The Boundaries of Affective Computing

Mario Verdicchio
University of Bergamo, Italy
Berlin Ethics Lab, Technische Universität Berlin, Germany

Introduction

Affective Computing (AC) is an interdisciplinary field at the intersection of Computer Science, Psychology, and Cognitive Science. The aim is to design, develop, and analyze computational apparatuses capable of detecting, processing, interpreting, and simulating human emotion. Since Affective Computing pursues its goals from a computational perspective, it relies on the core assumption that significant aspects of human emotion are amenable to a numerical description that is compatible with computing machines. Thus, questions arise regarding such compatibility, among which perhaps the most critical is about how to reconcile emotion's central feature of subjectivity, i.e., the first-person quality generated in and experienced by a human mind, with the nature of computational devices, which are objects in the external world.

Is there a way to study first-person experiences with a computational approach? This kind of enquiry has been dealt with for decades after the rise of Artificial Intelligence (AI) in the 1950s (McCarthy et al. 2006). Back then, the focus was not emotion but intelligence, and the question was whether machines could entertain intelligent thoughts the way humans do. The project of creating intelligent machines based on the rules of logic failed; however, a debate is still ongoing in this research field, where neural networks and Machine Learning (ML) systems based on statistics have taken center stage in cutting-edge AI technology (Ghahramani 2015). Such debate is about the boundaries of what can be achieved with computing machines with respect to capabilities that are traditionally attributed exclusively to humans. From this perspective, AC and AI seem to be similar, since both endeavors are about computational approaches to phenomena (emotion and intelligence, respectively) that are so intrinsic to a person's life as they are difficult to grasp and explain from an objective, scientific, and computational point of view.

However, there also seems to be a significant distinction: emotion and intelligence appear to coexist inside a human mind in a way that does not characterize computing machines. A recent achievement in AI sheds light on this difference. Computers are now able to best humans at poker (Brown and Sandholm 2019), but a machine does not seem to win at poker in the same way a consummate human poker player does: of course they both must follow the rules of the game to win it, but if the former only needs to perform probabilistic evaluations to achieve such goal, the latter also needs to keep emotions at bay and hide them from the other players by means of a "poker face". To be able to follow the rules of poker and to take probabilities into account is a facet, although a very narrow one, of what one may call "intelligence" and, in this specific sense, a computer winning at poker could be considered intelligent; the emotion-related side of the game, however, is missing from the machine.

This distinction between emotion-less machines and emotion-laden humans is reflected in two movements in AI research. "Strong AI" claims that one day AI technology will be able to overtake such distinction: computing machines will entertain thoughts in the same way humans do, with full-fledged, first-person, subjective, qualitative experience, also of emotion. "Weak AI", on the other hand, relies on the conviction that there is an ontological barrier between the phenomena occurring in a human brain and the workings of a computing machine, and AI must be content with reproducing only the appearance and the results of human actions, giving up the ambitious goal of creating the subjective experience of human thought inside a machine.

Despite the alarmist warnings of some scholars about sentient machines taking over humanity (Müller 2016), there is no evidence for the claims of strong AI, and machines that entertain thoughts the way humans do belong in science fiction. Indeed, many of the machines featured in Sci-Fi movies are not only intelligent, but they also seem to entertain emotions. Computer HAL 9000 in "2001: A Space Odyssey", for instance, says that it is afraid and begs astronaut Dave Bowman to stop unmounting its components (Kubrick 1968); Samantha, the intelligent operating system from the movie "Her", falls in love with its human user Theodore (Jonze et al. 2014); humanoid robot Ava finally achieves its goal of escaping the lab where it was built and can be seen smiling at the end of the movie "Ex Machina" (Garland et al. 2014).

All these fictional examples of strong AI have this in common: they express emotions by means of words, actions, and appearance, exactly how a human would do. This constitutes the foundations of Affective Computing today: it does not matter whether computing can eventually become affective or not, it does not matter whether machines will one day entertain thoughts the way humans do because whether we are dealing with humans or machines or even fictional machines, the only emotions we can feel are our own, and when it comes to the emotions of others, human or not, real or not, we can only rely on our capability to make inferences based on the words and tones of voice we hear, and the actions and the appearances we see. Since we are given access only to the expressions of the emotions of others, if computation in AC is to detect, process, interpret, and simulate something, that something is not emotion but expressions of emotion. Getting back to the game of poker, whether or not the machines of the future will have to put up a poker face, AC may help them already today read through ours.

The following sections will illustrate the computational technology behind the latest AC endeavors, how such technology is used to analyze expressions of emotion, and the different kinds (sociotechnical, automation- and ML-related) caveats that come with it.

The Computing Behind Affective Computing

AC is a relatively recent research field. What is universally considered its foundational book was published in 2000 (Picard 2000), and its first international conference and flagship journal came to be about a decade after that (Picard 2010). This decade is also the time during which the foundations for the latest AI peak, the one based on neural networks and ML, were laid. It may be fortuitously good timing or a synergistic convergence, but it is clear that computing in AC is significantly connected with the latest developments in ML.

Such relation, as with any other endeavor that adopts a computational approach, relies on one fundamental and essential requirement: whatever the phenomenon with which we are dealing, for computing to be of any use, it must be possible to describe such phenomenon in discrete numerical terms, because this is the only kind of input that computing machines are able to elaborate. From the simplest pocket calculator to the most advanced supercomputer, all computing machines share the same basic principles of binary arithmetic: $0 + 0 = 0$; $0 + 1 = 1$; $1 + 1 = 10$. Computing cannot escape its nature, and even the experimental machines based on quantum computing obey these rules while promising an exponentially more efficient way to perform computation.

