Machine Medical Ethics: When a Human Is Delusive but the Machine Has Its Wits About Him

Johan F. Hoorn

Abstract When androids take care of delusive patients, ethic-epistemic concerns crop up about an agency's good intent and why we would follow its advice. Robots are not human but may deliver correct medical information, whereas Alzheimer patients are human but may be mistaken. If humanness is not the question, then do we base our trust on truth? True is what logically can be verified given certain principles, which you have to adhere to in the first place. In other words, truth comes full circle. Does it come from empirical validation, then? That is a hard one too because we access the world through our biased sense perceptions and flawed measurement tools. We see what we think we see. Probably, the attribution of ethical qualities comes from pragmatics: If an agency affords delivering the goods, it is a "good" agency. If that happens regularly and in a predictable manner, the agency becomes trustworthy. Computers can be made more predictable than Alzheimer patients and in that sense, may be considered morally "better" than delusive humans. That is, if we ignore the existence of graded liabilities. That is why I developed a responsibility self-test that can be used to navigate the moral mine field of ethical positions that evolves from differently weighing or prioritizing the principles of autonomy, non-maleficence, beneficence, and justice.

1 Autonomous Agencies

In medical ethical issues, patient autonomy is a top priority [1]. Autonomy is habitually attributed to "agency," something that can undertake an action on its own behalf or that of others. An agency is something in pursuit of a goal or that

J.F. Hoorn (✉)
CAMeRA—Center for Advanced Media Research Amsterdam,
VU University Amsterdam, Amsterdam, The Netherlands
e-mail: j.f.hoorn@vu.nl

S.P. van Rysewyk and M. Pontier (eds.), *Machine Medical Ethics*,
Intelligent Systems, Control and Automation: Science and Engineering 74,
DOI 10.1007/978-3-319-08108-3_15

has a concern. It has intentionality. This is what sets it apart from stones, planets, and mopeds. An agency may be an organic system (plant, animal, human) or it may be artificial (commonly software) but it should potentially be capable of acting autonomously, at least in part. An agency does not necessarily have to possess "free will," because its behaviour may be determined by the circumstances. After all, the autonomic nervous system, which regulates glands and internal organs, is hardly controllable consciously but does pursue the goal of maintenance and continuity of the organism it is a part of. When an agency is simulated by a (semi) autonomous software system, it is a software agent. A robot, then, is a software agent that (inter)acts through electro-mechanical devices. When it is specialized in humanoid simulations, the robot becomes an android: A robot that simulates human behaviour but not that of other organisms. When the android is applied to healthcare tasks in a user-centred manner, it is a Caredroid and Caredroids are the main topic of our considerations.

The Caredroid's simulation of human behaviour typically may be in interaction with other agencies, commonly patients, care professionals, or informal caretakers. In the current chapter, we will not deal with human-human or robot-robot interaction but focus on Caredroids in interaction with patients, particularly those with a mental disability. There are a handful of software agents and robots that help autism patients (e.g., [40]), serve as depression therapists (e.g., [34]), or ease the loneliness of Alzheimer patients (e.g., Paro, see [45]).

2 Beliefs

If you were a patient, would you take advice from a robot; a machine without any understanding of what it is saying; something without a consciousness? If you were a robot, would you listen to a patient; an organism with incomplete information and bias in judgment? Who do you believe? What do we think that the other believes?

2.1 Cat in a Chinese Room, Opened by Ames

Suppose the robot laboratory at Hong Kong Polytechnic University constructs a Chinese Room out of steel and invites John Searle to spend a sabbatical year with them—for free—provided that he stays in the room and they hold the key. Every now and then, the researchers slip pieces of paper under the door, asking him how he is doing: Whether he is hot, cold, feverish, has chills, pains, hunger, is sweaty, sleepy, about his thorax and abdomen, and the like. To test him, they pose the questions in Chinese. Luckily, a soothsayer told John that "his natural wit would be his fortune." The room is packed with books, filing cabinets, pens, and stationary, and John figures out how to correlate the Chinese characters of the questions

to an appropriate response, also in Chinese characters, which he slips under the door for the researchers to read. Although John does not know what he is saying, the researchers think he has perfect command of Chinese because all the answers make sense to them. Moreover, they think they can diagnose what is the matter with him, thinking he is thirsty whereas in fact he has to urinate.

Then the robot engineers Mark Tilden and David Hanson walk in, asking the researchers how they like their new emotionally perceptive John Searle robot portrait, locked in that Chinese Room over there. The robot engineers hold the Chinese character writer for a computer because how to determine he is a human? Promptly, another piece of paper appears under the door, stating that John Searle is a cat weeping over the mouse that he just has caught, signed by Erwin Schrödinger. Now everybody is in great despair. If Schrödinger is in there together with John Searle as his cat, Schrödinger will try to kill him through radioactivity and hydrocyanic acid [42], which he hid in a flask in one of the cabinet drawers. There will be no telling whether John the cat is dead or alive in that room and it will be "smeared out in equal parts" over the floor of probability [42].

Everybody agrees that as long as there is no direct observation, there is no way telling whether John Searle is in there, his robotic portrait, whether he is a cat, or that he is imitating Erwin Schrödinger with his cat, whether Schrödinger is in there, whether that cat is dead or not, or everything together?

The soothsayer steps in, foretelling that a dead cat should not be buried in the ground or it becomes a demon [5, p. 65]. It would be safer to hang it from a tree (ibid.). In undertaking immediate action upon this delusion, the soothsayer whose name is Ames draws an apple drill from his pocket and punctures a viewing peephole into the steel wall of the Chinese Room. Of course, everybody is pushing everybody else aside to see what is in there. What they see is astonishing. The inside of the Chinese Room as opened up by Ames is a distorted problem space where relative to the frame of reference of each individual observer (cubic or trapezoid), the apparent and actual position of the information processer inside that room makes it a great man or a humble robot. The conventional frame of Searle's Chinese Room as a cube demands that the information-processer inside that problem space should be human—with a conscious understanding of what it does. With the unconventional notion of a trapezoid, the processor is always halfway a robot and halfway a human, robot-like—even when alive, humanoid—even when lifeless, because it is all the same or at least indiscernible.

By this, I mean that the judgment "human or non-human" depends on your frame of reference. Searle's Chinese Room is an underdetermined problem space [32] in which you do not know what happens. Is the man Searle in there or is it his robot portrait? For the observer, the information-processing agency inside the Chinese Room is always Schrödinger's cat because he cannot know whether Searle is a lifeless machine or a living creature. There merely is a certain likelihood that the information-processing going on inside is more consciously human than mindless automation or perhaps even something in the middle [16, p. 45]; something smeared out over the floor of probability between true, probable, or false as described by Schrödinger's [42] "psi-function of the entire system."

