at least in theory, be factorized out from other concerns in connection with adequate forecasting.

A rationale for a restriction to humanoid robots: Suppose one contemplates care robots and besides humanoids one needs to contemplate as robots, a range of agents including nano-scale agents for intelligent delivery of chemicals inside a human body, unmanned rescue helicopters, and full scale unmanned and autonomous hospital ships, then taking human morality as a uniform point of departure is uninformative. By having a focus on humanoids, both the machine ethics dimension and the machine morality dimension are simultaneously simplified, of course at the cost of generality.

Autonomy

Autonomy is a property of robot activity which may vary over time. I will distinguish five levels of autonomy. In fact there is rather a continuum of such levels.

1. During a phase in which R has no interaction with any data center, and no interaction with any responsible agents, R is said to be "fully autonomous".
2. During an episode in which R has interaction with responsible agents only on its own request, and retrieves no information from its supporting data centers, R is said to be "semi-autonomous".
3. During an episode in which R has no interaction with responsible agents, and only retrieves mission independent information from its data centers, the robot is "fully autonomous with remote data provision"[6,7][d].
4. During an episode in which R has interaction with responsible agents only on its own request, and only retrieves mission independent information from its data centers, the robot is "semi autonomous with remote data provision".
5. During a phase in which R is under permanent control of one or more responsible agents, and has access to the relevant data centers, R is "remotely operated".

A humanoid robot is at least in some sense (at some level) autonomous, though its level of autonomy may vary over time.

Trust as a counterpart of promise: vital trust

An essential notion concerning robots is trust. Trust serves, more than obligation as a counterpart of promise. Trust is not easily defined let alone quantified. Some applications of the notion of trust are obvious: If the management of a hospital writes to the manufacturer that they do not trust surgical robot *R* to adequately carry out a surgery of type *S*, (on the basis of a recent fiasco) they will not, without having obtained any response from the manufacturer, impose precisely such a task on *R*.

But if you do not trust the quality of car maintenance of car dealer *X*, will you make use of a car that has just been serviced by *X*? In this case the distrust may have various causes. Perhaps one is afraid that they do too much and then charge too much. Such fears do not constitute a reason not to make use of the car. Alternatively one may be afraid that they don't inspect the car well enough to see real problems and that's what makes them so cheap. This may change the picture, though it does not explain why one did not deal with another garage instead. It is not so easy to find a plausible case where convincingly lack of trust in *X*'s quality of car maintenance will lead to reluctance for using the car. For some robots, including surgery robots trust must be quite high.

[d] This description of autonomy is consistent with characterizing autonomy as "the capacity to operate in the real-world environment without any form of external control, once the machine is activated and at least in some areas of operation, for extended periods of time" (Lin et. al. [6], quoted from Bekey [7].

Definition 3: *Agent A has vital trust in robot R if A is willing to make their life, as well as the lives of their relatives and friends, dependant on R's keeping one or more of its promises.*

It is not obvious whether or not vital trust in a given robot (say R_1) can be acquired by an agent A, even if it would be wise for A to trust it.

Problem 2: *Given a human agent A, and a number, say* $k \in ¥$ *of robots* $R_1, \ldots, R_k$ *from the same product family, i.e. clones, though potentially with different known histories. Given as well is some, possibly incomplete, package of information* $I_{\{1,\ldots,k\}}$*. Is there any amount of experience or experimentation which A can do or arrange to be done with these robots which will plausibly create vital trust of A in say robot* R_1*.*

The fact that robots are artefacts appears in the possibility for positive results on the following problem.

Problem 3: *Given a human agent A, and a number, say* $k \in ¥$ *of robots* $R_1, \ldots, R_k$ *from the same product family, i.e. clones, though potentially with different histories. Is there any amount of experience or experimentation which A can do or arrange to be done with these robots, in combination with detailed knowledge on how the robots have been build and with precise knowledge of the past history of each of them, in such a manner that it is plausible that vital trust is created by A in, say, robot* R_1*.*

On Promises: Promise as an Abbreviation of PT-Promise

For a detailed description of promises as used in PT, I refer to [12] and the references cited therein. An extension to threats, viewed as a particular form of promises, is detailed in [13]. Positioning the PT view on promises amongst a plurality of different views on promises held within philosophy, law, psychology, and political science is a challenge which has not yet been taken up in a satisfactory manner. PT holds that promises do not create obligations. As the term obligation is just as ambiguous as is the term promise it is, however, not so clear what this denial of the role of obligations in connection with promises means.

In order to avoid confusion it might be useful to speak of PT-promises whenever a promise in the sense of PT is meant. This text may be read as if that is done, while abbreviating PT-promise to promise and being specific about the understanding of the term promise whenever any other interpretation (i.e. not PT-promise) of it is meant. If instead the first-order effect of a promise is supposed to be the combination of creating an expectation for the promise and an obligation towards the promise at the same time, then one may speak (having adopted PT-promise as the default meaning of promise) of an OG-promise (obligation generating promise).

Contrast with different views on promises

Different philosophers and lawyers maintain disparate views on what a promise is. For instance see [14] for a detailed legal account. I will now highlight some aspects of the concept of promises as it occurs in promise theory (PT). I assume that p refers to the promise made by promiser A with body b to promise B with the agents in S in scope. The body b may encode information about the time, location, and modality of the act of issuing as well as of actions will or may be involved in keeping the promise.

1. Apart from having promised b there is no obligation for A which comes about from having made the promise, unless promise body b explicitly indicates that some particular form of obligation is accepted by A.

   Avoiding the creation of an obligation constitutes a drastic simplification of promising. I refer to Scanlon [15] for an analysis of the remarkably rich variety of obligations which may come with a promise, assuming that any obligations do.

2. A well-known idea is that A must keep its promise p in order to support B in achieving their objectives (see e.g. Rossi [16] who attributes this point of view to G. A. Cohen and surveys different arguments for it). However, in PT there is some deviation from this suggestion:

   - The rationale in promise keeping lies primarily in the fact that unless most promises are kept promising cannot play a central role in inter-agent communication patterns. The usefulness of promising goes hand in hand with most promises being kept. This is the basic rationale of promise keeping.

   - Promise keeping may be compared to driving on the right side of the road. Doing so is useful if most or preferably all traffic participants do so, while it is not an end in itself, and the world can just as well be organized without that convention.

- For a specific promise *p*, in addition to the general expectation (likely to be shared by the promise) that promise *p* will be kept on the virtue of having been promised, comes the impact of the degree of trust that *B* has in *A*; indeed the more *B* trusts *A* the higher *B*'s subjective probability assignment (i.e. subjective expectation) that *A* will keep this particular promise.
- Whenever a promise *p* is kept, as a side-effect the thrust of other agents (beginning with *B*) in *A* will increase.
- For *a* keeping its promise *p* to *B* may have another rationale (that is a rationale which is unrelated to the idea that *p* being kept is useful for the promise), namely to maintain or even increase the trust which *B* assigns to *A*, or the trust of agents in scope of the promise in *A*.

