International Journal of Social Robotics (2021) 13:629–645
https://doi.org/10.1007/s12369-020-00653-w

Trust in and Ethical Design of Carebots: The Case for Ethics of Care

Gary Chan Kok Yew[1]

Accepted: 17 April 2020 / Published online: 23 May 2020

Abstract
The paper has two main objectives: to examine the challenges arising from the use of carebots as well as to discuss how the design of carebots can deal with these challenges. First, it notes that the use of carebots to take care of the physical and mental health of the elderly, children and the disabled as well as to serve as assistive tools and social companions encounter a few main challenges. They relate to the extent of the care robots' ability to care for humans, potential deception by robot morphology and communications, (over)reliance on or attachment to robots, and the risks of carebot use without informed consent and potential infringements of privacy. Secondly, these challenges impinge upon issues of ethics and trust which are somewhat overlapping in terms of concept and practice. The existing ethical guidelines, standards and regulations are general in nature and lack a central ethical framework and concrete principles applicable to the care contexts. Hence, to deal with these important challenges, it is proposed in the third part of the paper that carebots be designed by taking account of Ethics of Care as the central ethical framework. It argues that the Ethics of Care offer the following advantages: (a) it provides sufficiently concrete principles and embodies values that are sensitive and applicable to the design of carebots and the contexts of caring practices; (b) it coheres with the tenets of Principlism and select ethical theories (utilitarianism, deontology and virtue ethics); and (c) it is closely associated with the preservation and maintenance of trust.

Keywords Care robots · Trust · Ethics · Design · Ethics of care · Artificial intelligence

1 Introduction

Care robots are utilised in hospitals and homes to provide care and support for vulnerable persons such as the elderly, children and those suffering from physical and mental disabilities. They monitor the health conditions of patients, give medication, manually lift and aid the movements of disabled patients, and provide social companionship.

There are tangible advantages in using carebots to relieve human caregivers (whether healthcare professionals, social workers in nursing homes or family members and loved ones) from the significant assistive manual work and medication reminders to patients and care recipients. Care-O-bot—a home companion for the elderly—open doors and fetch household items, remind the elderly of his or her daily routine and seek help in the event of a fall or accident.[1] Robear acts as a nursing care robot to lift patients from the bed to a wheelchair. The US-based robotics corporation, Acrotek, advertise a robot for sale in an aged-care role, claiming on its website that the Actron MentorBotTM will keep track of the elderly persons or remind them to take pills. It is touted to have the ability to call the authorities and report a problem or make a call to loved ones when the patient has gone astray.

Apart from the obvious benefits in relieving the human caregivers of the manual work and routine administration of medication, the carebots can serve as social companions. Sony's AIBO robotic dog and NeCoRo (OMRON), a robotic cat covered in synthetic fur, come to mind. PARO—an interactive robotic seal developed by AIST in Japan to comfort dementia sufferers—makes high-pitched sounds, and responds to touch by moving its head and tail. It has been reported in studies that patients may prefer interacting with robots than human companions.[2] In addition, robots can help in facilitating therapeutic exercises and rehabilitation for patients suffering from both physical and mental disabilities. Since 1998, project Aurora (AUtonomous RObotic platform as a Remedial tool for children with Autism) has encour-

[1] Sorell and Draper [67].

✉ Gary Chan Kok Yew
garychan@smu.edu.sg

[1] School of Law, Singapore Management University, Singapore, Singapore

[2] Dautenhahn and Werry [19].

aged children with autism to become more engaged in human interactions.[3] In a pilot study, the autonomous humanoid KASPAR robot taught children with autism to engage in collaborative play through a video game. It found that the children were more engaged by the robot than the human counterpart (though the children showed more examples of collaborative play and cooperation when playing with the human adult).[4] Robots can also teach young children with illness or disabilities to deal with them.[5] The report on robot Matilda[6] that was trialed in Australian residential aged care facilities indicated a statistically significant improvement in the emotional and behavioural engagement of the older people suffering from dementia with the social robot [39]. To deal with physical disabilities, leg rehabilitation robots allow for physiotherapy[7] and gait assessment robots can capture gait parameters to customise rehabilitation interventions and exercises to cater to the patient's needs.[8]

Notwithstanding the substantial benefits (actual and potential) that carebots can bring to the vulnerable segments of the population, it is first and foremost, observed that there are unique challenges arising from the use of carebots. These challenges relate to issues of both ethical design and the public or users' trust in carebots. The second part notes that the existing ethical theories (of utilitarianism, deontology and virtue ethics) as well as ethical guidelines, standards and regulations on robots do not provide an overall ethical framework which developers can take into account in the design of robots or sufficiently concrete principles for application to the caring contexts. The third part will then argue for the Ethics of Care as the most appropriate ethical framework that can apply to the caring contexts. This central framework, together with the tenets of Principlism (comprising autonomy, non-maleficence, beneficence and justice) and other ethical theories (utilitarianism, deontology and virtue ethics) can deal with the current challenges presented by the use of carebots.

2 Challenges in the Use of Carebots

Robots may be defined as artificial beings possessing the four characteristics of mobility, interactivity, communication, and autonomy [UNESCO 76, p. 4]. These robots may be powered by artificial intelligence that includes the ability to reason, communicate, and learn from its perception of the surrounding environment, past experiences and even errors. It is artificial intelligence that enables them to respond in socially interactive ways with humans. Advancements in machine learning technology including deep learning and natural language processing have significant potential to enhance the functionality of social robots such as the care robots used in the healthcare sector. The specific roles that carebots are expected to play in healthcare however give rise to unique challenges.

1. Extent of robot care

Sparrow and Sparrow[9] have argued that robots are incapable of meeting the social and emotional needs of older persons under their care. Robots do not understand human frailties and therefore cannot show care for vulnerable people. Moreover, the use of carebots may result in a decrease in the amount of human contact experienced by the older persons, which would be detrimental to their well-being. It is unethical in their view to attempt to substitute robot simulacra for genuine social interaction. The inability of robots to care for humans and their increased role of robots in social interactions give rise to potential problems of deception and over-reliance on robots by users (which we will discuss below).

On this issue of the carebots' inability to care, we should take note of both internal and external aspects. The internal aspect is that carebots lack human consciousness and do not possess any human emotions.[10] As such, they lack human understanding as to how it feels to care for a human. It might be argued, somewhat ironically, that the robot's inability to feel pain and suffering or experience vulnerability as humans do accounts largely for their inability to show care to humans.[11]

Technological developments suggest that robots have the capacity to "perceive" human emotions and facial

[3] Dautenhahn and Werry [19, p. 5].

[4] Wainer et al. [85].

[5] Alemi et al. [1] and Meghdari et al. [47, 48] relating to children suffering from cancer; Meghdari et al. [49] on children with hearing disabilities.

[6] The social robot was jointly developed by NEC Japan and RECCSI (Research Centre for Computers, Communication and Social, Innovation) of La Trobe University in Melbourne. It possesses human-like attributes such as baby-face appearance, human voice, facial expressions, gestures and body movements and is able to recognise voices, human faces, emotions and possesses speech acoustics recognition.

[7] "Assistive Technologies to Improve Healthcare quality, productivity" MIS Today, 27 Feb 2016.

[8] "Charting the way to hospitals of the future" 23 January 2018 at https://edb.gov.sg/en/news-and-resources/insights/innovation/charting-the-way-to-hospitals-of-the-future.html (accessed on 20 February 2019). These rehabilitative robots are developed by The Centre for Healthcare Assistive & Robotics Technology (CHART) in Singapore.

[9] Sparrow and Sparrow [68].

[10] This does not necessarily mean robots cannot have internal states that mirror human emotions and which can be communicated to humans through facial expression, body posture and tone of voice: see Breazeal and Brooks [10] on the Kismet social robot with three-dimensional affect space of arousal; valence and stance.

[11] Metzinger [50].

expressions.[12] For example, Huggable was designed for therapeutic applications in nursing homes. The interactive social robot possesses the neural network to recognise nine different classes of affective touch and can react accordingly. Another example is the Model of User's Emotions which has been developed for the healthcare industry to assess patients' emotions from various sources (including heart rate, breathing pattern, temperature, vocal characteristics and facial expressions). As Wallach and Allen described, technology has enabled the translation of sensory data such that they can be received as "cognitive representations of emotions".[13] This does not mean, however, that robots are capable of perceiving real human emotions or capable of having emotions as humans do.

Despite this internal incapacity to care, carebots are capable of exhibiting the *external* or 'outward' aspect of care. Robots can be designed to respond to humans and human behaviours in a manner that may be perceived as 'caring' by the human care recipients. The external aspect of care may be demonstrated through words importing kindness and encouragement, a gentle pat on the back of the care recipient or a programmed smile on a humanoid carebot without the carebot possessing any corresponding (internal) emotions. The philosopher John Searle [62] has demonstrated in his famous Chinese room illustration that AI systems lack true understanding of the Chinese language terms being processed by them. The computer receives inputs on Chinese language terms and produces Chinese characters as outputs. Despite not understanding these Chinese terms, the computer would be able to pass the Turing test [75] (that is, to convince a Chinese speaker that it possesses knowledge of the Chinese language) by "simulating" an understanding of the Chinese language. Thus, the computer does not possess a "mind" or "think" the way humans do. Indeed, Turing [75] had remarked that the question "could a machine think?" was in itself meaningless.