In spite of not knowing whether true cognition is going on, we nevertheless bring it a glass of water although it has to urinate. We diagnose its health situation and take care of it, attributing it a sense of hunger and thirst and all different goal-directed behaviors, which make the machinery organic; make it "come to life." We tend to treat automated processes—including our own—as if they came from living creatures or real people (cf. the Media Equation Reeves and Nass [39]). And we do so because we were taught a frame of reference that says what the world is about and what the human condition is like [16, pp. 20–21]. Once apparent behaviors match our templates, we take them for actual. We work and take care of the agencies exposing those behaviors, according to our frame of reference or a priori belief system [16], containing, for example, certain moral principles (be good, don't hurt).

Because humans do not like to be confused, they prefer to force judgment into the direction of yes or no, in or out of category. We tend to apply logic-determinism to all possible problem spaces even in probabilistic cases. John Searle reproaches hard-nosed AI for looking at form and syntax alone to decide that the machine is "conscious," just like humans. Because the logics are the same, the difference between agencies becomes imperceptible in a Turing Test, so people themselves decide that the machine knows what it is doing. No, says Searle, if you look at semantic content, there is a difference because humans know what the forms and syntax refer to. Well, in the meantime researchers developed semantic Web techniques and machines that can reason through analogy and association, which is not far any more from what people do at a functional level because people also do not know eventually what the words stand for in the outside world—the world beyond their mental representations [16, p. 18]; a world that according to Schrödinger [43, p. 145] is just one of his cats again.

With respect to logics, then, it is hard to discern humans from machines; as it is increasingly with semantics. This position echoes Turing's [48] argument, stating that certain people may fail the Turing Test just as badly as the computer does. And because Schrödinger says that mental representation of the world is guesswork about the condition of a cat in a box that may die from poisoning, naïve empiricism does not do the trick either because our epistemology has no adequate access to the world about us except for what our senses, filters, and prejudices allow us to observe.

Ames teaches us that you have to look at it from a viewpoint of cognitive biases—illusory observations.[1] There is an unusual and perhaps even "trapezoid" psycho-logic to what we deem reality. The observer, however, has the bias—including our beloved philosopher John Searle—to apply conventional "cubic" logic to a probabilistic problem space, while staring at a cat that is fed by Erwin Schrödinger. That cat is an agency hardly discernible from a human or a robot when it processes logic-deterministic data. There are plenty of times that people are completely unaware of what they are doing and apply psychological schemas

[1] http://www.youtube.com/watch?v=hCV2Ba5wrcs.

or run scripts, delivering the right responses to a cue without "deep" understanding of the contents (cf. [26]). Think of chats about the weather, etiquette, or polite dinner conversations. People tell you what they are supposed to tell you and you think they are doing well, thank you… Think of officials that tick check boxes all day without thinking twice about the real-life consequences of their decisions. They behave like machines in Weber's [52] most rigid sense of the word. They continuously live in a Chinese Room.

If the cat agency processes probabilistic information, its best guesses are not distinguishable from a human's or a machine's either, because there is no way telling whatever guess is the right guess in the first place, whether the machine did a smart suggestion and the human a stupid one. In general, expert judgment in probabilistic situations does not exceed chance level [20, 47, p. 67] and is as good as a monkey's picking [37]. So there is only one thing that will discern a human from an animal from a machine, namely your own biased observations in relation to what you believe. Therefore, you need to realize what you think you know a priori about the agency in front of you (i.e., its ontological status or class), whether you believe your measurements and senses are right (i.e., your epistemology), and to know your own biases ("Am I a logician, an empiricist, or a cognitivist?"). Consequently, the more we think we know about the information processing capacities of a given agency, the more precise our classification will be (i.e., human, animal, machine)—without ever knowing its empirical correctness.

At the level of autonomous control, instinctive behaviors, and perhaps some aspects of memory, our behavior is nothing but machine-like. The more machines are capable of solving problems intelligently or even creatively, the more their behaviors become human-like. Finally, the two show indiscernible behaviors—passing the Turing Test brilliantly—that may have come from different processes but nobody can tell anymore, particularly when organic and digital circuitry becomes integrated (humans with electro-mechanical extensions, machines with biochips, a Robocop).[2]

Moral reasoning about healthcare dilemmas through machine computation yields judgments that are identical to those of medical ethical committees (e.g., [33]). Perhaps the machine has no clue what it is reasoning about and does not know about the real-life consequences of its decisions, it nevertheless delivers the same juridical quality as a human committee of professionals (6 cases rendered 6 identical judgments).

Searle's [44] final stronghold, then, is the lack of intentionality of a computer. It does not pursue goals and therefore, attaches no meaning to events and actions. The pursuit of goals and in particular the pursuit of self-maintenance and reproduction is what separates an organic system from a physical one. Searle's idea of intentions presupposes goal-directed behavior, resulting into emotions when goals are supported (☺) or obstructed (☹) (cf. [7, p. 494, 463]). In other words,

[2] Robocop meant as a metonym here.

reasoning logically about medical dilemmas from moral principles to deterministic outcomes in Searle's sense can only become humanoid if there is a trade-off with affective decisions, which by definition are intentional. The man in the Chinese Room should be capable of laughter and crying.

In Pontier et al. [36] we did exactly this: Provide the computer with goals to pursue and a mechanism for affective processing [19] and let this interfere with the reasoning from moral principles (i.e., from [1]). Greene et al. [12] state that moral issues with a personal accent ("My daughter is dying") involve more affective processing than moral-impersonal issues ("The patient is dying"). Such differences in emotional engagement modify people's judgments. Our system indicated to be more hesitant to sacrifice one person so to save five others if that person was someone who touched upon preset goals of the system that were not of a moral nature (e.g., that the person was "close by").

Now that we have a machine capable of moral behavior, that can reason, can deal with semantics, shows empathy, that is in pursuit of health goals, but that is unaware of doing it, we could place it in a care situation and confront it with a dementia patient. The moment this person puts herself and others at risk, cannot reason logically, does not understand a thing, shows little empathy with fellow patients or family, in pursuit of anything but health goals (cookies!), and hardly aware of doing it, the care robot might hold the patient for a dumb Searlean machine! How much cognition goes on in there? Get data, decode, execute a process, and respond? That is what a CPU does too. Moreover, the cognition that does go on is so delusive that even Ames would be shocked. How much "proper" judgment is still left? According to what distorted belief system? How much autonomy can be attributed to an information processing unit that merely correlates inputted symbols to other symbols, outputting something that the doctor may interpret as a proper response? "Are you thirsty?" "Yes, I have to urinate." Is this a living human or an organic machine, performing autonomous control over the vegetative system only?