Keeping these simple yet often forgotten considerations in mind, the claims of strong AI appear to be even more difficult to realize because they would entail that there exists a way to organize numbers and their functions that yields a sentient mind, endowed with thoughts and emotions. How our mind is created inside our biological brain is still a deep mystery, even for the most knowledgeable

neuroscientists. How such a result can be achieved by means of a computing machine is building another mystery on top of it.

Yet, computing has been and still is one of the biggest scientific and technological successes in the modern history of humanity, thanks to the fact that, despite the limitations imposed by the intrinsically numerical nature of this endeavor, engineers, physicists, and mathematicians have made discoveries and inventions that allow for the encoding of very important physical phenomena and data types, like images, sounds, texts, etc. Encoding these phenomena means obtaining a relevant numerical description so that computing machines can transmit, elaborate, and transform these numbers. Of course, numbers alone are not enough to complete useful tasks: computing machines need to be accompanied by relevant encoding/decoding apparatuses that work as an interface between the physical world and the numerical world, creating numerical descriptions of physical phenomena on the one side, and transforming numerical data into physical phenomena on the other. A computer monitor, for example, transforms the numerical output of a computer into images on the screen, whereas a digital camera captures light from the external environment through its sensor and creates a numerical description of the captured view in the form of a digital image.

Emotions that AC aims at studying are treated along the same lines. At its core, the computation in AC is based on the processing of digital signals, only this time, the signals are about how people express their emotions (Pentland 2007). Capturing affective expressions in the form of digital signals and transforming them into numbers can create a massive amount of data, and this is where ML comes into play. There is a 180-degree paradigm shift between traditional, logic-based, symbolic AI (the one called GOFAI, "good old-fashioned AI", Haugeland 1989) and ML. In the former, a top-down approach is taken, where axioms and general laws are encoded into a computing machine that is given the task of performing deductive reasoning; in the latter, computing machines are used with a bottom-up, data-driven approach where patterns, schemes, and laws are searched among a vast amount of data by means of statistical inductive processes. There are subtle differences between statistics and ML (Dangeti 2017), but they are not significant for the purposes of AC.

The basic idea underlying ML is to "train" a neural network, which is a complex mathematical function that is comprised of a network of "artificial neurons," that is, simple functions connected to each other in the sense that the output of some of them is the input of others (see Figure 1).

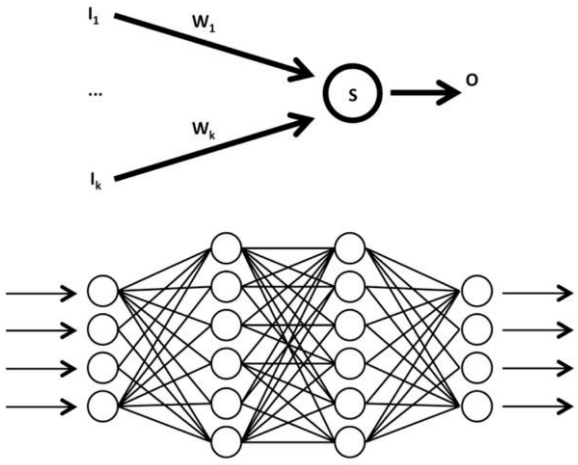

Figure 1: An artificial neuron (above) and a neural network (below).

An artificial neuron (Figure 1, above) is a mathematical function S that takes k different inputs I_i, each of which is weighted, multiplied by a factor W_i, and yields output $O = S(W_1I_1,\ldots,W_kI_k)$. The inputs for

a neural network (Figure 1) come from the data set we want to analyze, and the network outputs are the results of such analysis. "Training" in this context means that, following algorithmic rules, the weights of the neurons in the network are tweaked depending on the network's output. What is the network trained for? What is its output about? One very common aim for which a neural network is trained is classification. Given an array of categories predetermined by human trainers and a set of data that are meant to be classified (i.e., each datum is to be put into one of the categories), training the network means that by a series of trial-and-error, its weights are changed in a way that after a possibly very large number of attempts, the network satisfactorily classifies data, with the number of misclassifications below a certain threshold set by its trainers. At the beginning of this process, the attempts at classification by the network appear random. Every time the network makes an attempt, such attempt is verified against the correct answer previously provided by the trainers, also known as "tagged data," where the "tag" is the correct classification: The difference between the correct answer and the actual output of the network is taken as a basis to modify the weights throughout the network, starting from the weights that are the closest to the output, with a technique that is called "backpropagation."

In what is referred to in the field as "supervised" learning, the network trainers use only a part of the data for the training phase and then test the resulting function on the rest of the data. If everything works well, the network correctly determines the categories of the new data instances, those that were not used for the training. This happens when the neural network has "learned" to generalize from the training data to all available data. At this point, the network is ready to be deployed, and the classification task can be considered reliably automatized. In the context of AC, this method might be used to automatically analyze photographic portraits of people to classify them with respect to the emotion shown by the portraited person. The emotion categories (e.g., happiness, anger, sadness) and the correct classification of the images are given by the human trainers, and once the training phase is completed successfully, the network should be able to classify new data correctly, that is, it should be able to "recognize" the emotion of people in photographs. If we do an image search with a search engine with the input "happy person" and check the results, we notice that all images have some visual features in common (e.g., smiles, mouths open in laughter, thumbs up, fist pumps). These are the features that are "learned" by the mathematical function of the neural network. The network does not recognize a smile the way a human being does. Still, the shape and color of visible teeth and the contours of lips curved in a smile correspond to specific numerical patterns created in the image when translated into numbers (i.e., a digital image) to be fed to the network. Through training, the network shapes its function so that whenever such patterns are found in the input data, the probability for the output to be "happiness" (more precisely, the numerical encoding of that category) increases.