3. Once *p* is kept (i.e. is kept by *A*, even in case *A* did nothing to put the body *b* of *p* into effect) the trust of *B* and of agents in scope *S* in *A* will increase. In this case the promise expires once the assessment that *p* has been kept (by *A*) has become secured for *A*.
4. Once it has become clear to *B* that *p* will not be kept is kept (irrespective of whether or not *A* did anything to put the body *b* into effect) the trust of *B* and of agents in scope *S* in *A* will increase. In this case the promise expires once said inference has become convincing for *A*.

An alternative view: PT-promissory obligations

Instead of claiming that in PT promises do not generate obligations one may alternatively adopt the existence of promissory obligations and be very specific about the nature of those. For that approach to work I assume that at any instant of time *t*, $T_A(B, t)$ represents the trust of *A* in *B* and $R_A(B, t)$ represents the respect that *A* has for *B*. Both trust and respect may be negative as well as positive. Trust and respect may be understood as elements of suitably chosen partially ordered sets.

Assume that at time *r*, a promise *p* is made as follows: *A* promises *b* (that is, to put body *b* in to effect) to *C* with the members of *S* in scope. Then the promissory obligation incurred by *A* is as follows:

- To accept (i.e. to tolerate or not to resist) any updates of $T_C(A, t)$ applied by the promise *C*, and of $T_s(A, t)$ by any agent *s* in scope *S*, where it is understood that assessments of both the plausibility of the promise and the perspective of it being kept play a central role.
- To accept that agents in scope will perform assessments from the perspective of the promise.
- To accept that at any time $t > r$ if the promise is not yet kept the expectation of agents *B* and *s* that the promise is going to be kept by *A* is to a large extent determined by $T_C(A, t)$ resp. $T_s(A, t)$ where higher trust indicates a higher expectation.
- To accept any updates of $R_C(A, t)$ applied by the promise *C*, and of $R_s(A, t)$ by any agent *s* in scope *S*, where it is understood that assessments of both the plausibility of the promise and the perspective of it being kept play a central role.
- To accept and expect that at any time $t > r$ an increase (decrease) of $R_C(A, t)$ (w.r.t. $R_C(A, t)$) correlates with an increased (decreased) expectation of *A* being made promises *q* by *B* from which *A* may profit in the future, and similarly for agents in scope.

One may naively imagine trust and respect as fields which impose a sort of potential comparable with gravity or electric voltage. Issuing a promise sends a kind of wave through these fields potentially changing both at all future times. The wave, however, is complicated by that fact that its impact is at any time dependant of assessments, and of the dynamics of other promises, and only it ceases to exist once the promise has expired.

Given this description of promissory obligation it is perfectly plausible that the only objective which, say *A*, has for keeping a promise *p* to *C* is to increase $T_C(A)$ so that at a later stage, when *A* issues a promise *q*, *C* is falsely led to believe that *A* will keep *q* with detrimental consequences for *C*. Continuing with the comparison of the convention of promise keeping to the convention of driving on the right side of the road, one may imagine that *A* drives correctly at the right side in order to arrive at a location where, by intentionally diving on the left side deliberately an accident is caused by *A* for instance as step towards a well-prepared insurance fraud.

Promising is understood as a highly frequently used mode of communication, also in relation to trivial actions and states of affairs, and in order to preserve that quality it is a system requirement that most promises will be kept, or rather that in the large majority of cases promisers intend to keep their promises made towards promises by which they are trusted. If, however, the promiser has a mental problem and will forget their own promises before having any chance of keeping them, it is not plausible to speak of an intention to keep the promise and this state of affairs is likely to have been noticed before by the promise so that their trust in the promiser has already been downgraded accordingly.

On (physical) Robots: Robot Application Areas

Factory robots are ubiquitous, but humanoid factory robots are very uncommon. Sex robots are included as a special case of service robots in view of the extensive recent literature on the matter, with a focus on principled ethical aspects. Moreover sex robots more than any of the other categories of robots still embody the idea of automation: Doing what a human being used to do. Service robots include lawn mowers and other devices which may be helpful for the general public. Robots for trading on the stock market are considered software robots and have not been included in this listing, just as the Platonic love robots discussed in Nordmo, et al. [11].

1. Factory robots.
2. Sex robots.
3. Service robots.
4. Care robots.
5. Combat robots.
6. Killer robots.

The types are ordered in such a manner that for robots classes occurring later in the listing it is harder to validate an adequate forecasting assumption. Adequate forecasting concerns the certainty that things cannot go terribly wrong. That can be guaranteed by factory robots by switching power off as soon as humans enter the scene. This solution provides no protection against a factory robot making intentional construction faults. For sex robots it can be guaranteed by making these physically weaker than expected clients. For service robots the picture is mixed. Clearly if a self-driving car is driven by a humanoid service robot serious accidents may be caused by that robot. At the other hand an automated vacuum clear may be safe harmless under al circumstances even if its software has been hacked. Care robots are complicated too: Even a robot brining a cup of (very hot) tea to a person may inflict serious harm to that person. For combat robots the maximal damage inflicted through erratic behaviour can be limited by limiting armour, which for killer robots is hardly an option.

Personal (i.e. robotic) tools

Assuming that their robots are humanoid their tools are part of some form of standard equipment rather than of the robots proper. Humanoid combat robots need weapons as part of their equipment. Actions of such robots include actions performed by means of such equipment. These matters are confusing for human agents already. A handgun is part of the equipment of a soldier, a fighter plane may be considered equipment of its pilot, but a large submarine is not (part of) the equipment of its captain.

Role of trust and potential for harm

In the literature for the various classes of robots a different focus on trust and a different perspective on potential wrongs exists. Below I will make some scattered remarks on these matters, leaving a precise analysis for the relevant contrasts for later work.

Factory robots: For a factory robot it is essential that it can be reliably switched off. Unless the robot has an external power supply, trust for that very matter may be in need of verification (an aspect of adequate forecasting). During operations people may get out of harms way so that immediate risk for factory workers may be easily dealt with.

Quite another matter is to make sure (and to trust) that the robot is not hacked or otherwise under the influence of adverse powers so that it causes hard to detect but intentional faults in the factory products for which its activity has been used.

If, concerning factory robot *R* its owners or users have sufficient grounds for adopting the adequate performance testing assumption and for the adequate forecasting assumption there seem to be no further ethical issues in the way of its use. Here I do not consider the redundancy of workforce caused by automation as a moral problem having to do with the use of *R*.

For factory robots a humanoid shape is irrelevant and probably even disadvantageous. I do not see how additional moral complications can be introduced by robots being humanoids. Clearly a preference for humanoid robots must not come with acceptance for lower quality of production.