2.2 About Self, About Others

To make moral decisions, a fully functional Caredroid must have beliefs about itself and about others, in our case, the dementia patient. It needs to know the ontology that the patient works with no matter how distorted ("Doctor is a demon"). It needs to know how the patient got to this ontology, by authority of the doctor, priest, or family members or through personal experience ("I saw him rise from the ground"), and it needs to know the biases the patient has, what prejudices and preoccupations ("Cookies!").

This patient ontology is compared to a reference ontology (cf. [18, pp. 314–315]). That ontology is constituted by general socio-cultural beliefs and tells the Caredroid to what extent the patient is "delusive," can be taken seriously, and can be attributed autonomous decision capabilities. Also the Caredroid should know what the origin is of its robot ontology, whose goals it supports, and what

its biases are (what kind of logics, the quality of its sensors, whatever cognitive-emotional modules it has). And through this, the distinction between human and machine becomes obsolete because the problem is boiled down to affordances and the quality of information, not to being alive.

3 Affordances

Affordances in the context of a Caredroid are all the (latent) action possibilities (e.g., services) that the robot offers with respect to healthcare. This could be assistance, monitoring, or conversing. It could be a toy to pet in order to ease feelings of loneliness. There are designed affordances [10, 11], which are the features manifesting in the system or environment—even if they remain undiscovered by the user. For example, a Caredroid may demonstrate a workout but does not afford exercising if the user is paralyzed (cf. vascular dementia). There are also perceived affordances [27, 28], which are those possibilities that the user actually knows about, although much more options may have been designed. The latter remain "hidden" for the user [9]. The user also may falsely perceive certain affordances that the system does not offer [9]. An example is the assumption that the Caredroid will always be ready, which is wrong, because the servos heat up during usage and need a cool down period, which means that you cannot always count on them. False affordances may give rise to plenty of confusion; sometimes for the worse (e.g., Alzheimer patient panics when a robot enters the room: "It is going to eat me!"); sometimes for the best ("It is not a robot; it is a sweet animal: It talks to me").

For moral decisions, the Caredroid should know what affordances the user is capable of recognizing; or better, the Caredroid should store in its ontology what affordances the patient offers. To what extent are someone's capabilities intact or degraded? Are there periods of alertness or does someone suffer from visual hallucinations? Are the affordances designed in the Caredroid perceived at all? What are the false affordances? In "intact" users, the services that the robot has to offer will be recognized as such. If the robot is fork feeding a patient, the patient is supposed to open the mouth. That would be morally "good" behavior of the patient (i.e., beneficence), because it serves her wellbeing. But what if the patient recognizes an Afro fork-comb in the object that is pointing at her and starts doing her hair, smearing out in equal parts the mashed potatoes over her curls? Is this maleficence, according to moral principles? Or is the patient happy with the incredible surface contours of her new styling mold and should we leave it this way? What if during daytime activities, patients are cooking a meal and one of them uses his hair comb as a pasta tong to fish the spaghetti out of the soup? Maleficence because it is unhygienic? Beneficence because of fun? Or is this perhaps a sign of mental restoration as alternate uses are a known strategy for solving problems creatively (i.e., Guilford's Alternative Uses Task, [13])? After all, a gourmet tip to shape garganelli pasta in the right way is to roll it over an Afro comb instead of

using a pasta machine![3] With respect to diagnosis and its consequences for autonomy, are these patients just being creative or downright delusive?

If the Caredroid's ontology would follow Norman's [27, 28] conception of perceived affordances, the patient is to open the mouth because the fork is designed such that it "suggests" that you eat from it. Doing differently, then, would be "blameworthy:" "Don't do this! Stop it!" If the Caredroid's ontology follows Gibson's [10, 11] account, affordances may be there without intentionality. The tree may not grow to build bird nests in but it is extremely suited to carry them yet. The fork may not be designed to comb the hair, but nevertheless. In other words, features of a Caredroid may be designed without any purpose in mind, in fact, they may be a design flaw, but the user may put purpose to it in hindsight [51], so that we find ourselves sitting across Searle's Chinese Room again.

So as we can see, moral decision making may not depend on human or machine agency and may boil down to what an agency affords; *what* it affords is dependent on the way you interpret the offerings. In other words, what is considered an affordance follows from the belief system. How smart, creative, and capable is the user, how smart, creative, and capable is the machine itself? Are the things the machine has to offer convenient for the user (e.g., mechanical love over loneliness) and what is the quality of the information upon which those offerings are made? Is not correctness of information more important than the source being alive? A sign at the road is not alive but I do follow its directions whereas my traveling partner points out the wrong way all of the time. Affordances predict use intentions [29]: I won't ask my traveling partner again.

4 Information

The verification of correctness of information penetrates deeply into the question of truth-finding. It seems that only upon true information we can offer our services and make decisions that are ethical. Being sentenced to jail without being guilty may be right according to the information available at the time but is not seen as ethically just. Thus, controls, tests, and checks circumscribe what is regarded as "correct," relying on the integrity of those safeguards, coming full circle again, because who guards the guardians, who guarantees the method, who controls the controls, a vicious circle of infinite regress, a downward spiral of suspicion. If we cannot trust our methods, measurements, and senses, there merely is belief. Hoorn and Van Wijngaarden [18, p. 308] noted that in the literature, correctness of information supposedly is based on accuracy, completeness, and depth. But that just postpones the question to how much exactitude is needed, when something can be called complete, and how much detail is required?

[3] http://www.circleofmisse.com/recipes/garganelli-31102009.

That things are true is not the same as things having meaning. In healthcare, many things are not true but do carry meaning: placebo drugs, induced illness, health anxiety. If a Caredroid states that "Patient Searle is alive," this statement in Searle's medical dossier has the same truth conditions as "Patient Searle is alive and he is not a robot," although the meanings differ. [4]The truth condition is that Searle has to live for the statements to be true. What being alive means, however, is a matter of beliefs: medically, religiously, or otherwise [16, pp. 19–20]. The conditions under which truth is functioning can only be validated empirically, not verified logically. Hence, truth is attached to structure and syntax, to form, not to content. Thus, the truth of the medical statement "Patient Searle is alive" cannot be dependent on its source, the mindless Caredroid that is a not-alive robot. Truth is logical, not empirical. There is merely the idea of truth. Empirical truth is illogicality. There is only empirical meaning and meanings are connected to the goals of the meaning giver, the interpreter of life. We provide meaning to data but will never find truth in them.