Considering how easy it is for a human being to recognize the smile of a happy person, involving ML in such a task may seem overkill, but the usual advantages provided by automation hold: once trained, a neural network can perform a task in less time than a human being (speed). Moreover, provided that computing machines are endowed with the proper sensors and encoders, a neural network can analyze many more phenomena than a human observer (power). Finally, given the possibility to analyze encoded data at the smallest scale of numerical bits, a neural network can detect patterns even where human perception usually fails (precision).

Let us focus on what aspects of affect a computing machine can "observe." Starting from the widely accepted premise that emotion is built by internal processes of a person within the context of that person's interaction with the external environment (Lewis 2005), emotional states are to be considered multifaceted and comprised of neurobiological, physiological, bodily, action-oriented, cognitive, and phenomenological aspects (D'Mello et al. 2018). If we want to analyze these aspects with a computing machine, we need sensors to capture the relevant phenomena and encoders that create a numerical representation of such phenomena. The output of the encoders will work as input to the computing

machine, which will perform an analysis of the data, possibly to classify them by means of a trained neural network, as discussed before. Since the input to the machine is comprised of numerical data, whatever the emotive phenomenon under observation, one very significant task becomes evident: every effort in AC requires the construction of a computational model that is aimed at bridging the gap between the above-mentioned neurobiology, physiology, embodiment, physical actions, cognition, and phenomenology on the one side, and computable numerical data on the other.

However, this is not the only chasm researchers working with emotions must deal with. Computing machines are not the only entity unable to directly access the affective dimension of a subject: other human beings suffer from the same limitation. This is why it is not very difficult to suspend disbelief in front of machines (like Samantha or Ava) that act emotionally like humans: we observe affective expressions very similar to those of other humans, and we get tricked.

When it comes to research with real humans, what are the data, and how can observers gather them? Non-subjective data are literally "given" (from the Latin verb "dare," to give) about a phenomenon, and they are not the phenomenon itself. The closest data to what a subjective affective experience is would have to be descriptions provided by the subjects themselves. Even in this case, we are dealing with qualitative verbal or textual descriptions of emotions, not the emotions themselves. The subject is not the only source of affective descriptions in an experimental setting, where methodological characteristics such as objectiveness, repeatability, measurability, and others are required. More often, external observers are involved in gathering data, possibly with the use of machines and, in AC, also with computing machines.

Data collection, with or without computing machines, is guided by a simple and fundamental principle: observers need to gather data relevant to the affective states they are trying to analyze. Such relevance originates from all the disciplines that include affective analyses in their theories and practices, like psychology, cognitive science, and psychiatry, to name a few. ML comes at a later stage to enhance effective data analysis, but we cannot rely on computing machines to discover links between emotions and their relevant expressions. We need to remember how supervised learning works: a neural network is trained by means of data already correctly classified by human researchers beforehand. Nothing new can be discovered: Only new data can be classified according to criteria established by humans and "learned" by machines.

There are ML paradigms that are not based on training with data preprocessed by humans, like "unsupervised" learning. In unsupervised learning, the neural network can only find patterns based on the data's values in terms of numerical relative distance in search of clusters of similar instances or outliers. Without any correspondence to the phenomena from which the data are obtained, a computing machine can only indicate quantitative similarities or differences among data. This confirms that, whatever the paradigm adopted, ML in AC still fundamentally relies on knowledge acquired by researchers in disciplines in affective science where humans attribute meaning and names to emotions. Thus, the latest achievements in AI are extraordinary in automatizing and enhancing human capabilities of detection and classification, but they cannot be considered as a replacement for traditional analysis of emotions based on a variety of non-computational devices like films, pictures, music, and dialogue (Coan and Allen 2007).

Suppose one wants to harness the power of the latest computing machines in the context of AC. In that case, they need to keep in mind the inherent numerical nature of these devices: any phenomenon for which there exists a sensor and an encoder that provide a numerical description is a phenomenon that potentially lends itself to computational analysis. Images, videos, audio recordings, texts, and any phenomenon (like a heart's electrical activity) that can be described in the form of a curve in a Cartesian coordinate system, are all candidates. There have already been a great number of experiments in AC, where different emotion expressions like facial expressions, heart rates, gestures, and voice tones were captured by sensors, transformed into numerical data, and analyzed and classified

by computing machines (Calvo and D'Mello 2010, D'Mello and Kory 2015). In the next section, we will focus on one proposal in particular that points to the most critical aspects of applying ML in Affective Science.

An Application of Affective Computing

Mindstrong is a startup founded in the mid-2010s in Palo Alto, California, by three doctors, Paul Dagum, Tom Insel, and Rick Klausner. Insel is the former director of the U.S. National Institute of Mental Health, where he worked for 13 years before moving to Google-Alphabet's life sciences division in 2015. In 2016 Insel had several meetings with Dagum, who came up with the core idea of Mindstrong, and the startup's seed was sown (Metz 2018). The goal of Mindstrong is to help treat medical problems related to mental health, including depression, schizophrenia, and bipolar disorder, by means of a platform used by the startup's clinical team to deliver "evidence-based therapy and psychiatry in structured, goal-oriented messaging sessions" with the aim of "lowering the inpatient readmission rate, E.R. admission rate, mental health costs, and physical costs." (Mindstrong Health 2020a).

The basic idea is to "measure" human-computer interactions on a smartphone and analyze those measurements using ML to monitor the users' mental health. The assumption is that how a person uses their smartphone provides significant indications of their mental state. In particular, Mindstrong is an app that monitors how the person types, taps, and scrolls while using other apps. In his patent, Dagum claims the system can be expanded to include more data (Dagum 2016). Data could be recorded from the smartphone's GPS, accelerometer, and gyroscope to infer characteristics of the user's daily activities, including activity intensity, mobility and methods of travel, social engagement, and travel destinations. The patented invention can also include the option to record data from the device's phone, email, texting, and chat applications to capture incoming and outgoing calls, emails and texts, the length of conversations and messages, and possible differences between the number of emails received and those opened.