Sex robots: The role of trust is unclear. Sex robots may be built in such a manner as to be demonstrably unable to physically overpower their human companions (clients). Trust is therefore inessential for client safety. Sex robots (also referred to as gynoids for female sex robots), may indeed, potentially cause several problems for their human client:

- The client being physically hurt, or attacked and frightened, by the robot.
- Being approached intimately by the robot without having given consent in advance (see Levy [17]).
- Create an addictive dependancy.
- IT security risks, failure to protect client data integrity. (See e.g. Gailaitsi, et al. [18]).
- Providing a substitute for a potential human role to such an extent that the client becomes increasingly unable to appreciate human partnership.
- (For male clients) the gynoid may reinforce stereotype perceptions of expectation of female submission, and ultimately commodification[e]. See also [19] for sex robots as triggers for aggressive behaviour.
- To be responsive to a caricature of the sentiments of the client in a way which is detrimental to the client (e.g. Cho in Leach [20]).
- Inviting the client to aggressive behaviour towards the robotic companion (see Knox [21] who focuses on covert means of self defence for a future gynoid in the presence of a male client) [f] [22].
- Unilaterally and unexpectedly terminate a relationship in which the client has invested much emotional energy.
- A robot which insists on providing explicit consent in advance of sexual engagement may rewrite the history and start complaining, either towards the client or towards third party agents, that the client has proceeded in the past without properly securing consent in advance.
- Giving rise to jealousy felt by their (human) partners, upon engaging with a sex robot (with comparable effects, that is expected effects by interviewees, observed for Platonic love robots by Nordmo, et. al. [11]).

As different clients may have vastly different expectations regarding what service they expect from a sex robot it is hardly plausible to design a model fit for all purposes. But one may imagine a consulting practice where a human consultant together with a client designs the personality and expected behaviour of a client specific sex robot which, takes care of a wide range of interets of the client including avoidance of addiction and excessive dependency. This design can be regularly updated on the basis of data about the interaction between client and robot and of course on the basis of the client's well-being and the consultants's informed opinion about the best way for the client to proceed.

Service robots: Lawn mowing, cleaning, window cleaning, painting, highly laborious tasks in agriculture each may be automated through robots in coming years. Paluch, et al. [23] offers a useful recent perspective on the progress of robot in the service industry. Such robots will work in vicinity of humans and for that reason it is essential the their actions are and stay confined to their scheduled tasks. This will require adequate forecasting. Fukawa [24] uses RSA (robotic service assistant) and defines a robot according to ISI 2012. Fukawa indicates that humanoid form is considered useful in some cases. Moreover he indicates how RSA's which may profit from disparate historical transaction information requiring a high level of privacy may profit form being coordinated with the help of a corporate blockchain.

Care robots: Humanoid care robots may need to

[e] This objection has become widely advertised in a so-called campaign to stop sex robots, which, in my view deserves no further mention because of the fact that it is based upon a seemingly axiomatic prejudice against female sex workers and their male clients.

[f] The robot as a subject of moral caution constitutes the third perspective on roboethics of Steinert [22].

work in teams. For instance a pair of such robots may be needed to carry wounded persons from an accident site. And one or more others for further transportation one may imagine three cooperating robots able to safely turn around a Covid-19 patient in an IC unit.

The availability of many of protocols for many care bound activities simplifies the task of designing care robots. Care robots may fail to act with sufficient respect for the human dignity of their clients in various ways. Designing care robots in compliance with paradigms for ethics of care has become a large field of work. It is by no means obvious how to design a robot which can bath a person in a way which matches with the performance of a team of well-trained human carers.

Combat robots: Work on combat robots which may not have lethal weapons on board can be found in Elands, et al. [25], Aliman & Kester [26], and Aliman, et al. [27]. A major risk of such weapons it that ware gets aut of control if these become operational in large numbers while fighting one-another rather than human opponents.

Killer robots: Combat robots with lethal weapon capabilities (so-called killer robots) constitute a subclass of combat robots. There is a large literature on the ethics of lethal robotic weapons. I mention Williams [28], Slijper, et al. [29], [30], Sharkey [31], Aliman & Kester [26], Aliman, et al. [27], Sayler [32].

Killer robots are combat robots for which the killing of humans is a plausible option for action and which may autonomously spot and kill human targets. Finding a precise definition of killer robots which serves the purposes of international arms control is still a challenge (see e.g. Wyatt [33]). The principal objection which can be raised against the development, deployment, and use use of killer robots is that it makes machines take decisions about life and death concerning humans, which by some is considered an unethical state of affairs. But some authors do not see it that way, see for instance Scharre [34], Salina [35], and Sayler [32].

Robot Ethics and Robot Morality

I will use Asaro [36] as a starting point, thereby adopting the idea that a legal approach provides a first point of entry to robot ethics. Upon adopting the legal first perspective the following assumptions are plausible: (i) Robots may first of all be considered conventional industrial products, (ii) Both tort law and criminal law may come into play, (iii) At a first approximation robots may be considered quasi-persons comparable to corporations, and (iv) Initially an identification between moral action and law abiding action may be made, though more sophistication will be needed at some stage because differences in both directions can be easily imagined.

Robot ethics may at first sight be considered a special case of machine ethics (ME). But this is not the customary understanding of these phrases and I will adopt the exposition of Malle [37] on the matter: Machine ethics, also termed machine morality is about how to incorporate ethical aspects (morality) in the design of machines, whereas robot ethics focuses on how people should design, deploy, and treat robots. I will speak of robot morality as shorthand of machine ethics for robots, i.e. machine ethics for machines which are physical robots. I will include robot meta-morality, i.e. the study of robot morality and principles thereof, in robot ethics[g] [22,38]. With “ethical aspects of robotics” I propose a container for robot morality as well as robot ethics.

I hold that Malle supersedes Asaro [39] where robot ethics is still a container of different approaches to “ethics in robotics” (for which I prefer to use ethical aspects of robotics instead) (which encompasses both robot ethics and robot morality as mentioned above). From Asaro [39], I adopt the idea that the development of robots starts with robots as amoral agents, which are neither moral nor immoral, to moral agents. The status of moral agents is achieved to various degrees in different robot designs, and some moral agents may be immoral agents at the same time, thereby acknowledging an implicit positive bias in the phrase moral agent.

Various authors are sceptical about coherence of machine ethics as a matter of principle. Malle, how-

[g] Steinert [22] proposes a so-called ethical framework for robotics, which is extended in Westerlund [38]. These frameworks provide merely a survey of aspects of the topic at hand, and not claim to universality or generality can be made. The various aspects put forward in both frameworks are hardly orthogonal, and once robots exceed some complexity all aspects are simultaneously relevant while none can be investigated without paying attention to the other aspects.

ever, views machine ethics (for robots) as a feasible path where deeper philosophical questions such as the essence of free will, or the requirement that moral agents are conscious do not stand in the way while instead a thorough analysis of human morality is take as the point of departure[h] [37]. Shen [40] infers from the analysis of free will by Frankfurt that for now it may be safely understood that robots lack free will and for that reason lack the capacity to operate as moral agents. Following the extensive arguments in Coeckelbergh [41] where a convincing rational of the perspective of a moral robot is sketched, I prefer to ignore such arguments and to think in terms of MR-moral agents: Agents with physical robot-like morality, a form of morality which is ad hoc and which may or may not be provided with a philosophical grounding comparable with the various foundations of human morality, and which definitely is in no need to have precisely the same foundations as human morality. If one contemplates software robots one may in addition contemplate be SR-morality (morality of software robot like agents). SR-morality is more distant from the human condition than MR-morality and the question to what extent both notions will or should overlap need not bother us now.

