From traditional man-machine interfaces to empathic machines: Toward coadaptation

Laurence Devillers,
professor of AI, Sorbonne University/LIMSI/CNRS, member of the CNPEN & GPAI on the future of work

Abstract:
The multidisciplinary field of "affective computing" groups three sorts of technology: the recognition of human emotions; deliberation (reasoning and the making of decisions based on the information gathered); and the generation of emotional expressions. The sphere of emotions, which we might imagine as being specifically human, is encompassing machines as the latter are increasingly endowed with human capacities. When a robot is mistreated, a human being might feel empathy toward this artificial agent, which is merely simulating emotions and cannot suffer. The machine is programmed; it lacks intentionality and has no desires of its own. The coming of empathic robots raises ethical, legal and social questions…

To understand the particular characteristics of empathic robots, let us recall that a robot is characterized by three interacting features: gathering data via sensors, interpreting the data by using software, and moving within, and acting on, its environment. In addition, a robot might be anthropomorphic or be able to use language. A "chatbot" or "conversational agent" is a system for conversations between people and machines. Social robotics is trying to create robots endowed with social skills, among them the ability to take part in a conversation and thus replace people for certain tasks (DUMOUCHEL & DAMASIO 2016). A last point to emphasize: these social robots might be endowed with artificial empathy (AE).[1]

Empathy is an emotional response to a very particular situation. It is a personality trait, the ability to feel an appropriate emotion in response to an emotion expressed by someone else and to clearly distinguish the other person's emotion from one's own (*e.g.*, Your child is not well, and you feel the same way but do not suffer physically). People have the ability to project themselves. In children, affective empathy appears at the age of one; and cognitive empathy, later, toward the age of four and a half. A more "mature" empathy that includes a sense of reciprocity and serves to construct a sense of morals and justice arises between the ages of eight and twelve.

Robots and chatbots, like Siri on telephones or Google Home, might delude us into thinking that machines are empathetic when they are able to identify the emotions of their human contacts, to reason by using the thus detected emotions, and to produce emotions through their facial expressions, gestures, postures and acoustics. Detecting a person's emotions might lead the machine to change its strategies for responding. It might, for example, respond, "I'm sad too" when it has detected sadness in someone. This "deceit" could lead us to believe that the robot has empathy. However this is not really empathy, since these machines do not have a "phenomenal consciousness", *i.e.*, the experiences characteristic of a person's "life" or "feelings".

[1] This article has been translated from French by Noal Mellott (Omaha Beach, France). The translation into English has, with the editor's approval, completed a few bibliographical references. All websites were consulted in April 2021.

Phenomenal consciousness, unlike cognitive consciousness, is associated with a qualitative experience, such as the feelings of pleasure or pain, of hot or cold, etc. Phenomenal consciousness cannot be reduced to the physical or physiological conditions of its appearance. Nor can it be separated from the individual's subjectivity. There is a major difference between scientific descriptions of consciousness, which refer to publicly observable behavior or to operations in the brain, and the phenomenal consciousness specific to people as such. Antonio Damasio (1994), a neuroscientist, has proposed a new view of how emotions are expressed through the close interrelations between body and brain during the perception of objects. The "body" is a key idea. A person is not reduced to thinking, to self-awareness. A person is also a body whereby he/she is located in the world.

The mystery surrounding sciences such as artificial intelligence (AI), cognitive neuroscience and affective intelligence must be dissipated so that these artefacts, robots or conversational agents can be put to better use in society. Building computer models of affects raises the question of the societal consequences of living in an everyday environment with pseudo-affective objects (DEVILLERS 2017). We can imagine ethical principles for distinguishing artificial from human agents.

Affective computing

Affective computing, which has its origins in the work done by Rosalind Picard (1997) at MIT, brings together three fields of technology — the recognition of human emotions, reasoning, and decision-making — that use information and generate the expression of emotions. Affective computing is inherently multidisciplinary. Recognizing the social signals and, in particular, the emotions conveyed on faces or in voices is indispensable for communication with human beings and integration in society.