Despite the limitations on the types of data a smartphone, the possibilities appear to be ever-expanding: the patent points to the further step of recording data from an app used to scan the barcodes of purchased groceries and of consumed food and beverages to be matched against nutritional facts databases. With further wearable devices, even more phenomena can be turned into data, including heart rate, blood oximetry, body temperature, and even the brain's electrical activity. The system is also open to record data in terms of visited websites on the phone's browser or books read on its e-book reader, with further possibilities for content classification and complexity analysis.

All these data are gathered by the mobile device and sent in encrypted form to the company's servers, where they are analyzed using ML, and the results are sent back to the app on the phone. This analysis is meant to provide insights about a person's lifestyle, including their social engagements, level of activity, dietary habits, and cognitive functions, which can be indicative of good mental health or a problem. However, this is a vision of possible future developments of the app. Let us focus on the data types that the current version of Mindstrong is gathering, centered on human-computer interaction: the user's gestures on the phone's screen.

The app is based on ML, so if it is supposed to analyze such data and classify them as indications of good mental health or otherwise, this means that the neural network used for analysis and classification needs to "know" the connections between users' gestures and their mental state. In other words, there must be a computational model that describes those gestures in numerical terms (e.g., spatial coordinates of taps and swipes on screen, temporal measurements of the speed of typing, etc.) and relates them to emotions, moods, and states of mind. Where does that knowledge come from? In ML terms, how was the neural network of Mindstrong trained?

The starting point was a study based in the San Francisco Bay Area to verify the possibility of measuring a smartphone user's cognitive ability (or lack thereof) by means of checking how they use their device. Based on the assumption that higher-order brain functions are weakened in people with mental illnesses (McTeague et al. 2016), 150 research subjects were assessed with standardized neurocognitive and neuropsychological tests with respect to episodic memory and executive functions (e.g., impulse control, time management, focus). After the test, an app was installed on the subjects' phones that tracked and measured how they interacted with their phone's display regarding swipes, taps, and typing on the on-screen keyboard. The subjects were sent back to their normal lives, and for one year, the app ran in the background, recorded and encoded their behavior on their mobile devices, and sent the relevant data to the Mindstrong servers. After that, the subjects went back for another round of neurocognitive tests. What the researchers had at their disposal for their study were the results of the first round of tests, the results of the second round after one year, and all the data on the subjects' smartphone usage. All the ingredients for a successful attempt at ML-based AC were there:

- Previous knowledge from psychology, cognitive studies, and psychiatry in the form of tests, which allow for the measurement of expressions of emotions and mental health by the subjects;
- New data acquired on smartphone usage by the subjects through Human-Computer Interaction techniques;
- An experiment designed to keep track of the subjects and the correspondence between the state of their mental health (assessed through traditional tests) and their smartphone usage (captured through the app);
- Finally, a neural network ready to be trained and aimed at "learning" what patterns in smartphone usage correspond to what aspects of a user's mental health.

Human researchers and their interdisciplinary hypotheses provide the lion's share of the work, mapping mental health issues onto specific smartphone usage patterns. For example, speed of keystrokes, frequency of use of the "delete" key, speed of scrolling down the contacts list are all considered connected with memory problems, which are in turn considered an indication of brain disorders; another keyboard usage pattern emerges from how quickly the user is able to switch between keyboards to insert punctuation or special characters in their text. This amounts to a switch of tasks that is thought to be connected with the user's ability to focus, another hallmark of mental health.

The researchers at Mindstrong first determined the subjects' baseline by capturing their smartphone usage and combining the found patterns with the results of their neurocognitive and neuropsychological tests and the average measures from the literature. According to Dagum (2018), this process was extremely successful. It cemented his intentions to go on with the startup because not only were the data collected showing strong correlations with the results of the first round of tests, but they also allowed for successful predictions on the results of the second round. At Mindstrong, thus, smartphone usage patterns were considered a viable computational model of some aspects of mental health, able to support the investigation of human cognition and behavior over time. Such investigation, thanks to the exploitation of ML and mobile technology, benefits from all the enhancements mentioned in the previous section: the speed at which Mindstrong's servers can analyze usage patterns, the power of being able to continuously analyze gestures of millions of smartphone users at the same time, and the precision of quantified data that allows for the detection of the slightest deviation from usual behaviors. These enhancements go beyond the natural limitations of traditional psychological treatments: the time constraints of a regular schedule at a therapist or of a trial with new

medications. Dagum and his team claim that brain-disorder treatment has stalled partly because doctors become aware of someone's problems with mental health only when they are well advanced and that the Mindstrong app can provide much earlier detection and continuous monitoring.

Issues with Affective Computing

We have already seen a significant limitation in trying to enhance any detection, classification, or analysis of phenomena with computing machines: their intrinsically numerical nature allows only working on phenomena amenable to computational modeling. This also entails that we need to enrich computing machines with the appropriate tools, such as sensors, encoders, and decoders, to meaningfully translate real-world phenomena into numbers to enable computational analysis and *vice versa* to make the results useful for human users. In case computing machines are endowed with neural networks to harness the power of ML, additional requirements need to be met, because the networks need to be trained with tagged data to "learn" previous knowledge that the computing machines are meant to embody, automatize, speed up, and spread worldwide via telecommunication infrastructures. The Mindstrong initiative seems to check all the boxes: there is a computational model of the phenomena they are trying to tackle that seems to be working, and that computational model is based on a device that is ubiquitous and ensures a constant flow of data that can be used not only to detect problematic behavior, but also to improve the model itself with further training of the neural networks in the system.