Information becomes logically truthful if a deterministic problem space is assumed and some belief system is accepted with premises that are well-formed, following certain (e.g., moral) principles or rules. This is a very limited approach to the murkiness of daily moral behavior the Caredroid will be confronted with and is mute about semantics, empirical meaning, or "ecological validity." In moral argument, information is empirically meaningful if it satisfies certain goals of the arguer. Although not righteously, people do assume that a statement is logically truthful if it is plausible in a probabilistic sense. The plausibility is extracted from a mixture of belief systems (e.g., medical and religious) with sometimes conflicting and more-or-less lenient guidelines that may be ill-formed, that is, following certain principles or rules but not strictly—preferences and priorities depending on the situation. We may contend that pragmatically, truth values are attributed to data according to probability distributions, which are biased by, for example, what is "ethically good," that is, by something considered useful (see section *Moral priorities*). To put it colloquially: If it serves my purposes (meaning), I apply the rule (logics), and then my moral judgment is truthful. The reasoning may be flawed but its conclusion is convenient.

The integrity of information in the medical dossier that the Caredroid may interface is ascertained by the patient in a pragmatic way. That information should be reliable in the double sense of the word: true and ethical [16, pp. 23–24]. But as we now know, integrity of information is always compromised because truth is but a logical attribute and has little to do with reality, whereas what we call "reality" is a set of assumptions, a mental construct, based on biased sense perceptions [16, pp. 35–36]. The reference ontology of the Caredroid is as much a product of beliefs as the ontology of the delusive patient. The ethical side of this is that the patient supposes not to be lied to; that the source has enough authority to believe its propositions. It remains to be seen if robots are regarded as authoritative

[4] Also see: http://en.wikipedia.org/wiki/Truth_condition.

enough to base health decisions on the information they carry. The problem of an appeal to authority is that its reasoning is flawed ([24]: *argumentum ad verecundiam*). After all, who sanctions the mandate to be an expert ("Because I say so? By virtue of my right?"). Bottom line, if it is not correctness we can rely on by itself, and an appeal to authority is an unsound reason, then the certification of correctness must come from something as shaky as "trust."

5 Trust

How can a robot become trustworthy? How to place your faith in a robot? Trust is an ethical matter [49] but is extracted from highly peripheral cues. Van Vugt et al. [50] found that virtual health coaches that advised on food intake were thought to be more trustworthy when they were obese than when slim. If the user faces critical decisions, an intelligent look is preferred to a funny appearance [38].

Where correctness of information should be the central concern, but indecisive, the trust that deems the centrally processed information correct is derived from peripheral cues [30]. That suddenly makes the question whether the same message is delivered by a human or a robot a non-trivial matter again. It also brings into play the affordances that are perceived. If we see a doctor with a medical dossier, the doctor supposedly is sane and the dossier correct. If we see a robot clown with a baby cloth book, Norman [27, 28] would suppose that the clown is incapable of medical diagnosis and the cloth book does not contain serious medical information—whereas [10, 11] would argue: "Why not?" What if a delusive person provides correct information (according to some belief system)? Probably, the unreliability of the source overrides the correctness of the information. What if an intelligently looking and properly functioning Caredroid works with incorrect information? Probably, the perceived reliability of the source overpowers the incorrectness of the information, quite like the doctor who makes a mistake (which doctors do).

The peripheral cues (e.g., white jacket, stethoscope, diploma on the wall) that are used to build up trust are the stereotypical features that express authority, expertise, and reliability. They are stereotypical because they are repeatedly associated with those inner moral qualities and give rise to the prediction that similar trustworthy behaviors can be expected on future encounters.

Within a belief system, trust comes from predictability of behaviors. If not harmful (non-maleficence), those predictable behaviors persuade into cooperation (cf. [8]) to achieve common or at least non-conflicting goals given the affordances perceived in the other agency. Trust also comes if the Caredroid does something beneficial without expecting anything particular in return [49, p. 15, 24].

Predictable means that probabilities are high for a subset of behaviors to occur and not that of possible other behaviors. In a highly deterministic system such as a computer, this must be easy to achieve, vide the love of autism patients for robots. Not harmful but instead beneficial indicates that achieving goals and protecting concerns are not frustrated or blocked but supported [7, p. 207]. Fortunately, a robot can be programmed such that it expects no gratitude in return. And provided

that skills (affordances, functions) are strengthening or complementing the patient's own affordances, cooperation may happen.

Thus, if an Alzheimer patient sees that a Caredroid has better memory for appointments than she does, and the Caredroid does not stand in the way of other concerns (cookies!), and this process is repeatedly observed, then trust may transpire to collaborate with the robot. In other words, moral behaviors that repeatedly do not harm concerns (non-maleficence), but rather facilitate them (beneficence), are regarded as useful and constitute trust [49]. The point is, a bad memory does not store observed behaviors too well so that the Caredroid has to repeat its behaviors within the little time span of the working memory that the patient has left.

6 Moral Priorities

If we hold on for a second and mull over the simple contention that morality is a function of what we consider useful (cf. Spinoza's "foundation of virtue" in Damasio [3, p. 171], then the prioritization of moral principles [1] should be in line with the rank order of cultural values; in this case: Western values. "Having control" or "being in charge" would be the top priority (i.e., autonomy before anything), followed by being free from pain and threat (this is non-maleficence—nobody is eating me), then the need for nourishment and a mild climate, also socially (beneficence), and justice for all (that is, the group should be fine so to provide protection, on condition that the other three requirements are satisfied first). In other words, ethical reasoning is the legitimization of utility, which is provided by the affordances that an ecological or technological system has to offer [10, 11, 27, 25]. Therefore, Caredroids may afford functionality that make this animal called the user feel in control, keep him clear of danger, bring him the goods, and equally divide what is left over the others. In times of scarcity (no cookies!), this must lead to conflicts and friendships put under pressure [46] because the negotiation of the little supply that is left can only be held along the principle with the lowest priority: justice.

Because in Beauchamp and Childress [1] view, autonomy is the top priority of users and other stakeholders of the care system, a Caredroid should have diagnostics to test the capabilities of the person in front of him. Are the cognitive abilities of the other intact or degraded? If intact, normal prioritization of moral principles applies. If not, autonomy can be overridden by any of the three remaining values, justice included (Fig. 1).

To be frank, there should be some nuance to this position because the prioritization does not have to be so strict. In previous work (i.e., [33]), principles were more balanced and carried different weights. In that study, we also developed a rule to prevent decisions being taken against fully autonomous patients. On the other hand, we instantly questioned whether being fully autonomous actually exists.

A survivalist outlook as outlined above may raise objections from those with a more holistic worldview of being inseparable from all other beings or even from all other physical matter, robots included. An individual who is separated from

Fig. 1 Justice be done [collage created from picture justice (Southernfried, MorgueFile) and Robocop 1 (Verhoeven 1987)]

the collective can easily redistribute responsibilities to the higher (parent) nodes in the hierarchy because autonomy has been taken away from the lower (children) nodes. This is what happens in a Weberian organization structure [14]. It does that to make professional interventions and services predictable, reliable, and safe [14].