The consensual definition is that emotions are a reaction to an event or situation, real or imaginary that has several facets or components. Three components are usually accepted as essential to an emotional reaction: the emotional experience (subjective feeling), the physiological reaction and the emotional expression (facial, vocal or postural). During communications in objectively similar contexts, expressions change as a function of a set of sociocultural rules, which vary from one culture or social group to another. Some social situations might require suppressing certain expressions, whereas others might require showing other expressions or even exaggerating them. A spontaneous expression might be hidden when it is not deemed acceptable in the context. The social context can affect the expression of emotions as a function of the position and objectives of the person emitting the emotions. During communication, an individual can use his/her emotional expressions so as to influence — more or less unconsciously and more or less involuntarily — the other party's reactions.

Machines are going to increasingly have vocal exchanges with us in everyday situations. Conversational agents and social robots already have embedded systems of detection, reasoning and the expression of affects. They are, with a wide margin of error, able to interact with us. They are invading our privacy. In the United States, a household has up to six Amazon Alexa or Google Home speakers, one in each room. The market is enormous. These virtual assistants could become part of everyday life: keeping watch over our health, educating, helping and entertaining us, in short, looking after us. The machines for these tasks have been designed as digital companions, assistants or monitors.

Communications with a machine mainly involve exchanging information despite, too often, an "incommunicability" in circumstances when the communicated information contains no message or the receiver (*e.g.*, the robot) cannot decode the information contained in the message. Machines are not yet about to have semantic skills sufficient for holding a conversation and sharing ideas. However they will soon be detecting when we are restless or stressed, and they might even detect some of our lies.

Human empathy toward machines

According to the "*media equation*" (NASS & REEVES 1996), our social expectations are the same when communicating with artificial entities as with people. We unconsciously assign these entities rules of social interaction. Anthropomorphism is the attribution of the behavioral or morphological characteristics of human beings to objects. Given this innate but also socially reinforced reflex, an object seemingly in pain might arouse empathy.

Experimental studies have demonstrated this projection of affective and probably empathic reactions onto artificial entities, such as a damaged toy robot. Researchers observed that human beings felt empathy toward the mistreated robot. This empathy, though less intense than what is felt for mistreated human beings, is not aroused toward inanimate objects. As recent research based on brain imaging has shown, individuals respond in an amazingly similar way to emotional images of human beings and artificial entities. Since we cognitively represent these entities to be human, it should come as no surprise that we emotionally react toward them as toward human beings. It is not clear, however, whether the cognitive representations of robots and of human beings are of the same complexity. Fan Hui, the Go champion who trained DeepMind AlphaGo, explained that playing against a machine was somewhat like playing against oneself, since one was projecting one's own emotions onto a machine, which reflected them back like a mirror.

Nao, Pepper and Romeo are social robots capable of reflecting our emotions, but they are not, strictly speaking, empathic. Paro, a robot seal developed in 1993 in Japan, was brought to the market there in 2005 and then in the United States in 2009 (with FDA certification as a "therapeutic robot"). When Paro is used in nursing homes for the elderly, its reactions are not empathic but are more like a pet animal's expressive behaviors. During our tests in a nursing home with Nao and Pepper, which detect emotions and make adjustments to them, the elderly (more than fifty persons with an average age of 85 but who were not under custody) reacted with curiosity and amusement (GARCIA *et al*. 2017). However they did not think that the robots understood them even though it detected certain behaviors.

A two-year-old child knows that a teddy bear is not alive but talks to it as if it were. Psychologists refer to dolls and teddy bears as "transitional objects". A child chooses such an object, usually soft to touch, to cope with anxiety. It is the child's favorite and first possession. Not perceived as being part of the mother, nor as an inner object, it enables the child to move from the subjective toward the objective sphere. Paro, a furry robot, can be seen as a transitional object.

Social robots are primarily technical objects that record our data in order to transfer them (*e.g.*, to a doctor). Personal assistant robots fit into an ecosystem with several other parties: the family, care-givers, doctors.