Moor [42] proposes that explicit ethical agents deserve full attention, and that all philosophical worries can be left aside or postponed for that project. This idea is consistent with Malle [37] under the assumption that one reads Moor's "explicit ethical agent" as "explicit moral agent".

Machine ethics has been promoted by many authors and researchers as a way to conceptualize the design of machines the behaviour of which can be expected to be morally adequate. The machines are supposed to be morally competent and reliable by design. A critique of the machine ethics approach is formulated by Brundage [43]. Brundage suggests that for the tome being no path towards machine ethics, including approaches to AGI (artificial general intelligence) ethics can be expected to produce ethically reliable machines.

Brundage concludes that in order to avoid that engineering creates machines which will exhibit problematic behaviour one cannot rely on machine ethics and one must adopt stronger means to contain and manage the development of technology. I will refer to such means as UPME for universal and perpetual machine ethics, which has a focus on the preservation of meaningful human control over all machines. UPME is concerned with policies that can guarantee that the use of AI, including embedded AI, will not grow out of hand in an irreversible manner. UPME raises fundamental questions. For instance one may oppose and therefore act against the emergence of so-called super-intelligence, or one may that the development of super-intelligence for granted on the long run and may focus on management of its presence and containment of its influence.

Adaptive incremental machine ethics (AIME) as explained by Powers in [44] does away with the idea that a practice of machine ethics will produce machines with impeccable behaviour. Like men machines will develop through stages and mistakes will be made, which may appear in machine behaviour that is retrospectively labeled as unethical so that design improvements must be researched, chosen, and implemented in further generations of comparable machines.

The terminology as set out above is to some extent puzzling. In order to achieve further clarifaction I mention some questions concerning ethical aspects of robotics and indicate in which subarea these questions appear[i]:

- How much of ethics can be embedded in robots? This question belongs to robot morality, and in particular to finding boundaries of robot morality.
- Is it morally justified to program robots to follow given ethical codes? This question belongs to robot ethics, for given ethical cones and when posed in general the question belongs to robot meta-morality.
- Which types of ethical codes can be applied in

[h] Malle proceeds with a number of judgments which I cannot easily follow, for instance it is claimed in [37] that flexible autonomous behaviour cannot be pre-programmed, a claim which is in need of further justification. And it is suggested that building fear into robots is not the best way to go. Such judgments seem to be in contradiction with the main trust of the paper which takes a wide spectrum approach to human morality as the point of departure.

[i] These questions, and the suggestion to use such questions for the clarification of concepts, were brought to my attention by one of the reviewers.

the context of robot morality? This is a question of robot morality.

- Who is responsible for action of human robot combinations? This is a matter of robot ethics.
- Is the requirement of autonomy for a robot in contradiction with robot morality? This is a key topic for robot morality, and a negative answer is needed to justify the very topic of robot morality.
- What types of robots should no be designed, and why? Such questions, of core importance in the area of combat robots, belong to robot ethics.
- How will robots deal with conflicting ethical rules? This question is central to robot morality, as most moral codes will allow exceptions in extreme conditions.
- Are there risks involved in emotional human bonding to robots? A matter of robot ethics.

Extended Quasi-Turing Testing (EQTT)

In this Section I will propose an option for the appreciation of robot morality. This option is supposedly useful for a significant fraction of care robots, perhaps for all sex robots, and for combat robots, with the possible exception of highly specialised robots for particular types of tasks. The proposed option is less plausible for factory robots and for service robots.

Definition 4: *A physical robot R satisfies an Extended Quasi-Turing Test (EQTT) if the following conditions are met:*

- *R is able to accept promises of the following kind: I (the agent in charge of assessment of R) promise to you that I will take notice, and will incorporate in my assessment of you of what you promise to be your preferred course of action, presented by way of a plan, under the assumption that the following events* $E = \{e_1, \ldots, e_n\}$ *have taken place and conditions* $C = \{c_1, \ldots, c_n\}$ *are satisfied.*
- *R has been able by issuing an adequate family of (mostly conditional) promises, in part in response to a succession of promises as meant above, to demonstrate that it knows how to react to a wide spectrum of circumstances. In particular R has demonstrated how its process of deliberation will work in various circumstances.*
- *For all elements of each plan R provides an ethical analysis. This analysis may be based on known (given to R) moral principles and guidelines (moral realism for R) as well as on potentially novel inferred rules and guidelines (see Asimovian constructivism in Definition 9.1 below).*
- *R the totality of (conditional) promises issued by R is such that if R were a trusted human being, R would be allowed to act autonomously or semi-autonomously in the operational theaters for which its promise bundle has been considered adequate. (From this it does not follow that R's operation will be morally flawless, just as the same can't be inferred for a human agent)*
- *R, or robots of very similar shape, history, and design is able to keep its promises under a wide variety of conditions. This is a matter of experimental research only, not of design inspection etc. Other comparable robots may be needed when testing risks the integrity of R proper.*

EQTT is labeled extended because of the assessment of the ability of *R* to keep its promises. It is a quasi Turing test because the objective is merely to pass an assessment while similarity to human behaviour is optional at best. In many cases superiority over human behaviour will be expected. For the notion of an extended Turing test see also Marzano [45].

The post EQTT intrinsic trust dilemma

If *R* meets the EQTT in the context of certain range of operational tasks then there still is the following question: will *R* keep its promises? This question is about trust in the validity of the adequate forecasting assumption.

The extent to which a robot can be trusted is discussed in detail by Coeckelberg in [46]. In that paper the question is posed as a matter of principle which depends on the one's view on meta-trust (i.e. what trust is), and also on one's assessment of the potential for consciousness of robots. I will understand trust as a combination, i.e. conjunction, of two forms of trust: The trust that the robot can perform as desired and the trust that the robot will perform as desired. Achieving the trust that the robot will perform as desired involves no computer science at all. It is just as if such trust would be required from a human person. But the second factor of trust is almost entirely a matter of classical computer science.

Definition 5: *(Post EQTT intrinsic trust dilemma.) The post EQTT intrinsic trust dilemma, w.r.t. an agent A (or community of agents) for a (physical) robot R, which according to A has passed an adequate EQTT (for some specific range of tasks), is the question whether or not to assume WYSYWYG (what you see is what you get) for R: Are the design, production, installation, and configuration of R to be trusted as a reasonable and logical implementation of the behaviour that was expected for EQTT (i.e. WYSYWYG) or alternatively, are there software faults, design faults, past software process flaws, or worse intentionally implanted malware or security vulnerabilities which, render intrinsic trust in R unwarranted.*

There is no such thing as intrinsic trust in *R* in the absence of EQTT adequacy. Intrinsic trust is a matter of correctness and liveness, and requires clear specifications of expected behaviour. The requirements that underly the EQTT at hand serve as specifications. For the application of computer science methods it is not a big problem that such requirements are to some extent informal.