Drawing an analogy from human–human interactions, the fact that a human cannot read the (internal) minds and emotions of other humans does not prevent him from showing the external aspect of care to another human being. In truth, the "appearance" of emotions are not real human emotions. Nonetheless, the absence of (internal) human emotions in robots can be compensated for by the robots' external appearances and behaviour. Coeckelbergh [15] conjectured that advanced human–robot communications can be analogised to normal human–human interactions in that both can interpret "the other's appearance and behaviour as an emotion".[14] This approach requires a shift in mindset "from the 'inside' (what is 'in the mind' of robots) to the 'outside' (what robots do to us)" in personal, social and emotional contexts.[15]

2. Deception

Indeed, it is this ability of carebots to exhibit "external" care that gives rise to the problem of deception. Carebots that aim to imitate a human companion or caregiver raise the possibility that the user (especially the vulnerable) will be unable to judge whether they are communicating with a real person or with technology. This could be experienced as a form of deception or fraud.[16] Can the deception be regarded as benevolent or benign? Does that make the deception morally permissible?

The form that the carebot takes may be significant. Coeckelbergh et al. [17] indicated that zoomorphic robots may be less problematic than robots that look too much like humans. This does not imply that anthropomorphic designs should be prohibited. In fact, anthropomorphic designs of robots are arguably desirable if they advance the function of the technology[17] such as to encourage the use of or interactions with a robot for the patient's well-being.

Grodzinsky et al. [33] regarded deception by appearances as a question of trust.[18] A robot is deceptive where it misleads the human user into believing that or behaving as if the robot is a human or animal. They argue that where the software developer has used deception in the design of the robot in order to help the user, the deception is considered benign and does not involve a breach of trust. For example, manipulations or deceptions of a robot for the purpose of calming a patient with dementia which might otherwise cause injury to the hospital staff or others would be benign without involving any breach of trust and are thereby morally permissible.[19] The underlying basis for Grodzinsky et al. [33] appears to be the positive consequences generated or the avoidance of negative consequences or harms. This is analogous to the justification for therapeutic privilege exercised by doctors to withhold relevant information from the patient about their health condition if the disclosure of information is likely to cause serious physical or mental harm to the patient. But if the developer manipulates the user and does not act in the user's best interests, there is a breach of trust which cannot be condoned.[20]

[12] Pour et al. [57].

[13] Wallach and Allen [86, p. 152].

[14] Coeckelbergh [15, p. 238].

[15] Coeckelbergh [14, p. 219].

[16] Mittelstadt [51].

[17] Darling [18].

[18] See Burrell [12] on the economic and social inequalities that can arise from the opacity of machine learning algorithms including her examples concerning Internet fraud and spam filtering.

[19] Grodzinsky et al. [33, p. 97].

[20] Grodzinsky et al. [33, at p. 97].

The default position should be that deception is morally wrong due to deontological grounds (e.g., respect for humanity) and because they offend the virtues of honesty and trustworthiness. They argued that an exception should apply where the intention of the developer is ethical and the consequences are good.[21] Grodzinsky et al. [33] also added that the deceptive feature should be an essential functionality of the robot Isaac and Bridewell[22] argued that deception-capable robots can be ethical even when telling outright lies and that the ethicality of human or robot communication must be assessed with regard to its underlying motives (such as the achievement of pro-social goals and functions). However, Isaac and Bridewell were concerned with social interactions and conversations between robots and people generally as opposed to the relationship between carebots and the very young and the elderly who are vulnerable and require assistance.

Taking Grodzinsky et al. [33] as laying the basic requirements, I would argue that deception by carebots with regard to vulnerable persons may be allowed only if (a) the intention of the developer is ethical with respect to the care recipients, (b) the consequences are indeed positive for the care recipient, (c) there are no viable alternatives to the use of deception; and (d) the extent of infringement of the care recipient's autonomy is not more than that via other means. The additional two requirements (c) and (d) would further protect the vulnerable children and elderly from deceptive robots. Consider the analogy to lying to a dementia patient using Simulated Presence. This is a device developed for Alzheimer's patients comprising an audiotape that records the telephone conversation of the family member concerning his or her memories of the patient. Even if the tape is played back repeatedly via a device that looks like a telephone, the Alzheimer's patient may not realise and continue to regard it as a fresh conversation.[23] This is a form of deception; however, the underlying motive is arguably to improve the patients' emotional well-being. It must be shown that the device would in fact enhance the patients' emotional well-being. It must also be asked whether there are alternatives to Simulated Presence which do not deceive and whether it might be feasible to instead invest time to carry out real conversations with patients.[24] Where there are no viable alternatives to the use of the device, and the extent of infringement of personal autonomy is not more than via other means, such deception may be justified as benign vis-à-vis the care recipient and thus morally permissible.

The question of trust can also be assessed from the users' perspective which may in turn depend on the level of communications between the robot developers and users.[25] It is possible that though the patients know that the robot is not real, they have chosen to invest emotionally in the robots nonetheless. Humans can trust artificial agents on an emotional level instead of merely being dependent on them for practical functionality even if the humans know for a fact that they are only machines.[26] Mentally healthy humans who are aware that they are interacting with a robot are not deceived.[27] In such a case, there is arguably no breach of trust.

Hoorn and Winter [35] found in a study that, insofar as delivering bad health news to patients is concerned, the participants preferred the robot doctor and the robot's message to the human counterpart. Moreover, the robot garnered more compliance to the medical treatment. They noted that robots may outperform humans on emotional tasks and this can relieve physicians from the demanding duty of disclosing unfavourable information to a patient.

3. Over-reliance on and over-attachment to carebots.

Over-reliance can adversely affect both care recipients and caregivers. Caregivers may be over-reliant on robots to do the caring work, and technology (for example, to aid patients with motor impairment) may at times impede the improvement in health conditions of care recipients (where the care recipient refuses to make the attempt to walk without technological assistance). Vulnerable patients may also suffer from over-attachment to carebots. It is commonly found that children suffer from distress and grief when separated from their robot companions[28] and the reliance on carebots without adult guidance may impede the development of interactive abilities of babies and infants in the long term.[29] As mentioned above, carebots do not possess human attributes such as compassion.[30] Clinical practice often involves complex judgments and abilities that AI technology is currently unable to replicate, such as contextual knowledge and the ability to read social cues.[31] Concerns have been raised about a loss of human contact and increased social isolation if AI technologies are used to replace therapists or family time with patients.[32]

[21] Grodzinsky et al. [33, p. 99].

[22] Isaac and Bridewell [37].

[23] Schermer [61, p. 160].

[24] Schermer [61, p. 165].

[25] See Felzmann et al. [24] on relational understanding of transparency that is dependent on the communication between technology providers and users, where trustworthiness is assessed based on contextual factors that mediate the value of such communications including factors that make transparency meaningful and trustworthy in the users' eyes.

[26] Grodzinsky et al. [33].

[27] Coeckelbergh [16, p. 288].

[28] Riek and Howard [58].

[29] Sharkey and Sharkey [64].

[30] Parks [56].

[31] Loder and Nicholas [44].

[32] Sharkey and Sharkey [63].

In addition, if caregivers rely on carebots to take over the caring tasks, less time will be spent on human–human interactions. Vallor argued that caregiving teaches us reciprocity and empathy.[33] Thus, carebots should not "liberate" us from care but instead provide support that draw us into care-giving practices.[34] To mitigate the problem of over-reliance (and over-trust) by care recipients, human designers may consider suitable measures such as providing warning indicators or built-in tasks that requires the attention of the user from time to time.[35]

Given the challenges of using carebots, it is proposed that carebots should not *replace* human caregivers in interacting directly with the patient but merely *assist* the human caregiver in supporting or taking care of the patient.[36] Where the carebots are used as social companions and for social interactions *together* with human caregivers to care for those who are vulnerable, the risk of over-reliance or over-attachment to the robots is mitigated. Coeckelbergh et al.[37] suggested that we should develop and use robots based on the notion of "supervised autonomy". This would likely create more trust among stakeholders and improve the quality of the therapy. Vallor [77][38] has also noted that the use of carebots can deprive potential caregivers of goods (such as the virtues of reciprocity and empathy or, applying the capabilities approach,[39] the preservation and enhancement of their human capacities for affiliation, practical reason and emotion) that she considered to be central to care practices.

4. Informed consent to use of carebots and patient privacy

As the use of carebots may bring risks for the patients and users, the procedures of obtaining informed consent should be customised to ensure each patient understands the purpose and risks from using carebots. The patient's consent to the use of carebots has to be voluntarily and unequivocal. Greater caution should be exercised in the communication of the relevant information to children and elderly patients prior to obtaining consent. In particular, the patient's knowledge or lack thereof about the potential restraints on the patient's desire for independent living, and the deprivations of liberty or privacy that he or she may be subject to with the use of carebots should be taken into account; and such restraints or deprivations should be balanced against the potential benefits of safety and security that may be afforded by the carebots' role.[40] For persons who are not mentally capable of giving consent such as the very young children or elderly patients with advanced dementia, there may be a need to consider proxy decision-making with respect to the use of carebots.[41]

Carebots capture, store and process personal and sensitive data about the care recipient's health conditions and movements. They are networked devices that collect, store or process the data from various localities and in the cloud. The typical care recipient especially the young children and the elderly may not be aware of the significant data-processing capacity of carebots. Significant amounts of personal data and confidential information such as the health conditions and emotional responses of the care recipient during social interactions may be disclosed to and stored in the carebot[42] to which the robotics companies have access. There is a risk that the collection and disclosure of the data by carebots during such interactions including intimate situations would infringe privacy rights and thereby cause embarrassment to users. There is an additional risk that the database containing information on the patients' health conditions might be hacked or retrieved by unauthorised third parties. It is indeed timely to pay more attention to security and privacy issues with respect to data used for communication between people and robots and artificial intelligence [23].[43]

As a summary, it is pertinent to highlight that none of the abovementioned challenges (carebots' limited extent of care, the possibility of deception on vulnerable humans, (over)reliance and (over)attachment to carebots, and the lack of informed consent and potential infringement of users' privacy), serious as they are, would suggest that the use of all forms and types of care robots should be banned. Instead, we should pay close attention to the possibility of abuse and advocate the appropriate design of carebots according to ethical principles with a view to engender and maintain trust. We will now examine these concepts of trust and ethical design and their inter-relationship.