Yet, if the belief system tells that a person and her surroundings are manifestations of the same energy, even in a moral sense [21, p. 4], there is something like collective liability, a part-of-whole or "metonymic reflection" on individual and collective behavior. If I represent all else, all else is in me and I am in all else. My deeds act on the collective, the collective acts on me. Thus, I take responsibility for the deeds of others, also of robots. If others do not, robots included, they are "unaware." They are separated ego's incapable of introspection or better, of "part-of-whole reflection." A holistic worldview would work with a different prioritization of moral principles (Table 1). It is the idea that you do not control life but that life works through you. In a (probably Eastern) belief system of inner peace, harmony, and compassion, perhaps beneficence would take the lead as it induces a state of low-arousal positive affect [23], followed by not harming the collective (absence of negative affect, [23]), justice for all, and autonomy finishing last. This of course, is morally a completely different outlook than the one illustrated by Fig. 1 presented for comparison in Table 1. Take notice that Table 1 is a crude approximation of possible moral positions as rank orders may be fuzzier than Table 1 suggests (cf. [33]) and ties may occur if weights are introduced to the hierarchies tabulated here.

Whether we take a Cartesian position of "I know what I'm doing," a Buddhist perspective of "I am aware of being," or Robocop declaring "I am the law" (Fig. 1), the quintessence remains what Searle stressed with "consciousness," which is the point that people are capable of self-reflection; they can do internal diagnosis ("I must be insane!"). They can think about thoughts or be "mindful" of them. Maybe a robot that can adapt its behaviors according to its feedback loops and does self-testing this way might be suspected of another form of inner perception? And the dementia patient not capable of self-reflection perhaps may be

Table 1 Twenty-four possible priory configurations of autonomy, beneficence, non-maleficence, and justice, dependent on the belief system

By the power invested in me

1	A	A	A	A	A	A
2	B	B	N	N	J	J
3	N	J	B	J	B	N
4	J	N	J	B	N	B
			↑			

Survivalist outlook

Be good

1	B	B	B	B	B	B
2	A	A	N	N	J	J
3	N	J	A	J	A	N
4	J	N	J	A	N	A
				↑		

Buddhist view

Don't harm

1	N	N	N	N	N	N
2	A	A	B	B	J	J
3	B	J	A	J	A	B
4	J	B	J	A	B	A

Justice be done

1	J	J	J	J	J	J
2	A	A	B	B	N	N
3	B	N	A	N	A	B
4	N	B	N	A	B	A
				↑		

Robocop

Note: No ties assumed.

said to be (morally) "comatose" or put more mildly, "unaware"? Thus, inner perception, awareness, or conscience, or whatever you want to call it, is an agency's self-test circumscribing what is regarded as "correct" information to base a moral decision upon, its certification stamp being "trust," a false syllogism of authority, flawed but convenient, notwithstanding.

7 Responsibility Self-Test

The previous section hinged on two innate tendencies ostentatiously portrayed and alluded to throughout cultural history, all being appearances of good and evil, sin versus virtue. In Abrahamic religions, it would be Satan against God, the Greeks contrasted Dionysus with Apollo, Descartes separated passion from reason, Freud distinguished *id* (instinct) from *superego* (conscience), and Buddhist teachings say that ego detaches itself from awareness. As far as I am concerned, these are graphic descriptions of the brain's evolutionary architecture [31, p. 91]. It has an older mechanistic part, which it has in common with physical nature. That part is taken control of by the vegetative system, which executes genetically hard-coded behaviors (cf. a virus). Through the genome, a soft transition to the animalistic part is made. These are behaviors that certain animals also have, for instance, memory and learning (soft-coded information), communication, organization, and exchange of information across group members. In humans, that would count as language. This part of the brain can be retrieved to organic nature (e.g., cats and ants) other than what we think makes us human, which are the higher cognitive functions residing in the youngest brain lobes, the things we call "spiritual:" intelligence, creativity, wisdom (or awareness, for that matter).

The idea of course is that our animalistic lust is kept in check by our reason, conscience, or awareness of that "lower" behavior. If the higher cognitive functions (i.e., being an angel) fail to control the lower functions, we start behaving like animals (cf. the snake).

What should the Caredroid be aware of, then? We have been discussing a number of factors. The first was that of Agency (Ag), which could be human or robotic ($Ag_{(h,\,r)}$), then we examined the Beliefs the agency has about self and others ($B_{(s,\,o)}$), whether the (mental) affordances of both agents are regarded as intact or not ($Af_{(i,\,d)}$), to what degree the Information they express seems to be correct or incorrect ($I_{(c,\,i)}$), whether Trust in the source is high or low ($T_{(h,\,l)}$), and what Moral priorities apply to the situation ($M_{(1,\,\ldots,\,24)}$). Putting a value to each of these variables selects one path in a nested factorial design of:

$$Ag_{(h,\,r)} * B_{(s,\,o)} * Af_{(i,\,d)} * I_{(c,\,i)} * T_{(h,\,l)} * M_{(1,\ldots,24)}$$
$$= 2 * 2 * 2 * 2 * 2 * 24 = 768 \text{ constellations to base a moral decision upon.}$$

To navigate this wide grid of moral positions, I want to try a responsibility self-test that comes in 7 steps. It is valid for both individual and collective liability,

depending on an agency's affordance of inner reflection or more profane, a self-test that handles one or more of the said priority constellations of Table 1 as its yardstick.

We will race a number of agencies over the 7 hurdles and see who survives. The winner is the morally most responsible one that is accountable for its deeds—this "game" approach loosely follows the lead of Lacan [22]. The first step would be to see if an agency can act irrespective of the awareness of doing so.

1. **I do something**

This is to be taken as an act without awareness. One could imagine that a worm, an old-school robot, a cat, a patient with advanced dementia, as well as a sane person can pass this check, because all can do something. That may be on their own behalf or predicated by others, with or without knowing its own way of conduct, but at least they can act. In Moor's [25] taxonomy of moral agency, *normative agents* that can prove a theorem, *ethical impact agents* that are supposed to alleviate human suffering, and *implicit ethical agents* that take safety and legislation into account would pass this hurdle but not the next because they merely behave according to certain principles without knowing that those principles are "ethical."

2. **I know what I did was bad, good, or neutral**

This time, the agency is aware of its behavior as well as the rules of conduct, the rules of engagement under which that behavior is executed. It does not reflect about those rules, it operates within the boundaries of those rules. Those rules can be imposed upon by others; they may be one's own rules. A cat knows, for example, that it is not allowed to steal the meat from the kitchen. It fears punishment. The same is valid for certain dementia patients ("Don't steal cookies!") as well as a sane person. But also a robot with moral reasoning implemented may have a feedback loop built in that can judge whether certain principles were enforced or violated by its own actions. In Moor's [25] classification, *explicit ethical agents* would qualify for this test because they state explicitly what action is allowed and what is forbidden. Hurdle 2 is decisive to separate the sane and lightly demented people from the severely demented, who no longer have any clue about good or bad and start making nasty remarks, cold and insensitive, doing things that are inappropriate (e.g., aggressive behaviors).