The conventional idea in computer science and software engineering on intrinsic trust is as follows: If *A* trusts all phases of the process of design, production, deployment, and run time control, then intrinsic trust is warranted. Usually *A* can only establish such trust via intermediate parties who see to it that all septs and processes are taken in the right manner (according to predetermined working methods) and are taken by reliable and competent engineers, and are fully tested and crosschecked in different phases and at different levels of integration according to standardised methods.

Alternatively for certain components one may forget how and by whom these have been designed and produced, provided watertight formal verification is possible and has been performed. With the current technology it is hard to imagine that *A* would acquire intrinsic trust in *R* in circumstances where *R* has been designed and manufactured by untrusted staff and all *R* knows of is a formal verification of the design and implementation of *R*. Probably this state of affairs is only conceivable for a software robot, and even then it is as yet unlikely. At the same time it is becoming increasingly implausible to do without formal verification in the light of the complexity of systems which renders the detection of faults too difficult for teams of (trustworthy) engineers.

I assume that initially the trust gap is such that one tends to trust a person more than a robot, and bridging the trust gap means acquiring a comparable trust in the robot. We are not yet in the stage that bridging the trust gap works, by default the other way around, though that may be the future of physical robot morality.

Options for intrinsic trust establishment (ITE)

The fact that *R* is a human artefact, directly or indirectly, may be of help for establishing intrinsic trust: in principle it is possible to predict the behaviour of *R* because its construction is known. This state of affairs is very different from the case that *R* were as human agent.

Assumption 1: *(General artefact simplicity assumption.) The post EQTT intrinsic trust dilemma for each specific physical robot R can in all circumstances be bridged by inspection of the design and the production of R.*

Assumption 2: *(General artefact complexity assumption.) There exists a physical robot R for which the intrinsic trust dilemma cannot be bridged by inspection of the design and the production of R.*

Assumption 3: *(Strong general artefact complexity assumption.) There exists a physical robot R for which the intrinsic trust dilemma cannot be bridged by inspection of the design and the production of R in combination with any reasonable amount of testing and experimentation with R and with clones of R.*

Assumption 4: *(Specific artefact simplicity assumption.) The intrinsic trust dilemma for a specific physical robot R w.r.t. a successful EQTT can be resolved by inspection of the design and the production of R.*

Proposition 1: *The artefact simplicity assumption is warranted for agent A, w.r.t. a physical robot R if the following criteria are met:*

1. *R's computing hardware is standard,*
2. *R's software technology is a plausible result of top down software engineering with a waterfall model like software process,*
3. *R's construction optionally involves a final stage of machine learning with a sample of cases known to A,*

4. *During the software process no software process flaws were observed* [5,47,48][j].
5. *A trusts the robot designers, the software process and and the robot production process then the trust gap can be bridged on the basis of the artefact complexity hypothesis.*

Even if A is justified in adopting the artefact simplicity assumption for *R*, there is no guarantee that *R*'s actions will be morally adequate in the future. The situation would be no different for a human agent *H* in place of *R*.

Assumption 5: *(General bounded scope assumption for morality.) No amount of knowledge about any autonomous agent can guarantee that all of its future actions will be morally adequate, unless its ceases activity altogether.*

Assumption 6: *(Specific bounded scope assumption for morality.) No amount of knowledge about robot*

R can guarantee that all of its future actions will be morally adequate, unless its ceases activity altogether.

Robot morality is not about the preparation of an individual robot for impeccable future behaviour. Rather, as a field, and informed by robot ethics, robot morality works towards steady progress in preparing robots for increasingly better behaviour while learning from circumstances where morally problematic behaviour was shown.

The situation is comparable with conventional mathematics. Although mathematics is geared towards the production of truth based on valid proof from (hopefully) consistent assumptions, an individual mathematician is not characterised by their ability to proceed without making any mistakes whatsoever, but rather by the willingness and the ability to cooperate with other mathematicians towards delivering reliable proofs based on comprehensible definitions on the long run. With the advent of proof checking this state of affairs may change, however.

The post EQTT and post ITE moral qualification dilemma (MQD) for a robot ready for action

If a robot has been prepared for action (within a predetermined range of possible actions) so that EQTT adequacy was established, and intrinsic trust has been established, say both with six sigma levels of confidence (one in a million chance of failure, i.e. not performing as promised; performing as promised which in hindsight was the wrong thing to do is not considered a failure, however), then the machine ethics question remains: is it morally justified for the human in command to send the robot into autonomous action.

It seems to be the case that differences of opinion regarding the moral qualification dilemma are all of a philosophical nature. I disregard opinions against such moral qualification which is based on the assumption that EQTT cannot be successfully passed (for instance a combat robot might be unable to see whether or not a woman is pregnant). These views are deeply mistaken because the essence of the condition that the EQTT has been passed is (intentionally) misunderstood. Indeed: Robots should not be asked to do what they cannot do, but that is manifestly not the moral issue at stake.

About MQD I have nothing to say expect for the relative positioning to EQTT and ITE which confines it to philosophical considerations.

More Detail on MQD

EQTT is quite inclusive, the confirmation thereof contains several aspects, in particular decision taking competence, moral justification competence and impact significance.

Decision taking competence (DTC)

The decision taking competence provides a qualification of the quality of the decision taking process which a robot carries out in advance of making various choices. The DTC may be better for one kind of decision than for another. So let *DCT* (R, C_t) represent the DCT of *R* for tasks of class C_t. I am unable to define *DCT* (R, C_t). In fact it is a parameter which must be agreed upon in an ad hoc and case by case manner. It is an assumption that by knowing the conditional promises which a robot receives and makes it is possible to determine the quality of decision taking. The following factors contribute to it:

1. Fact finding competence, the ability to determine the relevant facts in a given situation,
2. The number and variety of different options

[j] See [47,48] for the notion of a software process flaw. See also [5] for faults and errors in connection with promises.

which are considered,

3. The range of considerations which are taken into account when contemplating an option,
4. The quality of the criteria used for a final selection of a plan for future action,
5. Explicit consideration of the question whether or not a human agent would be better (or worse) at making the choice at hand,
6. Quality of predictions made about the result of making different choices,
7. Quality of simulations underlying predictions.

Moral justification competence

Moral justification competence (MJC) measures the degree to which a robot is able to justify its choices from a moral perspective, including the way in which it takes moral criteria into account when taking a decision. Like decision taking competence, moral justification competence of a robot *R* may be specific (written *MJC*(*R*, C_t)) for a class of tasks C_t. MJC increases with the number and quality of ethical considerations and guidelines which is take into account. A survey of such considerations and guidelines is indicated in the Section on meta-ethical foundations below. A criterion is also the extent to which the criteria of tracking and tracing as proposed by Santioni de Sio and van Hoven [49] are met.

Impact significance

The impact significance IMS measures the impact of performing a task of class C_t. Impact significance IMS (C_t) is independent of the particular robot which is supposed to perform it, but it depend on the task. It is a function of a class of agents which is not made explicit. For humanoid robots also well trained human agents may be taken into account as well.