Behavioral disorders in dealings with robots have been the topic of psychological and psychiatric studies (TISSERON 2015). Kate Darling (2016), a researcher at MIT's Media Lab, studied people's empathic reactions in their dealings with robots (in particular, mistreated robots). Her idea is that robots should be granted legal protection like animals. Nonetheless, animals are living beings who feel emotions and suffer whereas this is absolutely not the case of robots, which only simulate emotions.

Machine emotions

Enabling machines to interpret and simulate emotions is indispensable for building systems capable of social interactions and of better communications with people. Many applications have the goal of helping people who are dependent, owing to their advanced age or to degenerative pathologies. The emotional sphere, which we might have thought was specifically human, is encompassing machines, which are coming closer to having human faculties. Hesitancy or a sigh from a machine might give the impression that a robot is alive. Donald Davidson (1980), an American philosopher, has described "anomous monism" as a union expressed by two different languages without translation. There is no causal relation between mind and body. The more a machine looks fragile, the more we can humanize it and be moved by it, even though it is not, strictly speaking, empathetic.

Spinoza, in particular his writings on ethics (1677), is a source of inspiration for explaining our contemporary world and mind-body relations. According to this philosopher, the organism makes itself. The affects are places where body and mind are united. Negative affects, like hate, are sources of alienation if we fall to them or of liberty if we understand the processes underlying them. Thanks to explorations of the brain, we can now prove Spinoza's propositions, which ran counter to the commonsense of his time. Bodily expressions precede feelings; and, as we learn from Spinoza, body and mind are mixed. However machines still do not have a body (in the sense of guts, hormones and skin). They do not have intentions, their own desires or sense of pleasure (what Spinoza meant by *conatus*). As a consequence, a living being can be defined as being autonomous and having the possibility of reproducing itself.

At present, human beings always program robots to be (relatively) autonomous. Programmed learning offers a variable degree of freedom to the machine. The ultimate quest of researchers in AI is to endow a robot with the capacity for learning on its own, in interaction with the environment and people. If robots learn by themselves, it would be well worthwhile to teach them the common moral values of life in society. Such a faculty would signal a technological breakthrough and a legal disruption. It would raise many an ethical question. Such robots might be somewhat creative and autonomous in their decision-making on condition that we program them for that.

How narcissistic to want to copy human intelligence in a machine! After all, what do we know about our own intelligence? We do not know what underlies thinking, and we have no awareness of the autonomy of some of our organs. We are aware of but a small proportion of our perceptions or our brain's activities. There is no word more polysemous and subject to interpretation than "consciousness". For some people, it refers to self-consciousness; but for others, to the awareness of others or to the phenomenal consciousness, moral conscience, etc.

In line with a materialistic philosophical conception of life, a computer and the human brain can be seen as comparable systems capable of processing information. The most effective AI techniques, like deep learning, rely on a simplified model of neurons integrated in a machine with discrete states and simulated on a computer. The depth is indicated by the number of layers hidden in this model's architecture. At present, this is still very far from the complexity of living organisms.

Current AI systems can calculate correlations of facts (thanks to deep learning, for example) and are capable of making decisions and learning but without any consciousness of what they are doing. Some robot prototypes already have the "seeds" of a level of consciousness, comparable to what Stanislas Dehaene (2014) has described, that are simulated through procedures of introspection and knowledge sharing. Nevertheless, these machines are not conscious like human beings. They have neither a moral conscience nor the phenomenal consciousness associated with qualitative experiences (such as the sensation of hot or cold, the feeling of anxiety, etc.), since they do not have guts or feelings (unless, of course, the latter are simulated).

Equipped with sensors for pain and pleasure (ASADA 2019), the first learning, communicating robots interact through simple procedures. Their sensors, cameras or microphones enable them to associate a face and voice with the expressive signals they pick up. In turn, their expressions (such as the sounds they emit) enable people to understand their "states". To produce rapid effects, physical actions are necessary. If petted on the head, the robot makes a positive association with the person it sees; if bumped on the head, a negative association. Such a machine learns and adjusts its behavior to the context; but its interpretations of the context are still very limited. For example, a robot assisting students in dental surgery should be well informed about the grimaces signaling pain and anticipate the dentist's gestures so as to warn about a nearby nerve (and thus a possibility of pain).