3. **I know that I know I was bad, good, or neutral**

Here we enter the realm of having meta-cognitions about knowledge. The agency becomes morally conscious and can agree about the rules although disobeying them. Or, one can disagree about the rules and nonetheless comply with them. The agency now is aware of having knowledge about its actual behavior and that certain rules limit the behaviors that could possibly be executed. This is Moor's [25] *full ethical agent,* having consciousness, intentionality, and free will. It is a Cartesian position of thinking about thinking. It is Searle's criterion of being a conscious agency. But it also reflects a Buddhist position once we replace "thinking" by "awareness." In that case, it is not so much "Thinking, therefore I am" but "Aware that I am."

Checking on hurdle 3 may actually discern light dementia and "correctable" behavior from advanced dementia and lack of liability. As far as I know, there is no robot yet that can take hurdle 3. It would require a feedback loop about its feedback on its behaviors: The robot should record how well the processes perform that monitor its moral actions.

4. **I also know why I was bad, good, or neutral**

From this point on, it is not about logical verification alone any more but also about empirical validation. At this level, the agency has knowledge about the way its behavior was tested. How it came to know what happened at hurdle 3. Number 4 is an epistemic examination in how far some judge, referee, auditor—which could be the agency self—can be trusted in matching the agency's behavior against some law, rules, agreements (individual or collective), in an empirically valid or meaningful way, according to belief.

Only a sane person can take this hurdle, because the agency should have knowledge of which rules of conduct, rules of engagement, social contract, or terms of agreement are relevant in the situation at hand, picking one or more priority constellations from Table 1 as appropriate to goals and concerns of multiple stakeholders. It requires an estimate of how well the Caredroid senses its environment, perspective taking, and being able to weigh aspects in a situation with different sets of goals in mind.

5. **That I am aware of me knowing what I did and why it was wrong, right, or neutral—even if I disagree—means my higher cognitive functions are intact**

This is the Cartesian and Buddhist stance taken in unison with empiricism as a self-test on the agency's affordances. By overseeing the entire moral evaluation process, the agency can decide whether it has intelligence, creativity, and/or wisdom.

6. **My "higher" cognitive functions are supposed to control my "lower" functions but failed or succeeded**

In many cultures throughout history, this trade-off between good versus evil has always been the overture of the final verdict (see 7). Point 1 up to 5 were there to feed or load this opposition and hurdle 6 does the definitive weighing of being capable of handling the animalistic tendencies.

7. **Therefore, I am responsible and can be punished/rewarded or remain as is**

The final verdict. Any agency that went through 6 can be held responsible for 1 and automatically will go to 7. That agency is entitled to receive what is waiting for him or her as agreed upon in a given culture.

In answering Searle's Chinese Room dilemma, then, steps 3 and 4 are to be modeled, formalized, and implemented before we can even think of robot "consciousness," and following from that, machine liability. It also shows that moderately demented patients are only partially and in severe cases not responsible for their behaviors. In that respect (and with all due respect), the more than moderately demented patients are comparable to mammals (e.g., cats) and robots that have some sort of command over what is obliged or permitted but have no

meta-cognition about that knowledge. Dementia in its final stage is even below that level. Cats, stota moral robots, or more than lightly demented elderly are in Searle's sense "unconscious" or in a Buddhist sense "unaware" of their own dealings. That is what makes them "primitive" or "animal-like," meaning that the youngest human brain functions or highest cognitive functions are defunct or missing.

8 Discussion

This chapter mixed ethical issues with epistemic considerations from the assumption that judgments of "morally good" are intertwined with "really true." When a dementia patient is confronted with a care robot that has reliable knowledge about the patient (e.g., according to a medical dossier), then we have a real person with delusions facing a virtual person with a realistic take on the world. Now, who should be controlling who? Should the robot comply with the demand of human autonomy and obey every command that the patient gives [41]? Or should it overrule certain proposals by the patient to protect her (and others) against herself? It all depends on what is regarded as the proper framing of the situation (the beliefs): The fiction inside the real person's head or the reality inside the fictitious person's CPU? Bottom line, what is more important: The correctness of information or the trustworthiness of the information carrier (the source)? And what would correctness be then?

Everything we do and stand for comes from belief, and morality is no exception. You cannot get it from logic, because the premises from which the logic start are empirically bound and hence, inaccessible epistemically. Put differently, it is incontrollable whether information is correct. Because truth telling is unknowable, it becomes intertwined with moral goodness: Trust is laid in the source that conveys the information. This state of affairs is no different for a human as it is for a robot. The doctor who tries to understand an Alzheimer patient is comparable to the user trying to find out what goes on in a robot's microchips. In the Chinese Room, the actual position of the information processer on the human-robot continuum will never be known. If an observer does make a choice (the reduction to a single eigenstate) biased perception made it so.

Trust in whatever source comes from peripheral cues that indicate expertise and authority on a particular matter, such as a white lab coat and glasses. They are cues to affordances that are perceived in the agency such as the ability to test information and high intelligence.

The list of four moral principles is the empirical aspect or "meaningful part" of moral reasoning. Contingent upon the belief system (e.g., Cartesian or Buddhist), what an agency understands under "autonomy" or "beneficence" may differ. Customarily, the meaning attached to a moral notion is related to goals and concerns of the individual and its community. But also the rank order (Table 1) or more sophisticated, the weighing of the principles, depends on goals and concerns. Thus, the reasoning may be straight but the semantics are biased.

Psychological biases are inescapable. Even the most rational of choices has emotional biases because the contents that are reasoned about pertain to empirical goals and concerns. One could even argue that moral reasoning without affect is to be lulled into a false sense of security.

That is why a medical ethical reasoning machine cannot do much more than "to know thyself," having meta-knowledge about its (reference) ontology (i.e., its belief system), epistemology (i.e., robot sensing and testing), and cognitive biases (i.e., the cognitive flaws in the system and the goals it has to achieve, for example, monitoring the patient). That is why for the moral domain I developed a self-test by which the agency can determine whether it is responsible for its acts or not.

8.1 Autonomae Servus

There is a difference between form, formalization, mechanism, syntax, structure, system, logics, verification, and truth on the one hand, and meaning, semantics, experience, empirical relevance, validation, and truth conditions on the other. It is what linguistic Structuralists (e.g., De Saussure [4, p. 121]) would have called the combinatory syntagmatic axis (i.e., the grammar) that needs to be filled by selections from the paradigmatic or associative axis (i.e., the lexicon). Whereas the formal part is relatively fixed, the meaning part is fluctuating. It is a problem of reference: What do the signals (e.g., words) stand for? This is not a logical but an empirical question. De Saussure would ask: "What is signified by the signifier?"