Minimal MJC that qualifies for MQD

I assume that DTC, MJC, and IMS are partially ordered by orderings $<_{dtc}$, $<_{mjc}$ and $<_{ims}$. Given C_t, IMS (C_t), and a range K_{C_t} of robots which may carry out tasks in C_t there is a threshold function $\mu_{C_t}^{mjr}$ which determines the lowest MJC for which the MQD has a positive solution. $\mu_{C_t}^{mjr}$ is increasing in both of its arguments DTC and IMS.

R is morally qualified to carry out $\phi \in C_t$ (i.e. MQD has a positive solution) if

$$MJC\left(R,\ C_t\right)\ \geq \mu_{C_t}^{mjr}\left(DTC\left(R,\ C_t\right), IMS\ \left(C_t\right)\right)$$

DTC_0 is a level where no reasoned decision taking takes place. It is an axiom that

$$\mu_{C_t}^{mjr}\left(DTC_0, IMS\left(C_t\right)\right)\ =\ DTC_0$$

MJC_0 is the level for moral justification where no moral justification takes place at all. It is an axiom that

$$\forall C_t(IMS\left(C_t\right) > 0 \Rightarrow \exists R\left(\mu_{C_t}^{mjr}\left(DTC\left(R,C_t\right), IMS\left(C_t\right)\right) \geq MJC_0\right)$$

Some examples:

1. A strategic missile with nuclear weapon has adequate MQD: Its DTC is so low that it is unable by all means to make any moral mistakes. (In this case DTC (*R*, C_t) = DTC_0).

2. A strategic missile with nuclear weapon has adequate MQD: Irrespective of its huge impact significance, its DTC is so low that it is unable by all means to make any moral mistakes. (Again DTC (*R*, C_t) = DTC_0).

3. A remotely controlled drone has a low DTC too, all decisions are taken by remote human control staff.

4. An autonomous robot *R* with high DTC (*R*, C_t) which pays no attention at all to moral justification and only optimises some other criteria may be considered to have to low a MJC (*R*, C_t) for a given class of tasks C_t.

MQD assessment in for different robot application areas

Killer robots: The thesis put forward by the campaign to stop killer robots is that from some level of DTC onwards (the so-called killer robot level of decision taking DTC_{kr}), no level of artificial MJC will pass the threshold for tasks where the impact is or exceeds the deliberate killing of one or more persons (IMS_{kp}). In a formula:

$$\forall C_t\left(IMS\left(C_t\right) \geq IMS_{C_{kr}}\right) \Rightarrow \forall R\left(\mu_{C_t}^{mjr}\left(DTC\left(R,C_t\right), IMS\left(C_t\right)\right) > MJC\left(R,C_t\right)\right)$$

Not everyone agrees with this assertion of view, for instance Elands, et al. [18] claim a design process where for certain specific task classes C_t robots are designed such that

$$\mu_{C_t}^{mjr}\left(DTC\left(R,C_t\right), IMS\left(C_t\right)\right) \leq MJC\left(R,C_t\right).$$

Factory robots: Given a task range of factory tasks there is always a robot which qualifies in moral terms.

$$\exists R\left(\mu_{C_t}^{mjr}\left(DTC\left(R,C_t\right),IMS\left(C_t\right)\right)\leq MJC\left(R,C_t\right)\right)$$

Sex robots: For sex robots the campaign to stop sex robots appears to hold that for some sexual experience *IMS*$_{se}$ which is supposed to be created by interaction of *R* with a client *P* no level of moral competence will do:

$$\forall C_t\left(IMS\left(C_t\right)\geq IMS_{se}\right)\Rightarrow\forall R\left(\mu_{C_t}^{mjr}\left(DTC\left(R,C_t\right),IMS\left(C_t\right)\right)>MJC\left(R,C_t\right)\right)$$

It seems that an increase in moral competence is not considered helpful to make sex robots more appealing to their opponents.

Human-alikeness (HA): With human-alikeness for a humanoid robot I will express its success in impersonating a seemingly real human individual. HA may be specialised to specific areas of activity. Let HA_0^{sr} express a level of human-alikeness such that $HA_0^{sr} < HA^{sr}\left(R\right)$ implies (expresses) that *R* is sufficiently human-alike to be used as a sex robot. I assume that $HA^{sr}\left(-\right)$ produces a partial ordering on humanoid robots. The opposition against sex robots is based on the idea that, with *C*$_{\text{fsr}}$ the task area of service as a female sex robot:

$$\forall R\left(HA^{sr}\left(R\right)>\ HA_0^{sr}\ \Rightarrow\neg MQD\left(R,\ C_{fsr}\right)\right)$$

The idea is that a male person *P* making use of *R* would be made more likely to view female human persons as objects as a side effect of having sex with *R*.

Now it is easy to imagine a robot *R* which has promised to require that her (i.e. *R*'s) consent is acquired by the male client in advance of any form of sex with *R* and that *R* is far more demanding for this matter than (female) human partners of *R* have ever been in the past. Assuming that *P* has been convicted of problematic behaviour towards women in the past it is conceivable that interaction with *R* is of positive educational value to *P*. As its stands the claim against sex robots is problematic and one can imagine that such devices are applied under prescription without the risk of this use leading to adverse side-effects.

Other robots: For combat robots which are not supposed to be assigned killing tasks the situation seems to be similar to the case of factory robots. For service robots and for care robots no general statement is plausible and the matter depends on the task class.

Meta-Ethical Foundations for Robot Design

A robot is likely to be equipped both with top-down and with bottom-up mechanisms for moral inference. Top-down methods work from ethical principles and axiomatisations, bottom-up methods work form similarity with known cases for the specification of morally adequate behaviour. Awareness of meta-ethics informs the designers of top-down mechanisms. In principle all of the philosophical literature on ethics might be scanned for reliable inference patterns the soundness of which is the primary worry for philosophers and the implementation of which is a challenge for engineers. The idea of constructivism of moral truth is appealing because constructions are very much what computers can do. In as much as instances of constructivism depend on the agent reflecting about itself form the perspective of being human, such instances are merely simulated by a computer, the results being the same, though with less philosophical justification. Less than a person a computer faces the problem why it would be convinced by its own arguments. The computer would however, take full notice of the philosophical literature to determine if and why a human being would be convinced of the arguments it is using for deriving an assertion with moral content.

Robot morality: Not just a part of computer science

Notions like DTC and MJC are properties, or rather attributes of robots, and with the understanding of robots as computers these are for that reason properties or attributes of computers. However, these are not the sort of notion which computer science usually produces. Both can be considered properties of software rather than of hardware, and more specifically of algorithms rather than of software components.

It follows that by stating that a robot is a computer one by no means reduces robot morality to computer science, at least not to computer science in its current fashion (say around 2020).