3 Trust

Human beings seek to adapt to their environment by reducing complexity and uncertainty. Trust is one mechanism that allows humans to cope with this complexity and uncertainty [45]. On one level, human trust in another person or thing is based on belief. Gambetta [29] stated that trust depends on the "subjective probability with which an agent assesses that

[33] Vallor [78, p. 223].

[34] Vallor [78, p. 226].

[35] Borenstein et al [9, p. 135; 84].

[36] European Parliament [22, at para 32].

[37] Coeckelbergh et al. [17].

[38] Vallor [77].

[39] See Nussbaum [54] cited in Coeckelbergh [15].

[40] Leenes et al. [42, at p. 22].

[41] Ienca et al. [36].

[42] Fosch-Villaronga and Albo-Canals [26 at p. 83].

[43] Para 125.

another agent or group of agents will perform a particular function".[44]

Taddeo [70] has laid further groundwork on the concept of trust (and e-trust). E-trust occurs in environments which there is no direct or physical contact and, in this regard, may not be entirely applicable to human interactions with social or care robots. However, certain concepts of trust remain relevant for our purposes. First, she noted that trust is a relation between A (the trustor) and B (the trustee). The trustee can be a human or artificial entity. Second, trust is a "decision" by A to delegate to B some aspect of importance to A in achieving a goal in which A's decisions are "designed and implemented with the assumption that there is a high probability that B will behave as expected". Third, Taddeo observed that trust involves risk and that "the less information the trustor A has about the trustee B, the higher the risk and the more trust is required". Fourthly, the A has the expectation of gain by trusting B. Fifthly, positive outcomes that are generated when A trusts B encourage A to continue trusting B.[45] Taddeo's model essentially relied on the 'rational agent" which is capable of making the "best option for itself, given a specific scenario and a goal to achieve".[46] In Taddeo [71], she referred to the quantification of risks underlying trust based on the ratio of successful actions to the total number of actions necessary to achieve the goal.

Ferrario et al. [25] noted that apart from *pragmatic* reasons for trusting an AI system dependent on the expectation of gain from trusting, there may also be *epistemic* reasons for trusting which are based on the trustor's belief in the trustee's trustworthiness. The AI system may be viewed as trustworthy in (a) a *relative* sense in the context of the trust relationship from a person's perspective but not another or (b) an *absolute sense* where there are objective reasons that make the AI system trustworthy for everyone whatever the contexts. The model on trust discussed thus far can be further extended in three ways: first, trust is not based on purely cognitive or rational beliefs but there is the other aspect of "attitude"; second, by incorporating normative features and thirdly, by extending the range of stakeholders involved.

First, human trust in care robots cannot be assessed based on rational grounds alone given the issues we have raised in the section above on subjective human perceptions of the roles, functions and appearances of care robots and potential deception by such robots. Jones [38][47] regarded trust as an "affective attitude" with the implication that we are justified to trust even when we are not justified in predicting a favourable outcome from the person being trusted. Gompei and Umemuro [31] have studied both concepts of cognitive and affective trust of humans as applied to their interactions with social robots.

On the second point relating to the normative features of trust, Buechner and Tavani [11] have described trust as "a (moral or nonmoral) normative relationship affecting two agents"—A and B—in which "A has the disposition to normatively expect that B will do such and such responsibly."[48] Tuomela and Hofmann [74] distinguished rational social normative trust (based on the trustor's social right to expect the trustee's intentional good-willed performance) from predictive trust and reliance which are reason-based beliefs about the trustee's intentional good-willed performance. Grounds for the abovementioned 'social right" to expect the trustee's intentional good-willed performance include friendship, sincere agreement, and a relationship governed by mutually respected social norms.

Trust should not be defined narrowly as the "decision" taken by the trustor with respect to the trustee or the delegated task. Rather, it can also explain the basis that underlies the decision (for example, as in the statement "I decide to delegate to B *because* I trust B to achieve the goal"). This will naturally lead to the question "why *should* I trust B" which in turns triggers the examination of the normative basis for trust.

The third point concerns the range of stakeholder interests[49] (that is, who do you trust?). Trustees in question may include not merely the artificial entity or carebot but also the manufacturer of the carebot or the software developers of the artificial intelligence system to be used for the carebot. The trustors cover consumers, users and members of the public. We should also consider the social norms applicable to the range of stakeholders whether it is between the consumers and users of care robots and/or the human designers of the care robots.

In sum, trust is a relational and normative concept. The idea of trust implies some uncertainty or risks that delegation of task to artificial agent may not proceed as planned. Whether trust exists and the extent thereof depends on the trustor as a rational agent weighing the potential benefits and losses from reliance on the artificial systems. The reasons may be pragmatic or epistemic. Apart from the rational aspects, human affective attitudes towards care robots should be taken into account. Moreover, the applicable social norms and stakeholders' interests are also part of the trust concept.

Trust is also an instrumental concept in that trust can lead to certain positive ends. With respect to care robots, engendering human trust in such robots will potentially facilitate its

[44] Gambetta [29, p. 217].

[45] I have omitted Taddeo's point that "The trustee B may or may not be aware that trustor A trusts B". This is merely a neutral feature.

[46] Taddeo [71, p. 244].

[47] At p. 15. See also Kirkpatrick et al. [40]

[48] Buechner and Tavani [11].

[49] This focus on stakeholders is also consistent with the notion of Responsible Research and Innovation that take account of societal impacts under the European Union's Framework Programmes for Research and Technological Development: see Von Schomberg [83].

widespread use in the healthcare sector. Conversely, distrust arising from a mishap in the use of care robots can easily hamper the applications of care robots in the healthcare sector. At the same time, we should also guard against "overtrust" of robots that can arise from automation bias where the human underestimates the "loss associated with a trust violation" and/or "the chance the robot will make such a mistake".[50]

4 Ethical Design

Ethical Design refers to the process by which ethical values or principles are taken into account or embedded in the design process of a product or device or technology. The ethical values and principles may be taken into account in two ways via a "top-down" approach (by feeding in advance a set of ethical principles embedded in the robot algorithm) or a "bottom up" approach (in which machine learning adapts and learns about an external set of values and principles based on the robots' observations of humans, human behaviours and the operative environment) or a combination of both approaches.[51] The first two methods represent ideal types. The top-down approach selects in advance a theory to apply and analyses the requirements for the design of algorithms and subsystems in order to implement the theory.[52] The task of determining the universal ethical values or principles in advance for carebots under the "top-down" approach is at best unwise (and at worst, impossible). We need to pay attention to the particular contexts in which these values or principles are meant to apply.

On the other hand, the "bottom-up" approach assesses a task based on a performance measure and the outcomes are analysed after fulfilment of the performance measure in order to yield a theory.[53] The expectation or demand that the machine learning platform in a carebot generate appropriate ethical values and principles from the carebot's perceptions of humans and human behavior would be impracticable. We would need to least supply the carebot a preliminary ethical framework or a set of ethical reference points.

In practice, ethical design is inevitably the product of both the "top-down" and "bottom-up" approaches.[54] In this regard, a basic ethical framework should be in place to guide the actions and decisions of the carebot in the initial stages; and the carebot should also be enabled to "learn" from the environment and contexts it encounters and be permitted to make adjustments to the ethical framework or even develop new moral norms within the parameters of the general ethical framework. As Wallach and Allen put it, it is the "dynamic interplay" between the analysis of project structure and testing of system designed to reach goals.[55]

The design process may cover aspects such as the physical form the robot takes, the way it communicates, the contents of the communication, the actions or practices it is capable of carrying out and the contexts in which they operate. To properly design carebots, the designer will have to be aware of the contexts in which the carebots will be utilised in the healthcare sector, the challenges arising from the use of robots in caregiving, and also their potential impact (in physical, mental and social terms) on humans (both care recipients and caregivers) in those specific contexts.

Similar to the concept of trust, there is also a need to take account of the range of stakeholders' interests—of the patient and their loved ones, the healthcare professionals, caregivers, manufacturers and software developers—in ethical design. Polls may, for example, be taken of potential users of carebots such as the aged[56] in order to build "democratic spaces" for the voices of stakeholders to be heard.[57] There is also a need to consider the cognitive dimension of the user and his needs in the design process so as to develop robot care technology or devices that he can better use or interact with.[58]

Public trust in technology is dependent on whether the technology generates benefits for humans, improves their well-being and ensures safety. Should there be an accident or mishap in its use, distrust can easily creep in. Whether trust can be regained will depend on the efficiency and reliability of investigations and the follow-up actions by the developer and authorities. Another aspect is the consistency of functionality and performance of the technology such as carebots. In this regard, artificial agents are generally more predictable than humans and hence more trustworthy. [59] Whether the technology performs its role and functions for which it is designed (such as the roles of caring for its users' health and the enhancement of human–robot interactions for users' mental well-being through, for example, displaying "socially acceptable" emotions)[60] will be important for engendering

[50] Wagner et al. [84 at 22].

[51] Banavar [6].

[52] Wallach and Allen [86, pp. 79–80].

[53] Wallach and Allen [86, p. 80].