In deterministic problem spaces the logics of robots equals or even emulates that of humans (e.g., [33]). In probabilistic cases, where logics fail, the robot's guess is as good as any human's. If in addition you can make it plausible that someone hardly can decipher dealing with a human or a robot [35], then robots can be applied to many healthcare tasks. The only thing missing would be the "meaningful" aspect, sharing the belief system with the patient, "what the words stand for," which is problematic in caretaker-patient transactions as well. Even so, a Caredroid becomes more organic in its behavior once it is driven by goals [19] because it will attach more meaning to a signal in the sense of consequences for its pursuit of a patient's wellbeing.

Where logics and semantics grow together by introducing intentionality to the machine, the distinction between human and machine ethical reasoning cannot be made any more except for peripheral cues in the interface, such as outer appearance, bodily warmth, tone of voice, etc. The distinction is already vanished in comparison with a patient that hardly has any conscience left (cf. advanced dementia). If we integrate human with machine circuitry and live as cyborgs, making that distinction becomes irrelevant. In assuming a Buddhist perspective, we could allow to ourselves that we are made from the same matter; human consciousness intermixed with a machine's self-test on morality.

That leads us to the creation of the *Autonomae Servus,* the autonomous slave. We want the reasoning to be autonomous but the content to be serving our

purposes: The human as master computer, the robot as his autonomous slave. The Caredroid may act autonomously because of its affordances (e.g., reasoning capacities) but is obedient to our goals of non-maleficence, beneficence, and justice. It moreover will be compassionate about our feelings of personal autonomy.

When the Hong Kong researchers finally took their metal-cutting shears, they saw that they had not kept Searle inside the Chinese Room but an old Chinese practitioner of Qigong. He wore a yin-yang symbol around his neck and looked more at an agency's functions than at its anatomy. When he came out, he orated that no life is complete without suffering from loneliness, illness, and decease. Most patients will not see it that way, he said, and try to cast out the bad experiences, but this is coping through ignoring. A robot could teach us to use those experiences for creation, the practitioner stipulated, as an alien source of information that can be integrated with known practice.

He then rummaged around in one of his filing cabinets, and next to a small flask there was a photograph. It showed a healthy 60-year old, walking his puppy dog called Mao while he was chatting with the neighbors along the way [5, pp. 63–64]. The next picture showed him as a 70-year old. He held a walker rollator and could hardly control the dog anymore. He admitted to have felt ashamed of the rollator and did not use it. In not going out of the Chinese Room anymore, he became lonely. Today, as an 80-year old with light dementia, he could not keep the dog any longer and ate it. The bones were spread around the floor in equal distributions. Now he was thirsty. He did not want a cat. He hated cats because they could see spirits in the dark [5, p. 65].

The Hong Kong researchers felt sorry for the old man and gave him a Hanson's Robokind Alice to ease the loneliness. That machine had a creativity module called ACASIA implemented ([15, 17], Chap. 4) that suggested to put her, the robot, in a children's wheelchair (Fig. 2). Now the old man strolled away behind the wheelchair, a support similar to a rollator,[5] without having to be ashamed of it. In fact, he was proud that he took care of the handicapped robot. Out on the street, he attracted a lot of attention and had quite a number of chats about his poor but cute Caredroid, and by the way, the Caredroid had a navigation system telling grandpa how to get back home again (cf. LogicaCMG's rollator navigator).[6]

As a kind of addendum, I would like to emphasize that people have their mechanistic side; physiologically (e.g., dehydration from a loss of electrolytes: "I am thirsty") as well as mentally (i.e., automated processes, routine behaviors: "Bring water glass to the mouth"). Are impaired people automata, then? One could argue that the more cognitive functions become impaired, the more people start to resemble automata. Giving creativity, intelligence, language, and memory back to the impaired through Caredroids, is making them more human again. Caredroids are factory robots made loveable [6]. They give something more relatable to their organic coworkers than today's practice of disembodied limbs put on a friendly face (ibid). We make computers understand how humans work. In dealing with

[5] Courtesy Robert Paauwe, personal communication, Oct. 15, 2013.

[6] http://www.camera.vu.nl/news/2007/021107news.html.

Fig. 2 Hanson's Robokind Alice in a wheelchair

moral dilemmas, they can share as a human-android team, the burden of potential information overload for the patient, forming a system of multi-agencies that can exploit the information universe to the fullest. As "human-technology symbionts" [2, p. 3], impaired patients will be able to explore more alternatives, exclude more dead ends, reckoning with more situational constraints. It would be morally unfair not to compensate the impaired with loveable automata that care.

Acknowledgments This chapter is part of the SELEMCA project (Services of Electromechanical Care Agencies), which is supported by the Creative Industries Scientific Program (CRISP) of the Ministry of Education, Culture, and Science, grant number NWO 646.000.003.

References

1. Beauchamp TL, Childress JF (2001) Principles of biomedical ethics. Oxford University, New York
2. Clark A (2003) Natural-born cyborgs: minds, technologies, and the future of human intelligence. Oxford University, New York