Is everything good then? Not really. Can anything go wrong with this kind of AC system? It depends on the scope of our observation. Suppose we focus strictly on the experiments and the processes within the startup. In that case, we might agree that the founder's vision has become true and that his intuitions on a connection between smartphone usage and mental health proved correct. Indeed, there seems to be scientific proof of this success, with randomized tests on the app's efficacy among college students (Melnyk et al. 2020). However, if we widen the scope and include some possible social ramifications of an initiative like Mindstrong, we need to paint a different, more complete, and less perfect picture. There are several interconnected issues, forming a complex network of problematic dependencies. We can try to shed some light by roughly dividing the issues into three main categories. Here listed from more general to more specific: sociotechnical, automation, and ML-specific issues.

Sociotechnical issues

By borrowing a concept from the field of Science and Technology Studies (STS), an ML-based AC system like Mindstrong could be considered a "sociotechnical" system (Bijker 1997), that is, a technological system in tight connection with its human designers and users, who make the technology useful and meaningful. Indeed, a technological artifact does not work in isolation: for it to be successful in the real world, it has to exist in a context where humans work with it to reach their goals. In some of the sociotechnical contexts in which Mindstrong exists, the app turns out to be problematic.

First and foremost, this private initiative has relied on and still relies on significant investments on which investors expect a return. This a very general problem technological endeavors incur: technology comes with a cost that needs to be covered for the initiative to be financially sustainable. In the case of Mindstrong, the startup guarantees "no extra costs, no copays, no coinsurance" by "partnering directly" with the user's insurance company (Mindstrong Health 2020b). This is a US-centric way to deal with the issue of the cost of healthcare; the management at Mindstrong is not responsible for it, and a discussion on the decades-long debate on public vs private healthcare is outside the scope of this work.

However, this points to a general issue with technological innovation: very often, if not always, it comes with a cost, and that cost leads to the exclusion of those segments of the population who are not able to afford it. Exclusionary phenomena happen at all levels in the context of technology and medicine. Simply put, an AC initiative based on smartphone usage excludes all those people who are not able, for one reason or another, to use a smartphone. More in general, despite the rhetoric on a globalized, "flat" world, a truly global phenomenon like the COVID-19 pandemic has shed light, once again, on very stark differences in access to healthcare not only between different countries around the world but even inside a single American state (Bibbins-Domingo 2020). An app like Mindstrong adds to such differences.

When technology is deployed in society, another issue arises with respect to ownership: who owns the technology and, thus, has the right to use it, manage it, and exploit its results. Contrary to before, this problem does not affect the excluded, but those who are included. For the app to be successful, the Mindstrong servers need to collect data on smartphone usage of their users, which the system maps onto mental health parameters. This process entails that a single company creates, hosts, and manages a huge database comprised of sensitive personal data, destined to become even more sensitive if the company's vision of using GPS, browser history, email, heart rate, and electroencephalogram data is implemented. Privacy is the main concern here, but with this ML-based, data-driven approach, there are several facets to this issue.

Mindstrong guarantees that they "protect your personal information and medical records like any other doctor's office or hospital" (Mindstrong Health 2020b), which is what patients normally expect from their therapists and clinics. However, there is a key difference in the number of users from whom a successful app that is as widespread as smartphone technology can collect data, compared to what a single therapist or hospital can do. Moreover, the data collection with a smartphone app is continuous and uninterrupted, and with the envisioned expansion of involved smartphone apps, accessories, and peripherals, it can expand its reach to aspects of the users' lives normally inaccessible to therapists and doctors. This is the main point of Mindstrong: exploiting Information Technology to have a quicker, more powerful, and more precise analysis of the patients.

When privacy is considered, it becomes clear that this is a double-edged sword. One may contend that the contact-tracing apps recently deployed for the COVID-19 pandemic suffer from the same problem, as several lawmakers tried to argue (Birnbaum and Spolar 2020). Still, there is a fundamental difference: contact tracing is performed via Bluetooth technology in a way that preserves personal identity (Ferretti et al. 2020), whereas in the case of Mindstrong, smartphone usage data are explicitly associated with a user exactly for the objective of an automated mental health check for which the app was conceived. The privacy issue does not depend on the fact that a private company manages these data. There are hotly debated controversies also regarding a similar ML-based initiative implemented by the government in China to experiment with metrics and quantification of the value and virtue of its citizens (Wong 2019). Even with a fair, privacy-aware, law-abiding manager, a huge database full of sensitive data about the population is affected by the inherent risk of falling into the wrong hands (Véliz 2020). We are not talking about a single malevolent hacker, identity thief, or profiler. A new authoritarian regime could install itself and exploit data on the users' location, political beliefs, religious background, and, in the case of AC systems like Mindstrong, mental health.

Automation issues

Other issues originate from another significant aspect of this AC project: automation, that is, the goal of automatizing a number of tasks that humans traditionally perform. This is not about the displacement of humans due to the introduction of machines in the job market, which is an issue in other sectors (Acemoglu and Restrepo 2020). Mindstrong does not provide automated therapy but

offers a more efficient detection of specific mental health hallmarks meant to support therapists in their activities. When a human operator is not substituted but assisted by a machine in their tasks, there is nevertheless a risk of other, more subtle issues to the detriment of the quality of the service provided.

What is arguably the most common issue is called "automation bias": the tendency to over-rely on automated systems. Studies on automation bias started in the aviation sector, to investigate the effects of autopilot systems on the performance of pilots (Wiener and Nagel 1988). More recently, research on automation bias in healthcare also developed since early studies showed that clinicians may drop their own correct decisions in favor of erroneous advice from computer-based clinical decision-making systems (Friedman et al. 1999). In particular, there are two kinds of pitfalls: errors of commission, in which the human user follows incorrect advice coming from the computer, and errors of omission, where the human user fails to act because the computer did not prompt them to do so (Goddard et al. 2012).