[54] For a macro-perspective of the regulatory process of robot governance, see Fosch-Villaronga and Heldeweg [27].

[55] Wallach and Allen [86, p. 81].

[56] Sparrow and Sparrow [68].

[57] Vandemeulebroucke et al. [81].

[58] Fosch-Villaronga and Özcan [28].

[59] Grodzinsky et al [32, p. 26].

[60] Ojha et al [55] on the development of the Ethical Emotion Generation System (EEGS) based on data structure of emotions represented in the form of (Name, Valence, Degree, Threshold, Intensity, Decay Time), where Name denotes the name for the type of the emotion, Valence specifies whether the emotion is positive or negative, Degree represents the extent of the positivity and negativity of the emotion, Threshold represents the minimum intensity required to trigger the emotion, Intensity represents the strength of the emotional experience and Decay Time denotes the time required to drop the emotion intensity back to 0.

trust in the technology. Finally, the ethical design of carebots can and should take account of potential risks arising from humans' automation bias that result in overtrust in robots.

Thus, in terms of concept and purpose, both trust and ethical design are well aligned insofar as the use of carebots is concerned. I will also argue that there are appropriate criteria we can consider for the promotion of trustworthy and ethically designed carebots in the healthcare context.

5 Finding the Right Ethical Framework for the Design of Carebots

The ethical design of carebots may involve the translation of abstract ethical principles into concrete rules of action to be programmed into or to be taken into account in the design of the carebot. As mentioned above, a combination of the "top-down" and the "bottom-up" approaches will be required in practice. Both Utilitarianism and Deontological Ethics (Kantianism) are essentially "top-down" theories. Virtue ethics is more akin to the "bottom-up" approach. I argue that Ethics of Care with its sub-level principles and complemented by Principlism can integrate the abstract and the particular in a manner that is sensitive to the various caring contexts and would be most appropriate for promoting trust and the ethical design of carebots.

Utilitarianism assesses ethical decision-making by reference to the sum of happiness to the greatest number. In so doing, a utilitarian would have to assess the relative weights of pleasures and pains arising from an action and determine the future effects from the perspective of an impartial spectator. Bernard Williams says the agent must be omniscient and benevolent to make a proper utilitarian assessment. Such a task is virtually impossible for a robot. If Utilitarianism were to be applied to the design of a carebot, it would at best be a truncated or partial version. Decisions would have to be made under "bounded rationality" so to speak. One method to circumvent the problem might be to resort to rule utilitarianism as a rule of thumb based on overall costs and benefits for similar actions or situations. However, the rules of thumb (for example, killing is wrong) must be determined in advance for them to be useful for the ethical design of carebots.

The Kantian deontological ethics is premised on two main categorical imperatives: (1) universalisation of standards ("act only according to the maxim by which you can at the same time will that it should become a universal law") and (2) to treat humanity, whether in your own person or that of another, always as an end and never as a means only. These are strict and universal rules applied without exception. For example, lying is wrong according to Kant as universalizing the maxim of lying would give rise to a fundamental contradiction and impossibility in communications. As such, lying would not be condoned even if the purpose were to save a life.

Designing algorithms to capture the universal principles would be daunting. With respect to the first categorical imperative, the carebot would first have to recognise goals for its own actions, assess the effects of all other moral agents trying to achieve the same goals by acting in the same way in comparable circumstances and would need to possess an understanding of human psychology.[61] As for the second categorical imperative, the carebot should have no problem with the concept of using a human being as a means to an end (which is based on the logic of "if X, then Y") but would also have to understand the wholly different concept of what it means to treat a human being as an end in itself (an ontological concept). The latter would be extremely challenging if not impossible for the robot.

Virtue ethics contains an entire array of different virtues for selection. The contexts in which a particular virtue (such as kindness) may apply are wide-ranging. The determination of the rational "mean" of a particular virtue depends on the subjective disposition of the actor in the circumstances he is placed in. Put in another way, the right action is one that would be selected by the virtuous agent. As the focus of virtue ethics is on building human character and humanity through practices and experiences, it is more in line with the "bottom-up" approach similar to "connectionism" in terms of training and development.[62] There are no definitive ethical norms to inform the moral actor what he or she ought to do.[63] According to Aristotle, virtue ethics did not provide a precise "algorithm" of action even though the moral actor in virtue ethics will aim at human flourishing as the ultimate good.

Virtue and care ethics overlap in that they both focus on practice and values. Slote [65], for example, treats care as virtue. Vallor in *Technology and the Virtues* speaks of "technomoral care" as a "skillful, attentive, responsible and emotionally responsive disposition to personally meet the needs of others who share our technosocial environment".[64] That being said, it is fair to regard care ethics as focusing essentially on relationships whilst virtue ethics emphasise disposition and human character.

In the Christian and Confucian versions, virtue ethics also takes account of the golden rule of reciprocity (do to others what you want others to do to you). The algorithm would have to be sufficiently complex to take into account the artificial agent's observed effect of others' actions on itself, and assess consequences of its own actions on the others' affec-

[61] Wallach and Allen [86, p. 98].

[62] See for example Patricia Churchland on connectionist learning for moral cognition.

[63] This was what Swanton [69, p. 273] termed the "problem of indeterminacy".

[64] Vallor [78, p. 221].

tive states.[65] With respect to care ethics, we will discuss below more concrete sub-level principles for application to the caring contexts.

With respect to the healthcare context in particular, *Principlism* advocated by Tom Beauchamp and James Childress has been influential. It comprises four principles:

1. Autonomy—a norm of respecting and supporting autonomous decisions.
2. Non-maleficence—a norm of avoiding the causation of harm.
3. Beneficence—a group of norms pertaining to relieving, lessening, or preventing harm and providing benefits against risks and costs.
4. Justice—a group of norms for fairly distributing benefits, risks and costs.[66]

From the above norms, derivative rules such as "tell the truth", "keep your promises", "protect the privacy of others and do not pass on information in confidence" are generated.[67] Applying to carebots in the healthcare context, the principle of non-maleficence acts as a reminder to ensure safety for the patients and vulnerable care recipients. Any serious injury caused to the patients and care recipients by the carebots would likely generate negative publicity and undermine trust. It is thus paramount to ensure safety in the use of carebots especially since carebots interact on a regular basis with vulnerable people. Harm may include physical or psychological harm. Sorell and Draper[68] advocated autonomy as a priority with respect to the use of carebots. Unless the elderly persons are cognitively impaired, they should have the autonomy to make their own decisions in the use of the robots. In exceptional circumstances, autonomy may be sacrificed if respecting autonomy would endanger the user's life or physical safety such as when the patient is making suicidal requests. The robot should also report to or warn the patient about accidents even if it is against the patient's wishes.

MedEthEx (Medical Ethics Expert), an AI medical advisor, adopted Ross'[69] *prima facie* duties of autonomy, beneficence and non-maleficence (which form part of Beauchamp and Childress' Principlism) with rules for weighing the different *prima facie* duties when they conflict in a particular situation. MedEthEx has three components: a *knowledge-based interface* to select "duty intensities" for a specific case, an *advisor module* which determines the correct action by consulting "learned knowledge" and a *learning module* that abstracts guiding principles from particular cases provided by a biomedical ethicist acting as a trainer.[70] In the context of reminding patients to take medication, for example, a balance may be struck between respect for patient autonomy to skip medicine on certain occasions and the potential harm to the patient if the reminders are repeatedly ignored by reference to a time-based formula.[71] The humanoid robot, Nao, which was developed by the French company Aldebaran Robotics, receives instructions from the physician as to when to take medication, the maximum amount of harm that will occur if the patient does not take the medication, the length of time for the maximum harm to occur, the maximum amount of benefit from taking the medication and so on. Similar to MedEthEx, the Nao robot decides based on the three duties of respecting patient autonomy, non-maleficence and beneficence.[72]

The final principle of justice may feature in both public (for example, in the allocation or distribution of healthcare services to needy members of society) and private domains (such as in the division of labour in a care-based institution and the relative treatment of different care recipients in a hospital or nursing home). How do we design principles of justice in carebots? The carebots could follow a simple "first-come, first-served" rule; however, in the event of a public health epidemic and when pharmaceuticals are in scarce supply, the design of the carebots would have to take into account and weigh the relative needs of the patients in the institutional setting.[73] Though there may be societal concerns re the inequality of caring distributed in society and the burdens on caregivers, such macro-justice considerations need not be taken into account in the design of carebots. The focus could instead be on interpersonal justice in terms of the reciprocity between caregiver and care recipients in an institutional setting.

5.1 Ethical Guidelines, Standards and Regulations on Robots

In addition to the ethical theories (utilitarianism, deontology, virtue ethics, and principlism) discussed above, there are several emerging ethical guidelines, standards and regulations pertaining specifically to the design of robots and other artificial intelligent systems.[74] Regulation is itself a multi-faceted

[65] Wallach and Allen [86, p. 98].

[66] Beauchamp and Childress [7, pp. 12–13].

[67] Childress [13, p. 68].

[68] Sorell and Draper [67].

[69] Ross [59].

[70] Anderson and Anderson [2].

[71] Wallach and Allen [86, pp. 128–129].

[72] Anderson and Anderson [3].

[73] Asaro [4, p. 14].

[74] E.g. European Commission [21] (that trustworthy AI systems should be (1) lawful, (2) ethical, and (3) robust from a technical and social perspective; and that 'key requirements for Trustworthy AI are: (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) environmental and societal well-being and (7) accountability" (p. 4).

exercise comprising the different modalities of law, market, social norms, and technology. Most if not all of these standards are connected to and resonate with the existing ethical theories. They are also consistent with the emphasis on ethical design in this paper.