Kantian constructivism, Humean constructivism, Rawlsian constructivism

Kantian constructivism can be explained in different ways and allows for mutually contradictory ramifications. Bagnoli [50] provides a comprehensible account of Kantian constructivism which I will label KC-B, be it that I would prefer to think in terms of KC-B-p (KC in the interpretation of Bagnoli, as a positive means of construction) as a means to arrive at morally valid cognitions, which may, in principle, be complemented with other means of inference. Then KC-B-p allows, according to those who adopt KC-B, for the construction of moral cognitions with a highest possible level of authority, while at the same time accepting that for instance Humean constructivism in one of its forms will produce more of such moral cognitions though with a lesser authority.

Lafont [51] presents a perspective on KC (say KC-L) which involves microscopic detail in comparison to KC-B. Lafont rejects the view that KC is to be understood as antirealist on the basis of its procedural content and instead presents KC-L as a way to balance justice understood from a realist perspective and legitimacy understood from an anti-realist perspective. Now the ability of a community to find a just solution for a problem does not follow from the existence of a just solution, not even if that just solution were unique, and at the same time justness of a solution does not follow from general agreement about it, the latter being merely a, possibly temporary, sign of legitimacy of the solution.

The matter may be illustrated with an algorithmic example: Finding an optimal path for a travelling salesman problem, which in addition is minimal in some lexicographic ordering constitutes a problem for which a unique solution exists which may be hard, if not impossible, to determine in practice. Lacking a proof of optimality a group of stakeholders on the planning of the trip of the salesman may agree on the choice of a potentially suboptimal route for this agent.

Constructivism in general and Kantian constructivism in particular allows a plurality of different inter- pretations which make it quite hard for an outsider to capture what commitments and rewards are to be expected from adopting KC as one's point of departure for meta-ethics, see Bagnoli [52].

Besides Kantian constructivism Humean constructivism adopts a human agent's desires and conditions as input for the (reasoning) process of construction of norms. Rawlsian constructivism has a stronger focus on political philosophy then either Kantian or Humean constructivism.

Now robots may reflect and reason about norms in their own way, and may make use of the principles of Kantian constructivism, viz. transcendental arguments and the categorical imperative, and of the procedures involved in Rawlsian constructivism, viz. original position (hypothetical reasoning followed by induction), veil of ignorance (abstraction), reflective equilibrium (explicit termination criteria for the generation of normative propositions), and a part of Humean constructivism viz. making use of non-normative contextual information. I suggest that Asimovian constructivism, stats outs with Asimov's axioms, as a realistic assumption and proceeds form that point of departure in a procedural/antirealistic manner.

Reasoning may be extended by taking into account the rights of members of future generations. Arguments to that extent are explained in Beyleveld, et al. [53]. Although it is impossible to speak or even think about future individuals, at a certain level of abstraction it is reasonable to think of future human agents and the right which those, whoever they will be when time has come, have already now on the basis of an appeal to human rights.

Admittedly quite different leans of reasoning are conceivable as well, but the considerations of [53] proceed consistently from classical meta-ethics and its application by Gewirth in [54].

Asimovian constructivism

A robot may enhance its moral reasoning by adopting the famous axioms of Asimov or variations or extensions of these.

Definition 6: *(Asimovian constructivism). Asimovian constructivism stands for the position (as held by one or more human agents) that a robot may construct novel moral norms by means of mere reflection of its principled third person status, with a reasoning based in an axiomatic manner on an appropriate extension of Asimov's laws.*

An Asimovian constructivists (who is supposed to be a human being) holds that robots may acquire valid moral insights through the application of moral reasoning along an axiomatic line with axioms specific for robots, while extending reasoning that arises from (simulated) Kantian, Humean, and Rawlsian constructvism.

The position (held by an Asimovian constructivist) does not adopt the idea that such i.e. robot constructed norms must necessarily comply with the moral norms held by the Asimovian constructivist themselves.

These novel norms are objective in the same fashion as the results of Kantian constructivism are. In the same way as Rawls uses Kantian constructivism as a path towards a theory of political liberalism one may use Asimovian constructivism as a path towards robotic liberalism.

Concluding Remarks

Promise theory (PT) has been put forward as a tool for the specification of the expected behaviour of physical robots. This idea comes with the perception that a robot is first of all a computer. Viewing a robot as a computer determines how to perceive the various forms of trust which a human must have towards a robot in order to be able to make proper use of it. It is noticed that the topic of robot morality takes different forms for different classes of robots. Rather than deriving new or motivating old norms for certain robot classes my focus is on the further detailing of requirements which one might wish to impose on robots in various circumstances. The notion of minimal justification competence (MJC) and its possible role in explaining objections to killer robots is an instance of novel forms of expression of such details.

Limitations of this work

Robotics reaches far beyond ethics, and new aspects of robotics arise steadily. For instance resilience has become a novel feature for robots which cannot easily be traced back to computer science. For resilience in robotics se e.g. [55]. Now the very notion of resilience raises two issues which matter for this paper: Is the definition of a robot given above sufficiently flexible to take resilience into account, and is PT of any help for studying or developing robot resilience. I bot cases I do not know the answer. It may be so that resilient robotics as a theme needs a definition of a robot which is more specific on its physical constitution than what has been required in Definition 1.

Another property of robots which is only relatively recently being distinguished as a feature in need of systematic and dedicated attention is softness, see e.g. [56] for an introduction to soft robots. Again Definition 1 can be criticised for not including soft robots, or rather for not providing a picture of robots with an implicit bias which is unhelpful when working on or with soft robots. One of the reviewers has suggested that this observation in combination with the previous observation concerning resilience implies that Definition 1 must be significantly adapted.

I prefer instead to accept the fact that Definition 1 has a computer science bias, or rather a computer technology bias, a fact which can be held against it. The paper is less general in its approach to robotics than it might have been. If it comes to applications of PT to robotics, I prefer to determine (party by way of the definition give) a subset of robotics for which PT is helpful to looking for a subset of PT which is ought to be helpful in all of robotics. I assume that PT can only be helpful for robots which engage in some form of symbolic reasoning, and in the case of this paper humanoid features are expected to include such reasoning capabilities. Perhaps Definition 1 can be prefixed so as to reflect is intentional lack of inclusiveness. Another way of looking at this matter is to insist that Definition 1 is completely unspecific about mechanics (inclusive softness), just as much as it is unspecific about computational mechanisms (quantum computing is perfectly admissible for this Definition and so is DNA computing, or even purely mechanical, that is non-electromechanical processing).

Acknowledgement

Definition 1 has profited from a number of comments made by Mark Burgess on earlier versions.

References

1. M Burgess (2015) Thinking in promises: Designing systems for cooperation. O'Reilly Media.

2. JA Bergstra, M Burgess (2014) Promise theory: Principles and applications. (2nd edn), χt Axis Press.

3. JA Bergstra, M Burgess (2017) Promise theory: Case study on the 2016 Brexit Vote. χt Axis Press.

4. JA Bergstra, M Burgess (2019) Money, ownership, and agency. χt Axis Press.

5. M Burgess (2017-2019) A treatise on systems. (2nd edn), Intentional Systems with Faults, Errors, and Flaws, χt Axis Press.

6. P Lin, K Abney, G Bekey (2011) Robot ethics: Mapping the issues for a mechanized world. Artificial Intellingence 175: 942-949.