In the context of AC-enhanced mental health checks and therapy, an error of commission may mean that unnecessary or even counterproductive drugs are prescribed because of a warning from the system that does not correspond to an actual condition in the patient; in turn, an error of omission may happen when a therapist neglects some possible telltale signs of psychological discomfort and fails to address the issue with the patient because their computational profile does not trigger an alert. These risks are significantly limited in the current version of the Mindstrong app because the automatic detection of deviations from the baseline is meant to help the therapist within the context of a regular cycle of sessions with their patient. This means that the human-to-human component of the interaction is still playing a major role, and the therapist is still invested with full responsibility in their professional activity.

Interestingly, studies have shown that automation bias is reduced in those working environments where human users feel more responsible and accountable for their actions (Skitka et al. 2000). However, if smartphone-usage-based initiatives become very successful, one might imagine that the whole endeavor gets scaled up in some facets, like the number of downloaded apps, users and ML-enhanced computing machines, and less so in others, like the number of employed therapists and clinicians. In a scenario where each therapist has a greater quantity of patients to attend to, it is reasonable to fear that therapists might give in to the temptation to trust the computational machine more and more to increase productivity, and automation bias might increase as a result.

Strictly connected to over-reliance on automated systems is another issue: "deskilling." This term refers, in general, to the reduction or even complete loss of the capacity to perform a task due to the habit of relying on technological tools that automatize that task. In the healthcare context, doctors depend more and more on technology for obtaining patient information and performing diagnoses and treatments, and there is evidence showing negative effects on doctor-patient communications, physical examination skills, and development of clinical knowledge (Lu 2016).

At first glance, there seems to be no need to worry about deskilling in therapists due to the current version of Mindstrong because the app adds a new kind of phenomenon that can be considered indicative of mental health problems, smartphone usage in terms of swipes, taps, and keystrokes, that therapists could not detect before. In other terms, one might consider this the opposite of deskilling since technology augments the therapist's view of their patients. However, a connection with automation bias exists: a therapist might tend to trust the computational system's automated detections and notifications rather than their active observations in the context of traditional, human-to-human therapy sessions. Even if technology does not directly interfere with the traditional tasks a therapist is called to, automation bias might still shift the therapist's focus towards the new automated tasks, causing deskilling in the old ones and, ultimately, a lowering in the overall quality of the therapy.

Machine Learning issues

The most specific issues of AC systems are imported from ML: they are not intrinsic to the attempt to build computational models of emotion expression but to the choice of building and processing those models by means of neural networks. For a neural network to "learn" the task for which it has been built, training with a great amount of tagged data is necessary. We have already seen that such a need entails privacy issues that may dangerously lean toward population control and surveillance. Sociotechnical considerations aside, within the context of ML, there are more technical issues connected with how the data used for the training can affect the learning process and the resulting characteristics of the neural network.

Two major Information Technology companies lost face because of training data issues. Microsoft deployed a chatbot on Twitter, a piece of software with the persona of a young woman called Tay that was supposed to exchange messages with other users, learn from those messages, and post relevant tweets. Some malicious users fed the chatbot with racist messages, and the software eventually started posting tweets in a similar tone, and it had to be taken down (Hunt 2016). Google's automatic image labeling service deployed in the company's online photo albums created another racially charged case. User Jacky Alciné, a black man, noticed that in his photo album, the pictures he had taken with his girlfriend were being sorted into a folder tagged "gorilla." No other images but those of him and his girlfriend appeared in that folder, according to Alciné. (Griffin 2015).

Both cases point to problems from training a neural network with inadequate data. In the Microsoft case, the connection between training data and network output is evident and, in a rather twisted way, speaks to the good operation of their ML system: it was fed racist messages, and it learned to make racist tweets. The Google case is more complex because the training data was selected and used within the company's walls. Still, we can refer to experimental evidence that limited exposure to diverse races and age groups by training only on data belonging to a specific category creates biased neural networks (Nagpal et al. 2019).

These considerations bring our focus back to the experiments that laid the ground for developing the Mindstong app: standardized neurocognitive and neuropsychological tests were used with 150 research subjects from the Bay Area to assess their mental health parameters and map them to their smartphone usage. Some questions arise: was this group of subjects well-chosen in terms of representation of the general population? Is there a risk that the neural network at the foundations of the app was instead trained with biased data so that it has developed a biased way to operate? A neural network can always evolve with new training sessions, so the more customers Mindstrong acquires, the more chances the company has to expose its ML system to new, diverse smartphone usage patterns to reduce a possible initial Bay-Area-centered bias. However, doubt remains: Who or what is in charge of determining whether a newly observed smartphone usage pattern, which presents significant differences from what was "learned" by the network as corresponding to a healthy emotional state, is to be interpreted as a new healthy pattern or an unhealthy one?

This sheds light on the fact that not only is population sampling critical, but there is also a fundamental need for constant supervision and interpretation by human trainers with both in-depth knowledge of the field where they intend to apply ML and a breadth of scope with respect to varieties in habits among different populations whose data are to be used. Human intervention in the context of ML is not easy, and the way the companies managed the above-mentioned racism incidents points to this issue: neural networks work following a "black box" paradigm, that is, only their input and output are visible, and meaningful to the trainers and the users, whereas the internal parameters, the weights associated with the connections between the artificial neurons that are modified during the training phase, do not provide any understandable indication on which parts of the network attend to which features of the performed task. When a neural network that has been trained to tweet starts tweeting

in a racist way, even its creators and trainers are not able to determine in which part of the network racist characteristics are encoded. When a neural network that is meant to tag images starts mistagging only images of a specific group of people, even the computer scientists that supervised its learning phase cannot immediately shut down the specific components of the network making the mistakes. This is why Microsoft had no other choice than retiring the chatbot altogether, and Google had to disable the "gorilla" tag for all image classifications (including the correct ones) while attending to their neural networks, presumably with some extra training with images of people of colour.
This is a general problem with "black boxes": all is well when they work well, but when faulty behavior happens, given the complexity and non-explicit knowledge inside, correction is far from simple or immediate. The above-mentioned cases were easy to detect, thanks to the clear evidence. Still, machine error detection might be much less straightforward in case of more subtle deviations from expected behavior. In the context of an app like Mindstrong, where therapist-patient sessions still play a major role, classification and detection errors of the ML system may not have a great impact since the therapist can have a direct reassessment with the patient. However, in future scenarios where the automated computational part gets increasingly important in the overall healthcare service, the possibility of misdiagnoses and wrong treatments may increase should there be a negative synergy between automation bias and the limitations of the black box paradigm.