The European Parliament [22],[75] as mentioned above, whilst acknowledging the potential capacity of robots to enhance the mobility and integration of people with disabilities and elderly people, stressed that "humans will still be needed in caregiving and will continue to provide an important source of social interaction that is not fully replaceable". Subsequently, the European Parliament [23] noted the enabling capacity of AI and robotics to allow doctors and nurses to spend more time in high value activities including patient interaction,[76] and at the same time, recognised the impact of the increased use of sensors in robotics to enable patients to have "more personalised treatment and services and receive care remotely from their own homes, while also generating more meaningful data".[77]

The IEEE Global Initiative on "Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems"[78] focus on the following general principles:

- *Human Rights* Ensure they do not infringe on internationally recognized human rights.
- *Well-being* Prioritize metrics of well-being in their design and use.
- *Accountability* Ensure that their designers and operators are responsible and accountable.
- *Transparency* Ensure they operate in a transparent manner.
- *Awareness of misuse* Minimize the risks of their misuse.

On human rights, according to the UN Convention on the Rights of Persons with Disabilities, the care recipient's desire for independent living should also be respected.[79] A report by the Rathenau Instituut has recommended an individual's right to choose to have meaningful human contact rather than with a robot.[80]

The Foundation for Responsible Robotics[81] highlights that "[r]esponsible robotics starts before the robot has been constructed. Ethical decisionmaking begins in the R&D phase." The British Standards Institute Ethical Design of Robots (BS 8611) identifies potential ethical harms arising from robots and autonomous systems and provides guidelines on risk management associated with the ethical hazards. The standard covers safe design, protective measures and information for the design and application of different types of robots, including those used for industrial, personal care and medical purposes. According to the Engineering and Physical Sciences Research Council (EPSRC) Principles of Robotics 2011,[82] in addition to ethical design to ensure safety and security, "[r]obots should be designed; operated as far as is practicable to comply with existing laws & fundamental rights & freedoms, including privacy." Special reference was made to robots used in the healthcare sector and the importance of data privacy and protection:

> a robot used in the care of a vulnerable individual may well be usefully designed to collect information about that person 24/7 and transmit it to hospitals for medical purposes. But the benefit of this must be balanced against that person's right to privacy and to control their own life e.g. refusing treatment. Data collected should only be kept for a limited time; again the law puts certain safeguards in place. Robot designers have to think about how laws like these can be respected during the design process (e.g. by providing off-switches).

Further, "[robots]… should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent." Two of Murphy and Wood's Three Laws of Responsible Robotics state that:

- A human may not deploy a robot without the human–robot work system meeting the highest legal and professional standards of safety and ethics.
- A robot must respond to humans as appropriate for their roles.

In the Future of Life Institute Asilomar Principles for Beneficial AI (Jan 2017), the two principles with a focus on design are:

> *Value Alignment* Highly autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation.
>
> *Human Values* AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity.

[75] Para 32.

[76] Para 71.

[77] Para 82.

[78] See https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf (accessed on 22 April 2019).

[79] Article 19.

[80] Van Est et al. [79] at para 3.9.2.

[81] See https://responsiblerobotics.org/wpcontent/cache/page_enhanced/responsiblerobotics.org//aboutus/mission//_index.html_gzip (accessed on 22 April 2019).

[82] See https://epsrc.ukri.org/research/ourportfolio/themes/engineering/activities/principlesofrobotics/ (accessed on 20 February 2019).

In addition, ISO 13482[83] specifies guidelines for safe design and protective measures in respect of personal care robots (mobile servant robot, physical assistant robot and person carrier robot).

Drawing from the above discussion on ethical guidelines, standards and regulations, the ethical design of carebots should focus on a few crucial aspects: safety and security, respect for human rights such as the right to independent living, meaningful contact with humans, dignity and privacy, transparency, freedom from deception and the accountability of human designers of AI.[84] Following Murphy and Wood, it is also important to consider the proper "role" morality of carebots in response to humans especially care recipients. Notwithstanding that the above ethical guidelines, standards and regulations are consistent with the main challenges with respect to carebots, they are fairly general in nature and do not provide sufficiently concrete guidance as to how the ethical design of carebots should be approached. I propose below the Ethics of Care as a central ethical framework with sub-level principles applicable to the caring contexts in the healthcare sector.

6 The Ethics of Care

The story of Ethics of Care began with Gilligan's *In a Different Voice* in reaction to Kohlberg's scale of moral development from children to ethics grounded in universality and rationality in adulthood. Gilligan [30] argued that women tended to focus more on empathy, emotions and compassion when discussing moral dilemmas at the expense of arguments grounded in universality and rationality. One is not superior to the other; they are merely different approaches to morality. Several commentators have since built upon Gilligan's ideas and arguably established Ethics of Care either as an ethical approach that supplements existing ethical theories or as a standalone ethical theory.[85]

Noddings [53] viewed ethics of care as imposing an obligation on us to act with an attitude or motive of caring towards others. Instead of abstract general ethical principles, the Ethics of Care is focused on the individual's emotional sensitivity to particular others. In addition, a person who cares for another is "engrossed" in that other person; the former does not impose his viewpoint on the other but rather gives due attention to the latter's perspective concerning the world and his or her relationship to the world.[86] Thus, both Gilligan and Noddings take the position that "care ethics aims to meet the concrete needs of individuals in context-specific and responsive ways".[87]

Care Ethics, according to Virginia Held [34], is premised on a "caring relation" and care is a "practice" of responding to needs.[88] For her, care is both a practice and a value. Care Ethics focuses on the agent's ability to engage in the practice of care and the exercise of that ability and not merely the agent's motives of caring.[89] Persons are relational and interdependent not individualistic autonomous rational agents.[90] Care Ethics also values emotions [34, p. 10].

In contrast to Held, Slote [65] advocated an agent-based virtue ethics of caring. He viewed the caring person as a benevolent person who may prefer those who are near and dear. In *Ethics of Care and Empathy*, Slote [66] affirmed that ethics of care is rooted in "moral sentimentalism" and that empathy is the "primary mechanism of caring, benevolence, compassion, etc.".[91]

Tronto [72] arguably provides the most important set of concrete principles applicable to the caring contexts. First, she defined good care as comprising both a caring attitude and a caring activity. To her, care practices—as opposed to mere tasks to be carried out—involve both "thought and action" which are directed towards some end.[92] Second, caring entails four phases: caring about, taking care of, care-giving and care-receiving.[93] Tronto [73][94] argued for institutionalised caring based on the purpose of care, a recognition of power relations and the need for pluralistic and particular tailoring of care to meet individual needs. In essence, it is about "purpose, power and particularity". Tronto [73] also made the point that potential care recipients should be able to "state what their needs are" and to make choices as to how their needs can be met.[95] For the caregivers, consideration should be given to the negative aspect of the heavy burdens of caregiving to the extent that "there was no space in our lives outside of the circles of care".[96] Tronto laid down four ethical elements of care:

1. Attentiveness—care requires a recognition of others' needs in order to respond to them.
2. Responsibility—that we take upon ourselves to care for others.

[83] See ISO at https://www.iso.org/standard/53820.html. Accessed April 22, 2019.

[84] Winfield and Jirotka [87, p. 3].

[85] E.g. Slote [66].

[86] Slote [66, p. 12].

[87] Engster [20, p. 2].

[88] Held [34, p. 546].

[89] Held [34, p. 51].

[90] Held [34, p. 72].

[91] Slote [66, p. 4].

[92] Tronto [72, p. 108].

[93] Tronto [72, pp. 105–108].

[94] Tronto [73].

[95] Tronto [73, p. 167].

[96] Tronto [73, p. 167].

3. Competence—skills in which care is given.
4. Responsiveness—the "responsiveness of the care receiver to the care" in order to guide the caregiver.

Tronto's four elements connect well with the challenges encountered in the use of carebots as well as the tenets of Principlism and the basic ethical theories of Utilitarianism, Kantianism and Virtue ethics. The first element of attentiveness to patient needs constitutes one important step towards advancing respect for patient autonomy. This element recognises the need for the care recipient's privacy (both physical and informational privacy) to be respected. This would mean that the informed consent of care recipients to the use of carebots should be obtained. Yet the promotion of patient autonomy can at times conflict with the aim to meet their other needs (for example, when patients in exercising their autonomy refuse medical treatment or medication to cure their illness). Furthermore, some vulnerable patients may be mentally incapable of exercising autonomy in which case the decision to use carebots to cater to their needs must be made in their best interests upon taking into consideration their important physical and emotional needs Care ethics remains particularly relevant for such vulnerable and dependent persons such as patients, elderly, young children and disabled. In *Love's Labor Revisited* [41], Kittay argued that dependency is not something exceptional but is in fact "integral" to human life. At the same time, care recipients may not wish to be overly dependent on human caregivers and in this regard, the use of carebots could "liberate" the care recipient from this feeling of dependence [8, p. 282].

The concept of "beneficence" in Beauchamp and Childress' Principlism is related to the second element of "responsibility". Though we cannot always be certain that human caregivers are exhibiting genuine care for their care recipients, they must act professionally according to their roles to benefit the care recipients in catering to their physical and emotional needs. Vanlaere and Gastmans[97] have sought to justify care ethics by reference to the notion of "our own personhood and the personhood of the vulnerable other". It is ultimately about human "dignity" in the relationship between the caregiver as well as the care recipient.

Competence is linked to the skills and efficiency of the caregiver in providing care. The achievement of excellence in one's craft or work (in this case, in providing care) is an Aristotelean virtue.