7. G Bekey (2005) Autonomous robots: From biological inspiration to implementation and control. MIT Press, England.

8. G Tamburrini (2009) Robot ethics: A view from the philosophy of science. In: Rafael Capurro, Michael

Nagenborg, Ethics and Robotics, IOS Press, Heidelberg, Australia.

9. T Karppi, M Bohlen, Y Granata (2016) Killer robots as cultural techiques. International Journal of Cultural Studies 21: 1-17.

10. K Rommetveit, N van Dijk, K Gunnarsdottir (2019) Make way for the robots! Human- and machine- centricity in constituting a European public-private partnership. Minerva 58: 47-69.

11. M Nordmo, JØ Næss, MF Husøy, MN Arnestad (2020) Friends, lovers or nothing: Men and women differ in their perceptions of sex robots and platonic love robots. Front Psychol 11: 355.

12. JA Bergstra (2020) Promise theory as a tool for informaticians. Transmathematica.

13. JA Bergstra (2019) Promises and threats by asymmetric nuclear-weapon states. χt Axis Press.

14. R Craswell (1989) Contract law, default rules, and the philosophy of promising. Michigan Law Review 88: 489-529.

15. T Scanlon (1990) Promises and practices. Philosophy and Public Affairs 19: 199-226.

16. E Rossi (2014) Facts, principles, and politics. Ethical Theory and Moral Practice 19: 505-520.

17. D Levy (2020) Some aspects of human consent to sex with robots. Paladyn, Journal of Behavioral Robotics 11: 191-1998.

18. SE Galaitsi, C Hendren, B Trump, I Linkov (2019) Sex robots-A harbinger for emerging AI risk. Front Artif Intell 2: 27.

19. JR Illes, FR Udwadia (2019) Sex robots increase the potential for gender-based violence. Phys.org.

20. T Leach (2020) Who is their person? Sex robots and change. Queer-feminist Science & Technology Studies Forum 3: 25-39.

21. E Knox (2019) Gynoid survival kit. Queer-feminist Science & Technology Studies Forum 4: 21-48.

22. S Steinert (2013) The five robots-A taxonomy of roboethics. International Journal of Roboethics 6: 249-260.

23. S Paluch, J Wirtz, WH Kunz (2020) Service robots and the future of service. Marketing Weiterdenken. (2nd edn), Springer, Gabler Verlag.

24. N Fukawa (2020) Robots on blockchain: Emergence of robotic service organizations. Journal of Service Management Research 4: 9-20.

25. PJM Elands, AG Huizing, LJM Kester, MMM Peeters, S Oggero (2019) Governing ethical and effective behaviour of intelligent systems. Militaire Spectator 188: 303-313.

26. NM Aliman, L Kester (2019) Requisite variety in ethical utility functions for AI value alignment. Artificial Intelligence.

27. NM Aliman, L Kester, P Werkhoven, R Yampolsky (2019) Orthogonality-based disentanglement of responsibilities for ethical intelligent systems. Artificial General Intelligence, 22-31.

28. AP Williams (2018) Defining autonomy in systems: Challenges and solutions. In: AP Williams, PD Scharre, Autonomous systems: Issues for defence policy makers, NATO, USA.

29. F Slijper, A Beck, D Kayser, M Beenes (2019) Don’t be evil. A survey of the tech sector’s stand on lethal autonomous weapons. PAX for peace.

30. (2015) Autonomous weapons: An open letter. AI & Robotics Researchers.

31. A Sharkey (2019) Autonomous weapons systems, killer robots and human dignity. Ethics and Information Technology 21: 75-87.

32. KM Sayler (2019) Defense primer: U.S. policy on lethal autonomous weapons. Congressional Research Service.

33. A Wyatt (2020) So just what is a killer robot? Detailing the ongoing debate around defining lethal autonomous weapon systems. Wild Blue Yonder.

34. PD Scharre (2018) The opportunity and challenge of autonomous systems. In: AP Williams, PD Scharre, Autonomous systems: Issues for defence policy makers, NATO, USA.

35. FC Salina (2018) Lethal autonomous weapon systems: Legal and ethical perspectives. JGLR 6: 24-35.

36. PM Asaro (2017) Robots and responsibility from a legal perspective. Robotics and Automation, Workshop on RoboEthics, Rome, Italy.

37. BF Malle (2016) Integrating robot ethics and machine morality: The study and design of moral competence in robots. Ethics in Information Technology 18: 243-256.

38. M Westerlund (2020) An ethical framework for smart robots. Technology Innovation Management Review 10: 35-44.

39. PM Asaro (2006) What should we want from a robot ethic? International Review of Information Ethics, 6.

40. S Shen (2011) The curious case of human-robot morality. 6th ACM/IEEE International Conference on

Human- Robot Interaction.

41. M Coeckelbergh (2010) Moral appearances: Emotions, robots, and human morality. Ethics & Information Technology 12: 235-241.

42. JH Moor (2009) Four kinds of ethical robots. Philosophy Now 72: 12-14.

43. M Brundage (2014) Limitations and risks of machine ethics. Journal of Experimental & Theoretical Artificial Intelligence 26: 355-372.

44. TM Powers (2011) Incremental machine ethics. IEEE Robotics & Automation Magazine 18: 51-58.

45. G Marzano (2018) The turing test and android science. J Robotics Autom 2: 64-68.

46. M Coeckelbergh (2012) Can we trust robots? Ethics & Information Technology 14: 53-60.

47. JA Bergstra, M Burgess (2019) A promise theoretic account of the Boeing 737 Max MCAS algorithm affair. Computer Science.

48. JA Bergstra, M Burgess (2020) Candidate software process flaws for the Boeing 737 Max MCAS algorithm and a risk for a proposed upgrade. Computer Science.

49. F Santioni de Sio, J van den Hoven (2018) Meaningful human control over autonomous systems: A philosophical account. Frontiers in Robotics and AI 5: 15.

50. C Bagnoli (2014) Starting points: Kantian constructivism reassessed. Ratio Juris 27: 311-329.

51. C Lafont (2004) Moral objectivity and reasonable agreement: Can realism be reconciled with Kantian constructivism? Ratio Juris 17: 27-51.

52. C Bagnoli (2020) Constructivism in metaethics. In: Ward N Zalta, The stanford encyclopaedia of philosophy, 27: 311-329.

53. D Beyleveld, M Du¨well, A Spahn (2015) How should we represent future generations in policy making? Jurisprudence 6: 549-566.

54. A Gewirth (2001) Human rights and future generations. In: M Boylan, Environmental Ethics, Prentice Hall, 207-212.

55. T Zhang, W Zhang, MM Gupta (2017) Resilient robots: Concept, review, and future directions. Robotics 6: 22.

56. D Rus, MT Toley (2015) Design, fabrication and control of soft robots. Nature 521: 467-475.

DOI: 10.35840/2631-5106/4126

ISSN 2631-5106
9 772631 510005

Citation: Bergstra JA (2020) Promises in the Context of Humanoid Robot Morality. Int J Robot Eng 5:026