3. Damasio A (2003) Looking for Spinoza: joy, sorrow, and the feeling brain. Harcourt, Orlando, FL
4. De Saussure F (1916/1983) Course in general linguistics (trans: Harris R). Duckworth, London
5. Eberhard W (1986) A dictionary of Chinese symbols. Routledge & Kegan Paul, London
6. Fingas J (2012) Rethink delivers Baxter the friendly worker robot, prepares us for our future metal overlords (video). Retrieved 13 Jan 2014 from http://www.engadget.com/2012/09/19/rethink-delivers-baxter-the-friendly-worker-robot/
7. Frijda NH (1986) The emotions. Cambridge University, New York
8. Gambetta D (2000) Can we trust trust? In: Gambetta D (ed) Trust: making and breaking cooperative relations. Blackwell, Oxford, UK, pp 213–237
9. Gaver WW (1991) Technology affordances. In: Robertson SP, Olson GM, Olson JS (eds) Proceedings of the CHI '91 SIGCHI conference on human factors in computing systems. ACM, New York, pp 79–84. doi:10.1145/108844.108856
10. Gibson JJ (1977) The theory of affordances. In: Shaw R, Bransford J (eds) Perceiving, acting, and knowing. Towards an ecological psychology. Wiley, Hoboken, NJ, 127–143
11. Gibson JJ (1979) The ecological approach to visual perception. Erlbaum, Hillsdale, NJ
12. Greene JD, Sommerville RB, Nystrom LE, Darley JM, Cohen JD (2001) An fMRI investigation of emotional engagement in moral judgment. Science 293(5537):2105–2108. doi:10.1126/science.1062872
13. Guilford JP (1967) The nature of human intelligence. McGraw-Hill, New York
14. Harrison S, Smith C (2004) Trust and moral motivation: redundant resources in health and social care? Policy Polit 32(3):371–386
15. Hoorn JF (2002) A model for information technologies that can be creative. In: Hewett TT, Kavanagh T (eds) Proceedings of the fourth creativity and cognition conference. ACM, Loughborough, UK, New York, pp 186–191
16. Hoorn JF (2012) Epistemics of the virtual. John Benjamins, Amsterdam, Philadelphia, PA
17. Hoorn JF (2014) Creative confluence. John Benjamins, Amsterdam, Philadelphia, PA
18. Hoorn JF, Van Wijngaarden TD (2010) Web intelligence for the assessment of information quality: credibility, correctness, and readability. In: Usmani Z (ed) Web intelligence and intelligent agents. In-Tech, Vukovar: Croatia, pp 305–324
19. Hoorn JF, Pontier MA, Siddiqui GF (2012) Coppélius' concoction: similarity and complementarity among three affect-related agent models. Cogn Syst Res 15–16:33–49. doi:10.1016/j.cogsys.2011.04.001
20. Horrobin D (2001) Something rotten at the core of science? Trends Pharmacol Sci 22(2):1–22
21. Keown D (2005) Buddhist ethics. A very short introduction. Oxford University, New York
22. Lacan J (1945/2006) Logical time and the assertion of anticipated certainty: a new sophism. In: Écrits: the first complete edition in English (trans: Fink B, Fink H, Grigg R). Norton, New York, pp 161–175
23. Lee Y-C, Lin Y-C, Huang C-L, Fredrickson BL (2013) The construct and measurement of peace of mind. J Happiness Stud 14(2):571–590
24. Locke (1689) An essay concerning human understanding. Holt, London
25. Moor JH (2006) The nature, importance, and difficulty of machine ethics. IEEE Intell Syst 21(4):18–21
26. Nilsson NJ (1984) A short rebuttal to Searle [unpublished note]. http://ai.stanford.edu/~nilsson/OnlinePubs-Nils/General%20Essays/OtherEssays-Nils/searle.pdf. Accessed 9 Oct 2013
27. Norman DA (1988) The design of everyday things. Doubleday, New York
28. Norman DA (1999) Affordance, conventions, and design. Interactions 6(3):38–43
29. Paauwe RA, Hoorn JF (2013) Technical report on designed form realism using LEGO Mindstorms [Tech. Rep.]. VU University, Amsterdam
30. Petty RE, Cacioppo JT (1986) Communication and persuasion: central and peripheral routes to attitude change. Springer, New York
31. Pfenniger KH (2001) The evolving brain. In: Pfenniger KH, Shubik VR (eds) The origins of creativity. Oxford University, New York, pp 89–97
32. Poincaré H (1913) The foundations of science. Science, Lancaster, PA

33. Pontier MA, Hoorn JF (2012) Toward machines that behave ethically better than humans do. In: Miyake N, Peebles B, Cooper RP (eds) Proceedings of the 34th international annual conference of the cognitive science society, CogSci'12. Cognitive Science Society, Sapporo, Japan, Austin, TX, pp 2198–2203
34. Pontier MA, Siddiqui GF (2008) A virtual therapist that responds empathically to your answers. In: Prendinger H, Lester J, Ishizuka M (eds) 8th international conference on intelligent virtual agents, LNAI 5208. Springer, Berlin, GE, pp 417–425
35. Pontier MA, Siddiqui GF, Hoorn JF (2010) Speed dating with an affective virtual agent. Developing a testbed for emotion models. In: Allbeck JA, Badler NI, Bickmore TW, Pelachaud C, Safonova A (eds) Proceedings of the 10th international conference on intelligent virtual agents (IVA) Sept. 20–22, 2010, Philadelphia, PA, Lecture Notes in Computer Science (LNCS) 6356. Springer, Berlin, Heidelberg, DE, pp 91–103
36. Pontier MA, Widdershoven G, Hoorn JF (2012) Moral Coppélia—combining ratio with affect in ethical reasoning. In: Pavón J et al (eds) Lecture notes in artificial intelligence, vol 7637. Springer, Berlin-Heidelberg, DE, pp 442–451
37. Porter GE (2004) The long-term value of analysts' advice in the wall street journal's investment dartboard contest. J Appl Finan 14(2):1–14
38. Prakash A, Rogers WA (2013) Younger and older adults' attitudes toward robot faces: effects of task and humanoid appearance. In: Proceedings of the human factors and ergonomics society annual meeting September 2013, vol 57(1), pp 114–118. doi:10.1177/1541931213571027
39. Reeves B, Nass CI (1996) The media equation: how people treat computers, television, and new media like real people and places. Cambridge University, New York
40. Scassellati B, Admoni H, Matarić M (2012) Robots for use in autism research. Ann Rev Biomed Eng 14:275–294
41. Schreier J (2012) Robot and Frank [movie]. Samuel Goldwyn Films, New York
42. Schrödinger E (1935/1980) Die gegenwärtige Situation in der Quantenmechanik, *Naturwissenschaften* 23, 807. In: Paper, proceedings of the American Philosophical Society (trans: Trimmer JD) The present situation in quantum mechanics: a translation of Schrödinger's "Cat Paradox", vol 124, p 323. Available at:http://www.tuhh.de/rzt/rzt/it/QM/cat.html
43. Schrödinger E (1944/2010) What is life? Mind and matter. Cambridge University, New York
44. Searle JR (1980) Minds, brains, and programs. Behav Brain Sci 3(3):417–457
45. Shibata T, Wada K (2011) Robot therapy: a new approach for mental healthcare of the elderly—a mini-review. Gerontology 57:378–386. doi:10.1159/000319015
46. Silver A (1989) Friendship and trust as moral ideals: an historical approach. Eur J Sociol 30(2):274–297. doi:http://dx.doi.org/10.1017/S0003975600005890
47. Tetlock PE (2006) Expert political judgment: how good is it? How can we know?. Princeton University, Princeton
48. Turing AM (1950) Computing machinery and intelligence. Mind 59(236):433–460
49. Uslaner EM (2002) The moral foundations of trust. Cambridge University, New York
50. Van Vugt HC, Konijn EA, Hoorn JF, Veldhuis J (2009) When too heavy is just fine: creating trustworthy e-health advisors. Int J Hum Comput Stud 67(7):571–583. doi:10.1016/j.ijhcs.2009.02.005
51. Ward TB, Smith SM, Finke RA (1999) Creative cognition. In: Sternberg RJ (ed) Handbook of creativity. Cambridge University, Cambridge, pp 189–212
52. Weber M (1922/1947) The theory of social and economic organization (trans: Henderson AM, Parsons T). Free, New York