Reciprocity is a common value in various ethical theories namely Utilitarianism, Kantianism and Virtue theory. Such a notion is captured in Tronto's fourth moral element of "responsiveness" of the care recipient to the caregiver. The relationship is reciprocal in order to ensure that the caregiver's objectives to provide care to the care recipient are met. Held [34, p. 34, 35] opined that "[c]aring is a relation in which carer and cared-for share an interest in their mutual well-being". Hence, Tronto's theory aims to benefit both caregiver and care recipient.

Justice at the societal level may at first blush be perceived as alien to the Ethics of Care with a focus on personal relationships. Noddings commented that caring is "engrossment" in the specific needs and desires of another person whilst justice is a commitment to abstract principles and rules. But Ethics of Care and justice are not diametrical opposites. Engster's theory of justice, for instance, is based on caring practices.[98] In fact, he goes further to advocate that the government ought to establish institutional frameworks for supporting and accommodating caring practices.[99]

The partiality towards or prioritisation of close relationships in the Ethics of Care does not present a problem for carebots as long as the needs of the patients under their assigned scope of duty are attended to. After all, similar to the role of the doctors, nurses and healthcare professionals, the carebot's role should be to act in the best interests of specific patients under its care. However, basic justice concerns should be adhered to. For example, patients in an institutional setting with similar needs and placed in similar circumstances should receive similar level of care from the assigned carebots within the institution.

7 Care Ethics for Trust in and Ethical Design of Carebots

In view of the limitations of carebots in providing care to humans, it is suggested that carebots should play an assistive role in caregiving instead of replacing human caregivers. The patient should continue to take care of, to the extent possible, his or her own health and well-being in conjunction with the use of assistive robots[100] for specific tasks forming part of the care practice.[101] The nurse should continue to play a supervisory role in taking care of the patient including exercising supervision over the robots delegated with the caring tasks. In addition, the designer of carebots has to take into consideration the specific caring tasks and ethical framework in the design of carebots. Thus, in the overall caring context, the human caregiver or nurse, the designer of carebots and even the care recipient himself remain morally responsible for the implementation and use of carebots.

[97] Vanlaere and Gastmans [82].

[98] Engster [20, p. 4].

[99] Engster [20, p. 11].

[100] Lehoux and Grimard [43, p. 334].

[101] European Parliament [22, at para 32].

Care Ethics coheres well with the World Health Organisation's framework for people-centred health[102] based on: patient safety, patient satisfaction, responsiveness to care, human dignity, physical and psychological well-being. In addition, the selection of Care Ethics as the central ethical framework is justified by the following reasons:

(a) the availability of sub-level and sufficiently concrete principles for application to the (health)care context,
(b) its coherence with Principlism (non-maleficence, beneficence, autonomy but less emphasis on macro-justice) and basic tenets of existing ethical theories, and
(c) the close association between trust and Ethics of Care in the caring context.

7.1 Care Ethics Offers Fairly Concrete Sub-level Principles for Application to (Health)care Context

van Wynsberghe [80][103] advocated the use of care values as the foundational values to be integrated into carebots (which she refers to as "care centered value-sensitive design"). The value-based framework is meant to evaluate care robots in an ethical manner both retrospectively and prospectively. The retrospective stance allows the designers to consider the impact of their design on care practices. According to the prospective angle, the designers seek to integrate the ethical framework in the design process. Van Wynsberghe's analysis is premised on Tronto's care ethics, which are in turn based on the moral elements of attentiveness, responsibility, competence and reciprocity. In addition to providing a normative account of the values in care, she took the view that the ethical design process can "foster trust between the public and the resulting robots".

According to van Wynsberghe [80], the ethical framework should cover the following components: context (hospital, nursing home or home), practice (such as lifting and feeding), the actors involved (nurse, patient and robot), the types of robots (enabling, assistance versus replacement) and the manifestation of moral elements (attentiveness, responsibility, competence and responsiveness).[104]

Attentiveness to care recipients' needs cover not only their physical needs but also their emotional needs for social interaction with human caregivers and carebots. Care recipients should be sufficiently informed of the scope of use of carebots and potential use of confidential information. The care recipient's right to independent living and meaningful contact with humans rather than robots if he so desires should not be ignored. Caregivers and carebots ought to respect the care recipients' physical privacy (such as by not eavesdropping on human conversations, and helping to mitigate the patient's potential embarrassment or distress in having to depend on their family members or relatives to take care of their toileting or bathing). Other examples include deactivating the video monitors during intimate procedures,[105] and programming the carebots to announce their presence in the midst of a human being especially in private spaces.

Although carebots cannot exhibit genuine human care from the "internalist" perspective, they can nonetheless discharge their *responsibility* towards care recipients by showing outwardly caring behaviour towards them using AI sensors and technology. Meacham and Studley [46] note that a caring relation is based on a care environment formed by gestures, movements and articulations that express attentiveness and responsiveness to vulnerabilities within the relevant context. This partially endorses Tronto's elements of attentiveness and responsiveness. Carebots should not be used to deceive vulnerable care recipients into thinking they are human caregivers unless the underlying motive or purpose is to improve the well-being of the care recipient, the well-being of the care recipient is indeed enhanced, there are no viable alternatives to achieve that purpose and the infringement is not more serious than other means.

Carebots should be designed to have the *competence* to carry out their tasks to fetch items for care recipients, aid their transfer from bed to wheelchair and be equipped with adequate machine learning to interact effectively with the care recipients. This may also include the development of acceptable AI models of emotions displayed by the robots in their interactions with the users [55].

The *responsiveness* of care recipients to guide caregivers in their caring roles suggest a notion of limited reciprocity which is likely to depend in practice on the mental and physical condition of the care recipients to respond to the caregiver. In line with Vanlaere and Gastmans' "concept of "personhood" above, Khosla et al. [39] also referred to the need for social robots to be designed to enable personalisation of its services and contents to suit the preferences and health conditions of aged people with dementia. They opined that designing carebots with the social context in mind can "facilitate a long-term meaningful reciprocal relationship between social robots and people with dementia".[106]

7.2 Coherence with Other Ethical Theories and Principlism

We have already encountered the material differences amongst the ethical theories. Nevertheless, Care Ethics com-

[102] See http://iris.wpro.who.int/bitstream/handle/10665.1/5420/9789290613176_eng.pdf. Accessed on April 22, 2019.

[103] van Wynsberghe [80]. See also Salvini [60, p. 436].

[104] van Wynsberghe [80, p. 420].

[105] Riek and Howard [58].

[106] Khosla et al. [39, para 6.3].

plemented by Principlism cohere with certain aspects of the ethical theories discussed above. For example, Utilitarianism—to the extent that it recognises the importance of satisfaction of needs as part of the principle of maximising the sum of happiness to the greatest number in the weighing of inter-subjective preferences—is relevant to Tronto's Ethics of Care that incorporates the moral element relating to Attentiveness to Needs. Utilitarianism also resonates with certain aspects of Principlism in that the former considers the avoidance of harms (non-maleficence) and the generation of benefits (Beneficence) in the overall utilitarian calculus.

Deontology is not merely about rational and universal principles. The root of deontology also arises from the sentimental and emotional aspects of human nature.[107] The Kantian categorical imperative to respect humanity as an end in itself is a case in point. Humanity can only be properly appreciated when we consider the rational, intellectual as well as emotional aspects of being human. In this regard, we need to respect the autonomy of the mentally capable patients and elderly to make decisions for their own health, an important ingredient in Principlism. The responsibility to take care of others in the Ethics of Care is a duty or obligation of beneficence that is derived from Kantian categorical imperatives and is again in line with Principlism.

Bearing in mind the critique of virtue ethics that it depends entirely on what a virtuous agent would choose as a right action without any epistemological grounding, Swanton [69][108] proposed a target-centred virtue ethics according to context. The aim of virtues of practice—which according to Swanton includes virtues of focus on a problem, creative virtues and "Dewey's imaginative deliberation" and virtues of dialogue[109]—is to arrive at right solutions by acting in a virtuous way to solve problems. Every problem or dilemma comes with constraints on its solutions (such as time and costs, conflicts amongst virtues and their targets). Two virtues may seem to conflict thus preventing the agent from making a decision without sacrificing a virtue. The virtuous agent, when he or she encounters an ethical problem requiring a decision, will seek to "integrate constraints on solutions" to the problems. This may sometimes involve modifying the initial constraints to solutions in a way that increases the possibilities for resolution. Swanton [69] gave the example where the initial constraint of "keeping promise to children" may be modified to "Be sincere. Show respect to the children and consult with them. Maintain trust, enjoyment".[110] This modification is made pursuant to the exercise of virtues of practice. Resolving open-ended problems requires practical wisdom, experience and expertise.[111]

This target-centred approach to virtue ethics does not require the agent to act from inner motives. This approach would *prima facie* suit robots which actions are not actuated by inner motives. However, to apply this target-centred virtue ethics, the algorithm would have to be capable of modifying the initial constraint from a strict principle (or prohibition) into one that is more general and contextual based on the virtues of practice. The added complication is that the processes for exercising these latter virtues are themselves open-ended or ill-structured rendering the task of designing algorithms to capture the nuances difficult indeed. At the same time, this complication reflects the unpredictable workings of human consciousness and behaviour. The human mind, according to Swanton [69, p. 280], is a "pattern completer" rather than a "logic machine" applying strict rules.