But do robots have to be as much as possible like human beings? An artificial consciousness equipped with feelings, thoughts and free will and without being programmed by human beings has, given current computer architectures, little chance of spontaneously emerging.

Conclusion

In the long run, everyday life with robots might bring social risks that will have to be anticipated in order to benefit from these machines. Addiction, isolation, the transference of autonomy toward the machine, and the confusion between machines and people are deviancies that require our attention. One of the risks, in particular for the frail, is to forget that a robot is connected and programmed. A robot capable of adjusting to its human owner could probably be used to lead the person make certain choices rather than others. In particular, it might help manage the owner's deviant behaviors (sexual pathologies, addiction to drugs) and, too, less praiseworthy acts as a consumer. Another risk is to forget that a robot feels nothing, has no emotions or conscience, and is not alive. We might feel empathy for a robot and even talk about its "suffering"; but the elderly who might endanger their lives to rescue their robot should realize that the machine does not suffer, not even if it falls. They must be conscious that it is but a programmed device.

Empathic robots spawn many meaningful questions, ethical, legal and social (BOSTROM 2014), which have only very recently come under discussion. The spectacular progress in digital technology will someday serve to improve human well-being if, instead of thinking of what the technology can be used for, we think about what we want to use it for. Certain ethical values are important: codes of ethics, the responsibility of designers and engineers, the emancipation of users, the evaluation, transparency, explicability, loyalty and fairness of systems, and, finally, studies over the long term of human/machine "coadaptation", so that machines adapt to people, and people to machines.

Human control will always be essential. It is necessary to develop ethical guidelines for social robots, especially in the field of health, and understand the level of human/machine complementarity. We need to demystify, to train artificial intelligence and to place human values at the center of the design of robotic systems.

References

ASADA M. (2019) “Artificial pain: Empathy, morality, and ethics as a developmental process of consciousness” in A. CHELLA, D. GAMEZ, P. LINCOLN, R. MANZOTTI & J. PFAUTZ (editors) *Towards Conscious AI Systems: Papers of the 2019 Towards Conscious AI Systems Symposium*, Stanford, CA, March 25-27, 8p.,available via http://ceur-ws.org/Vol-2287/paper19.pdf.

BOSTROM N. (2014) *Superintelligence: Paths, Dangers, Strategies*, (Oxford University Press).

DAMASIO D. (1994) *L’Erreur de Descartes. La Raison des émotions* (Paris: Odile Jacob).

DARLING K. (2016) “Faut-il accorder une protection juridique aux robots de compagnie?” in A. BENSOUSSAN, Y. CONSTANTINIDES, K. DARLING, J.G. GANASCIA & O. TESQUET (editors) *En compagnie des robots* (Paris: Premier Parallèle).

DEHAENE S. (2014) *Le Code de la conscience* (Paris: Odile Jacob).

DEVILLERS L. (2017) *Des robots et des hommes. Mythes, fantasmes et réalité* (Paris: Plon).

DAVIDSON D. (1980) “Mental events” in *Essays on Actions and Events* (Oxford: Clarendon Press.

DUMOUCHEL P. & DAMIANO L. (2016) *Vivre avec les robots. Essai sur l’empathie artificielle* (Paris: Seuil).

GARCIA M., BECHADE L., DUBUISSON-DUPLESSIS G., PITTARO G. & DEVILLERS L. (2017) “Towards metrics of evaluation of Pepper Robot as a social companion for elderly people”, International Workshop on Spoken Dialogue Systems (IWSDS), 8p.

PICARD R. (1997) *Affective Computing* (Cambridge, MA: MIT Press).

SPINOZA B. (1677) *Éthique*, Livre III, Proposition II translation by Bernard Pautrat (1998) (Paris: Point-Essais), pp. 207-209.

REEVES B. & NASS C. (1996) *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places* ((New York: Cambridge University Press).

TISSERON S. (2015) *Le Jour où mon robot m’aimera. Vers l’empathie artificielle* (Paris: Albin Michel).