Conclusions

Excluding subjective experience of emotion from the context of Affective Computing moves the field away from the unproven claim of strong AI that a properly complex computational system can create a mind with consciousness, thoughts, and emotions. However, even with the decision to give up on the imaginings deriving from a computational theory of the mind, AC is left with a lot to deal with about emotion in terms of behavior: how a person acts when experiencing a certain kind of emotion. By adopting a behavioristic perspective, AC can be interpreted as aimed at creating devices that detect, process, interpret, and simulate human behavior in terms of emotion. AC needs to rely on sophisticated devices with the appropriate sensory and reasoning apparatuses to build a computational model that creates and holds correspondence between the numerical data encoded by the captured human behaviors and the psychological and neurophysiological theories that map those behaviors onto emotions. The sensory part can be considered a technological evolution of tools that have been long used in the tradition of emotion elicitation and assessment in psychology, whereas the reasoning apparatuses are connected with a radical change that AI has undergone in the last decade, when ML has taken over traditional logic and rule-based reasoning systems. With an approach based on massive quantities of data and statistical analysis, ML has already shown that computational machines fare much better than the best humans in a number of contexts where classification and detection tasks are automated, augmented, and enhanced. Given a platform that allows for the collection of data from human subjects and a theory that connects data to the affective states of those subjects, ML seems to offer just the right approach and technology for deploying AC on a massive scale.
This work has focused on the issues in ML that AC imports by adopting this technological approach for its endeavors: exclusion, privacy, social control, automation bias, deskilling, and biased output are some of the problems that impact data-driven systems based on neural networks and that will become a permanent fixture of AC if the two disciplines become inextricably linked. We have analyzed the case of an AC app that depends at its very core on massive data analysis, since not only is it based on ML, but also smartphone usage data. Given the nature of this endeavor, the above-mentioned issues

might be inevitable and a necessary evil that the proponents have the moral duty to contain with as many countermeasures as possible to ensure that their project accomplishes all the beneficial goals for which it seems to have the potential.

Those countermeasures have already been discussed for some years in many circles, including the AI community, where all the downsides of a purely data-driven approach have become clear. The more technological the issues are, the more focused the efforts. For instance, now there exists a subfield of AI called Explainable AI (XAI), dedicated to countering the problem of the black-box paradigm by means of an enhancement of ML with symbolic-AI based systems that are meant to make the inner workings of neural networks more explicit and understandable to their designers, so that the designers have more control on how to train them and correct them, if necessary (Adadi and Berrada 2018).

The mission of XAI is not at all trivial from a technological perspective; moreover, it seems to become even more complex from a socio-technological perspective. ML is deployed in so many contexts on a global scale that we have not fully comprehended its socio-economic impact yet, but there are already many social groups that are negatively affected in terms of inequality, and there is potential for even greater harm on a more general societal level (O'Neil 2016).

Is this what AC is about? Will computational models of emotional expression become a use case of the latest developments in AI and ML and cater to all the relevant debates involving social injustice and human rights? These are questions that all researchers, scholars, and therapists in the field need to ask every time they engage in AC that is enhanced with ML techniques. Depending on the context, this technology may just bring in the benefits of computation: speed, power, and precision. Or, it may usher in problems that not only were not generated in the field of Affective Science but may also strengthen the issues Affective Science was called to solve. Despite the speed at which technology seems to be evolving, the reins of Affective Science are still in the hands of human experts. These experts now face both the opportunity of harnessing the power of computing and the risk of being harnessed by computing machines.

Acknowledgment

I am grateful to Martina Vortel of the Berlin Ethics Lab at the Technische Universität Berlin, Germany, for pointing me in the direction of the Mindstrong initiative.

References

Acemoglu, Daron, and Pascual Restrepo. "Robots and jobs: Evidence from U.S. labor markets." *Journal of Political Economy* 128, no. 6 (2020): 2188-2244.

Adadi, Amina, and Mohammed Berrada. "Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI)." *IEEE Access* 6 (2018): 52138-52160.

Bibbins-Domingo, Kirsten. "This time must be different: disparities during the COVID-19 pandemic." *Annals of Internal Medicine* 173, no. 3 (2020): 233-234.

Bijker, Wiebe E. *Of bicycles, bakelites, and bulbs: Toward a theory of sociotechnical change.* MIT Press, 1997.

Birnbaum, Michael, and Christine Spolar. "Coronavirus tracking apps meet resistance in privacy-conscious Europe." *The Washington Post* (2020). https://www.washingtonpost.com/world/europe/coronavirus-tracking-app-europe-data-privacy/2020/04/18/89def99e-7e53-11ea-84c2-0792d8591911_story.html (Last visited January 2025).

Brown, Noam, and Tuomas Sandholm. "Superhuman A.I. for multiplayer poker." *Science* 365, no. 6456 (2019): 885-890.

Calvo, Rafael A., and Sidney D'Mello. "Affect detection: An interdisciplinary review of models, methods, and their applications." *IEEE Transactions on affective computing* 1, no. 1 (2010): 18-37.

Coan, James A., and John JB Allen, (editors). *Handbook of emotion elicitation and assessment*. Oxford University Press, 2007.