Principlism (with the norms of non-maleficence, beneficence and autonomy) can be integrated into the care practices. The example of robots administering medication to patients in a hospital or nursing home is apt. The design of the carebot must strike a judicious balance between giving the patient autonomy to decide and ensuring that potential harms to the patient's physical and mental health are avoided and that his or her health needs are met. Though carebots are not capable of exhibiting genuine human care towards others, they must be designed to exhibit caring behaviour towards the care recipients and promote their well-being (beneficence) at least from the "external" angle. Moreover, through empathic caring for others which also involves an obligation to respect them, the "relational" character of autonomy is emphasised in Care Ethics [66, p. 56].

For carebots, the idea of justice in the macro-societal context may not be necessary. The carebot is typically assigned to an individual patient in a home setting or a group of patients within a hospital ward. Within the institutional care context, for example, carebots can exhibit a limited form of justice in the fair allocation of care to the care recipients. It may be based on a few factors: the proper assignment of a carebot to an approximately equal number of patients with similar needs, a first-come-first-served rule to serve new patients, and the possibility of re-allocation of assigned tasks in emergency situations.

7.3 Close Associations Between Care Ethics and Trust

We have already discussed the close relationships between trust and ethical design. Baier [5] specially underscored trust, a basic relation between particular persons, as the fundamental concept of morality.

[107] Slote [66, p. 43].

[108] See chapter 12 on "Virtues of Practice".

[109] Swanton [69, p. 253].

[110] Swanton [69, p. 256].

[111] Swanton [69, p. 258].

It is argued that adherence to Care Ethics promotes trust. Held [34, p. 92] stated that "care is not the same thing as trust, but caring relations should be characterized by trust, and caring and trust sustain each other".[112] For her, trust is important for care but it is not enough. There is still the need to do the work of care. Tronto's care ethics emphasised the relationship between the caregiver and care recipient and the latter's needs. This focus on the relationship between the parties advances mutual bonding and trust. Kittay [41, p. 248] noted that when a person has to depend on another, he or she also learns to trust.

Nissenbaum [52] related trust to security in the online context; however, it is not a straightforward and linear relationship. For care robots, the safety aspects of the care robots and the protection of the privacy of the patients and users will enhance their physical and psychological security and build trust. Nevertheless, it is merely one of the factors. Other factors include the competence of carebots, the promotion of wellbeing (beneficence) and the responsiveness of care recipients towards the caregivers and the care provided.

In sum, trust as a normative concept is promoted when care ethics is based on the competence of the carebots in an assistive role to benefit care recipients, the proper allocation of tasks and shared roles between the carebots, human caregivers, designers and even the care recipients themselves as well as the expectations engendered by care practices which should be conducted in a responsible manner towards the care recipients.

8 Conclusion

Carebots serve as a source of companionship and reminders for patients, the elderly and children and are able to alleviate the manual burdens and work of caregivers in taking care of such patients and vulnerable persons. The use of carebots encounters certain challenges as to the extent of care they are capable of giving, the problem of deception arising from the anthropomorphic form and appearance of robots, the concern with overreliance and attachment to the carebots, and the need to obtain informed consent and to show respect for the privacy interests of patients and vulnerable persons.

Trust is needed in order to enable the adoption of carebots in the healthcare context so that the benefits from the use of carebots can be optimised and risks minimised. Trust and ethical design are intertwined in the context of carebots in the healthcare for vulnerable patients, children and the elderly.

Care Ethics as the central ethical framework is capable of dealing with the abovementioned challenges. It has the potential to form the bedrock of an appropriate ethical framework for the design of carebots in the healthcare context. Ethical design and trust, though separate concepts, are mutually reinforcing insofar as carebots are concerned. Care Ethics is particularly relevant for vulnerable and dependent persons such as patients, elderly, young children and disabled. Tronto's version of Care Ethics offers concrete sub-principles based on the moral elements of attentiveness to needs, responsibility, competence and responsiveness that are applicable to carebots in the various caring contexts, and are not inconsistent with certain aspects of the main ethical theories (Utilitarianism, Deontology and Virtue Ethics). The three facets of Principlism (namely non-maleficence, beneficence and autonomy) and to a lesser extent, justice concerns, can also be integrated into the care practices.

It is an on-going process to ensure proper fit between Care Ethics in the design of carebots and public and user trust according to the different contexts of use in the healthcare sector. To promote the feasibility of the ethical design, the interests and needs of multiple stakeholders namely the needs and viewpoints of care recipients and their family members and the impact on human caregivers and healthcare professionals should also be taken into consideration at the design stage.

Acknowledgements This research is supported by the National Research Foundation, Singapore under its Emerging Areas Research Projects (EARP) Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author and do not reflect the views of National Research Foundation, Singapore. The author thanks Javier Han, a law student at Singapore Management University, for his research assistance.

[112] Held [34, p. 42].

References

1. Alemi A, Ghanbarzadeh A, Meghdari A, Moghadam LJ (2016) Clinical application of a humanoid robot in pediatric cancer interventions. Int J Soc Robot 8:743–759
2. Anderson M, Anderson SL (2005) MedEthEx: toward a medical ethics advisor. https://www.researchgate.net/publication/229043957_MedEthEx_Toward_a_medical_ethics_advisor. Accessed 22 Apr 2019
3. Anderson M, Anderson SL (2010) Robot be Good: a call for ethical autonomous machines. Sci Am 303(4):72–77
4. Asaro PM (2006) What should we want from a robot ethic? Int Rev Inf Ethics 6(12):9–16
5. Baier A (1987) Hume: the woman's moral theorist? In: Kittay EV, Meyers D (eds) Women and moral theory. Rowman & Littlefield, Lanham
6. Banavar G (2016) Learning to trust artificial intelligence systems: accountability, compliance and ethics in the age of smart machines. IBM. https://www.alain-bensoussan.com/wp-content/uploads/2017/06/34348524.pdf. Accessed 22 Apr 2019
7. Beauchamp TL, Childress JF (2009) Principles of biomedical ethics, 6th edn. Oxford University Press, New York
8. Borenstein J, Pearson Y (2010) Robot caregivers: harbingers of expanded freedom for all? Ethics Inf Technol 12(3):277–288
9. Borenstein J, Howard A, Wagner AR (2017) Pediatric robots and ethics: the robot is ready to see you now, but should it be trusted? In: Lin P, Jenkins R, Abney K (eds) Robot ethics 2.0: from autonomous

cars to artificial intelligence. Oxford University Press, Oxford, pp 127–141
10. Breazeal C, Brooks R (2005) Robot emotion: a functional perspective. In: Fellous J-M, Arbib MA (eds) Who needs emotions? The brain meets the robot. Oxford University Press, Oxford, pp 271–310
11. Buechner J, Tavani HT (2011) Trust and multi-agent systems: applying the 'diffuse, default model' of trust to experiments involving artificial agents. Ethics Inf Technol 13(1):39–51
12. Burrell J (2016) How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data Soc 3(1):1–12
13. Childress JF (2012) A principle-based approach. In: Kuhse H, Singer P (eds) A companion to bioethics, 2nd edn. Wiley-Blackwell, Hoboken, pp 67–76
14. Coeckelbergh M (2009) Personal robots, appearance, and human good: a methodological reflection on roboethics. Int J Soc Robot 1:217–221
15. Coeckelbergh M (2010) Healthcare, capabilities and AI assistive technologies. Ethic Theory Moral Pract 13:181–190
16. Coeckelbergh M (2012) Care robots, virtual virtue, and the best possible life. In: Brey P, Briggle A, Spence E (eds) The good life in a technological age. Taylor & Francis, Abingdon, pp 281–292
17. Coeckelbergh M, Pop C, Simut R, Peca A, Pintea S, David D, Vanderborght B (2016) A survey of expectations about the role of robots in robot-assisted therapy for children with ASD: ethical acceptability, trust, sociability, appearance, and attachment. Sci Eng Ethics 22(1):47–65
18. Darling K (2017) Who's Johnny? An anthropomorphic framing in human–robot interaction, integration, and policy. In: Lin P, Jenkins R, Abney K (eds) Robot ethics 2.0: from autonomous cars to artificial intelligence. Oxford University Press, Oxford, pp 173–188
19. Dautenhahn K, Werry I (2014) Towards interactive robots in autism therapy: background, motivation and challenges. Pragmat Cogn 12(1):1–35
20. Engster D (2007) The heart of justice. Oxford University Press, Oxford
21. European Commission (2019) Ethics guidelines for trustworthy artificial intelligence. High-Level Expert Group on AI
22. European Parliament (2017) 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))
23. European Parliament (2019) A comprehensive European industrial policy on artificial intelligence and robotics. European Parliament resolution of 12 February 2019 (2018/2088(INI))
24. Felzmann H, Fosch-Villaronga E, Lutz C, Tamo-Larrieux A (2019) Transparency you can trust: transparency requirements for artificial intelligence between legal norms and contextual concerns. Big Data Soc January to June 6(1):1–14
25. Ferrario A, Loi M, Viganò E (2019) In AI we trust incrementally: a multi-layer model of trust to analyze human-artificial intelligence interactions. Philos Technol. https://doi.org/10.1007/s13347-019-00378-3
26. Fosch-Villaronga E, Albo-Canals J (2019) Ï'll take care of you", said the robot. Paladyn J Behav Robot 10:77–93
27. Fosch-Villaronga E, Heldeweg M (2018) "Regulation, I presume?" Said the robot—towards an iterative regulatory process for robot governance. Comput Law Secur Rev 34(6):1258–1277
28. Fosch-Villaronga E, Özcan B (2019) The progressive intertwinement between design, human needs and the regulation of care technology: the case of lower-limb exoskeletons. Int J Soc Robot. https://doi.org/10.1007/s12369-019-00537-8
29. Gambetta D (1998) Can we trust trust? In: Gambetta D (ed) Trust: making and breaking cooperative relations. Basil Blackwell, Oxford, pp 213–238
30. Gilligan C (1982) In a different voice: psychological theory and women's development. Harvard University Press, Cambridge
31. Gompei T, Umemuro H (2018) Factors and development of cognitive and affective trust on social robots. In: Sam-Ge S, Cabibihan J-J, Salichs MA, Broadbent E, He H, Wagner AR, Castro-González Á (eds) Lecture notes in computer science, vol 11357. Springer, New York, pp 45–54
32. Grodzinsky FS, Miller KW, Martin MJ (2011) Developing artificial agents worthy of trust: "Would you buy a used car from this artificial agent?". Ethics Inform Tech 13(1):17–27
33. Grodzinsky FS, Miller KW, Wolf MJ (2015) Developing automated deceptions and the impact on trust. Philos Technol 28:91–105
34. Held V (2007) The ethics of care. Oxford University Press, Oxford
35. Hoorn JF, Winter SD (2018) Here comes the bad news: doctor robot taking over. Int J Soc Robot 10(4):519–535
36. Ienca M, Jotterand F, Vica C, Elger B (2016) Social and assistive robotics in dementia care: ethical recommendations for research and practice. Int J Soc Robot 8:565–573
37. Isaac AMC, Bridewell W (2017) White lies on silver tongues: why robots need to deceive (and how). In: Lin P, Jenkins R, Abney K (eds) Robot ethics 2.0: from autonomous cars to artificial intelligence. Oxford University Press, Oxford, pp 157–172
38. Jones K (1996) Trust as an affective attitude. Ethics Inf Technol 107(1):4–25
39. Khosla R, Nguyen K, Chu M-T (2017) Human robot engagement and acceptability in residential aged care. Int J Hum Comput Interact 33(6):510–522
40. Kirkpatrick J, Hahn EN, Haufler AJ (2017) Trust and human–robot interactions. In: Lin P, Jenkins R, Abney K (eds) Robot ethics 2.0: from autonomous cars to artificial intelligence. Oxford University Press, Oxford, pp 142–156
41. Kittay E (2002) Love's labor revisited. Hypatia 17(3):237–250
42. Leenes R, Palmerini E, Koops B-J, Bertolini A, Salvini P, Lucivero F (2017) Regulatory challenges of robotics: some guidelines for addressing legal and ethical issues. Law Innov Technol 9:1
43. Lehoux P, Grimard D (2018) When robots care: public deliberations on how technology and humans may support independent living for older adults. Soc Sci Med 211:330–337
44. Loder J, Nicholas L (2018) Confronting Dr Robot: creating a people-powered future for AI in health. Nesta Health Lab, London
45. Luhmann N (1979) Trust and power. Wiley, Chichester
46. Meacham D, Studley M (2017) Could a robot care? It's all in the movement. In: Lin P, Jenkins R, Abney K (eds) Robot ethics 2.0: from autonomous cars to artificial intelligence. Oxford University Press, Oxford, pp 97–112
47. Meghdari A, Shariati A, Alemi M, Vossoughi GR, Eydi A (2018) Arash: a social robot buddy to support children with cancer in a hospital environment. Proc Inst Mech Eng 232(6):605–618
48. Meghdari A, Shariati A, Alemi M, Nobaveh AA (2018) Design performance characteristics of a social robot companion "*Arash*" for pediatric hospitals. Int J Humanoid Rob 15(5):1850019
49. Meghdari A, Alemi M, Zakipour M, Kashanian SA (2019) Design and realization of a sign language educational humanoid robot. J Intell Rob Syst 95:3–17
50. Metzinger T (2013) Two principles for robot ethics. In: Hilgendorf E, Gunther JP (eds) Robotik und Gesetzbung. Nomos, Baden-Baden, pp 263–302
51. Mittelstadt B (2017) The doctor will not see you now. In: Otto P, Graf E (eds) 3TH1CS: a reinvention of ethics in the digital age?. IRights Media, Berlin, pp 68–77
52. Nissenbaum H (2001) Securing trust online: wisdom or oxymoron. Boston Univ Law Revew 81(3):635–664
53. Noddings N (2013) Caring: a relational approach to ethics and moral education. University of California Press, Berkeley

54. Nussbaum MC (2006) Frontiers of justice: disability, nationality, species membership. Belknap Press, Cambridge
55. Ojha S, Williams M-A, Johnston B (2018) The essence of ethical reasoning in robot–emotion processing. Int J Socl Robot 10:211–223
56. Parks JA (2010) Lifting the burden of women's care work: should robots replace the "human touch"? Hypatia 25:100–120
57. Pour AG, Taheri A, Alemi M, Meghdari A (2018) Human–robot facial expression reciprocal interaction platform: case studies on children with autism. Int J Soc Robot 10:179–198
58. Riek LD, Howard D (2014) Code of ethics for the human–robot interaction profession. https://www3.nd.edu/~dhoward1/a-code-of-ethics-for-the-human-robot-interaction-profession-riek-howard.pdf. Accessed 22 Apr 2019
59. Ross WD (2003) The right and the good, 2nd edn. Clarendon Press, Oxford
60. Salvini P (2015) On ethical, legal and social issues of care robots. In: Mohammed S, Moreno J, Kong K, Amirat Y (eds) Intelligent assistive robots: recent advances in assistive robots for everyday activities. Springer, New York, pp 431–445
61. Schermer M (2014) Telling the truth: the ethics of deception and white lies in dementia care. In: Foster C, Herring J, Doron I (eds) The law and ethics of dementia. Hart Publishing, Oxford
62. Searle J (1980) Minds, brains and programs. Behav Brain Sci 3(3):417–457
63. Sharkey N, Sharkey A (2010) Living with robots: ethical tradeoffs in eldercare. In: Wilks Y (ed) Close engagements with artificial companions: key social, psychological, ethical and design issues. John Benjamins, Amsterdam, pp 245–256
64. Sharkey A, Sharkey N (2011) Children, the elderly and interactive robots. IEEE Robot Autom Mag 18(1):32–38
65. Slote M (2001) Morals from motives. Oxford University Press, Oxford
66. Slote M (2007) The ethics of care and empathy. Routledge, New York
67. Sorell T, Draper H (2014) Robot carers, ethics, and older people. Ethics Inf Technol 16:183–195
68. Sparrow R, Sparrow L (2006) In the hands of machines? The future of aged care. Mind Mach 16(2):141–161
69. Swanton C (2003) Virtue ethics: a pluralistic view. Oxford University Press, Oxford
70. Taddeo M (2009) Defining trust and E-trust: from old theories to new problems. Int J Technol Hum Int 5(2):23–35
71. Taddeo M (2010) Modeling trust in artificial agents: a first step toward the analysis of e-trust. Mind Mach 29(2):243–257
72. Tronto J (1993) Moral boundaries: a political argument for an ethic of care. Routledge, New York
73. Tronto J (2010) Creating caring institutions: politics, plurality, and purpose. Ethics Soc Welf 4(2):158–171
74. Tuomela M, Hofmann S (2003) Simulating rational social normative trust, predictive trust, and predictive reliance between agents. Ethics Inf Technol 5:163–176
75. Turing A (1950) Computing Machinery and Intelligence. Mind LIX(236):433–460
76. United Nations Educational Scientific and Cultural Organization (UNESCO) and World Commission on the Ethics of Scientific Knowledge and Technology (COMEST) (2017). Report of COMEST on Robotics Ethics. 14 September 2017
77. Vallor S (2011) Carebots and caregivers: sustaining the ethical ideal of care in the twenty-first century. Philos Technol 24:251–268
78. Vallor S (2016) Technology and the virtues. Oxford University Press, Oxford
79. Van Est R, Gerritsen JBA, Kool L (2017) Human rights in the robot age: challenges arising from the use of robotics, artificial intelligence, and virtual and augmented reality—expert report written for the Committee on Culture, Science, Education and Media of the Parliamentary Assembly of the Council of Europe (PACE). Rathenau Instituut, The Hague
80. van Wynsberghe A (2013) Designing robots for care: care centered value-sensitive design. Sci Eng Ethics 19:407–433
81. Vandemeulebroucke T, de Casterle BD, Gastmans C (2018) The use of care robots in aged care: a systematic review of argument-based ethics literature. Arch Gerontol Geriatr 74:15–25
82. Vanlaere L, Gastmans C (2011) A personalist approach to care ethics. Nurs Ethics 18(2):161–173
83. Von Schomberg R (2013) A vision of responsible innovation. In: Owen R, Heintz M, Bessant J (eds) Responsible innovation. Wiley, London, pp 51–74
84. Wagner AR, Borenstein J, Howard A (2018) Overtrust in the robotic age: the ethical challenge. Commun ACM 61(99):22–24
85. Wainer J, Dautenhahn K, Robins B, Amirabdollahian F (2014) A pilot study with a novel setup for collaborative play of the humanoid robot KASPAR with children with Autism. Int J Soc Robot 6(1):45–65
86. Wallach W, Allen C (2009) Moral machines: teaching robots rights from wrong. Oxford University Press, Oxford
87. Winfield AFT, Jirotka M (2018) Ethical governance is essential to building trust in robotics and artificial intelligence systems. Philos Trans R Soc A 376:20180085

Gary Chan Kok Yew is Professor of Law at the School of Law, Singapore Management University. His research interests cover the ethical and legal aspects relating to artificial intelligence & health as well as Tort Law and Comparative Legal Systems. He currently serves as the project head for "Trustworthy AI" at the Centre for AI and Data Governance, Singapore Management University. He has obtained postgraduate degrees in both law (London) and philosophy (Birmingham).