Dagum, Paul. "Method and system for assessment of cognitive function based on mobile device usage." U.S. Patent 9,420,970, issued August 23, 2016.

Dagum, Paul. "Digital biomarkers of cognitive function." *NPJ Digital Medicine* 1, no. 1 (2018): 10.

Dangeti, Pratap. *Statistics for machine learning*. Packt Publishing Ltd, 2017.

D'Mello, Sidney, and Jacqueline Kory. "A review and meta-analysis of multimodal affect detection systems." *ACM Computing Surveys (CSUR)* 47, no. 3 (2015): 1-36.

D'Mello, Sidney, Arvid Kappas, and Jonathan Gratch. "The affective computing approach to affect measurement." *emotion Review* 10, no. 2 (2018): 174-183.

Ferretti, Luca, Chris Wymant, Michelle Kendall, Lele Zhao, Anel Nurtay, Lucie Abeler-Dörner, Michael Parker, David Bonsall, and Christophe Fraser. "Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing." *Science* 368, no. 6491 (2020).

Friedman, Charles P., Arthur S. Elstein, Fredric M. Wolf, Gwendolyn C. Murphy, Timothy M. Franz, Paul S. Heckerling, Paul L. Fine, Thomas M. Miller, and Vijoy Abraham. "Enhancement of clinicians' diagnostic reasoning by computer-based consultation: a multisite study of 2 systems." *The Journal of the American Medical Association* 282, no. 19 (1999): 1851-1856.

Garland, Alex, Andrew Macdonald, and Allon Reich. "Ex Machina." [Motion Picture] Universal Pictures, 2014.

Ghahramani, Zoubin. "Probabilistic machine learning and artificial intelligence." *Nature* 521, no. 7553 (2015): 452-459.

Goddard, Kate, Abdul Roudsari, and Jeremy C. Wyatt. "Automation bias: a systematic review of frequency, effect mediators, and mitigators." *Journal of the American Medical Informatics Association* 19, no. 1 (2012): 121-127.

Griffin, Andrew. "Google Photos Tags Black People as Gorillas." *The Independent*, 2015. https://www.independent.co.uk/life-style/gadgets-and-tech/news/googlephotos-tags-blackpeople-as-gorillas-puts-pictures-in-special-folder-10357668.html (Last visited January 2025).

Haugeland, John. *Artificial intelligence: The very idea*. MIT Press, 1989.

Hunt, Elle. "Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter." *The Guardian* 24, no. 3 (2016): 2016. https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter (Last visited January 2025).

Jonze, Spike, Megan Ellison, Vincent Landay, Joaquin Phoenix, Amy Adams, Scarlett Johansson, Chris Pratt, Rooney Mara, Olivia Wilde, and Owen Pallett. "Her." [Motion Picture] Sony Pictures, 2014.

Kubrick, Stanley. "2001: A space odyssey." [Motion Picture] Metro-Goldwyn-Mayer, 1968.

Lewis, Marc D. "Bridging emotion theory and neurobiology through dynamic systems modeling." *Behavioral and Brain Sciences* 28 (2005): 169-245.

Lu, Jingyan. "Will Medical Technology Deskill Doctors?" *International Education Studies* 9, no. 7 (2016): 130-134.

McCarthy, John, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon. "A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955." *AI magazine* 27, no. 4 (2006): 12-12.

McTeague, Lisa M., Madeleine S. Goodkind, and Amit Etkin. "Transdiagnostic impairment of cognitive control in mental illness." *Journal of Psychiatric Research* 83 (2016): 37-46.

Melnyk, Bernadette Mazurek, Jacqueline Hoying, and Alai Tan. "Effects of the MINDSTRONG© CBT-based program on depression, anxiety and healthy lifestyle behaviors in graduate health sciences students." *Journal of American College Health* (2020): 1-9.

Metz, Rachel. "The smartphone app that can tell you're depressed before you know it yourself." *Technology Review*, October 15, 2018. https://www.technologyreview.com/2018/10/15/66443/ (Last visited: January 2025).

Mindstrong Health. "Fixing mental healthcare to empower everyone." https://mindstrong.com/about-us/ (Last visited: October 2020; the website is no longer active as of January 2025), 2020a.

Mindstrong Health. "Your resources as a Mindstrong member." https://mindstrong.com/our-services/ (Last visited: October 2020; the website is no longer active as of January 2025), 2020b.

Müller, Vincent C., ed. *Risks of artificial intelligence*. CRC Press, 2016.

Nagpal, Shruti, Maneet Singh, Richa Singh and Mayank Vatsa. "Deep learning for face recognition: Pride or prejudiced?" (2019). *arXiv preprint arXiv:1904.01219*.

O'Neil, Cathy. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2016.

Pentland, Alex. "Social signal processing [exploratory DSP]." *IEEE Signal Processing Magazine*

24, no. 4 (2007): 108-111.

Picard, Rosalind W. *Affective computing*. MIT Press, 2000.

Picard, Rosalind W. "Affective computing: from laughter to IEEE." *IEEE Transactions on Affective Computing* 1, no. 1 (2010): 11-17.

Skitka, Linda J., Kathleen Mosier, and Mark D. Burdick. "Accountability and automation bias." *International Journal of Human-Computer Studies* 52, no. 4 (2000): 701-717.

Véliz, Carissa. *Privacy is Power - Why and How You Should Take Back Control of Your Data*. Bentam Press, 2020.

Wiener, Earl L., and David C. Nagel, (editors). *Human factors in aviation*. Gulf Professional Publishing, 1988.

Wong, Karen Li Xan, and Amy Shields Dobson. "We're just data: Exploring China's social credit system in relation to digital platform ratings cultures in Westernised democracies." *Global Media and China* 4, no. 2 (2019): 220-232.