Ethical Considerations on Affective Computing: An Overview

This article summarizes ethical considerations on affective computing. Dangers include oversimplification of affective states, lack of grounding in reality, potential addiction to affective systems and implicit dependence on those controlling the systems, and potential manipulation among others.

By Laurence Devillers, *Member IEEE*, and Roddy Cowie, *Member IEEE*

ABSTRACT | Affective computing develops systems, which recognize or influence aspects of human life related to emotion, including feelings and attitudes. Significant potential for both good and harm makes it ethically sensitive, and trying to strike sound balances is challenging. Common images of the issues invite oversimplification and offer a limited understanding of the moral consequences and ethical tensions. Considering the state-of-the-art shows how pervasive and complex they are. In many areas, the discipline can potentially bring ethically significant benefits and hence has a duty to try. They include making interactions with machines more effective and less stressful, diagnostic and therapeutic roles in emotion-related disorders, intelligent tutoring, and reducing isolation. However, the limits of recognition technology mean that actions are likely to be based on impoverished representations of people's affective state, particularly with certain groups; systems are liable to arouse feelings that are positive, but not well grounded in reality, affectively engaging systems can become addictive and manipulative, and they confer dangerous power on those who control the technology. We offer an overview of those and other particular ethical issues, positive and negative, which arise from the current state of affective computing. It aims to reflect the complexities inherent in both the technology and current ethical discussions. Establishing appropriate responses is a challenge for society as a whole, not only the affective computing community.

KEYWORDS | Affective computing; affective computing applications; ethical/societal implications; influencing human emotional state.

I. INTRODUCTION

The term "affective" describes states where rational processes are influenced by nonrational elements of mental life, such as feelings, attitudes, inclinations, and bonds. There are multiple kinds of affective states. Emotions are the central example: others include passions, moods, stances, states of arousal, and so on. They pervade human life: states with no affective components are very much the exception [1]. Affective processes interact intimately with most psychological mechanisms: as Damasio [2] has argued, feeling grounds rationality rather than being opposed to it.

Affective computing responds to that issue. It is a branch of artificial intelligence (AI), which enables machines to take account of the affect that pervades human life—including, but not only, in human–machine interactions. The term was introduced in a book by Picard [3] in 1997, and the HUMAINE network of excellence (2004–2007) built a research community in the area [4], [5]. There is now extensive work in the field. Several technologies are involved: emotion recognition, sentiment analysis, emotion interpretation, and generation of synthetic emotions. Between them, they enable systems with some form of AI to recognize, interpret, and simulate emotions in their interactions with humans.

The aim is rarely to have dramatic emotional effects. Emotional coloring is part of communication between

Manuscript received 2 November 2022; revised 14 March 2023, 28 April 2023, and 27 June 2023; accepted 5 September 2023. Date of publication 2 October 2023; date of current version 11 October 2023. The work of Laurence Devillers was supported in part by the ANR French Chair HUMAAINE "HUman-MAchine Affective INteraction & Ethics." *(Corresponding author: Laurence Devillers.)*

Laurence Devillers is with LISN-CNRS, Sorbonne University, 91405 Orsay, France (e-mail: devil@limsi.fr).

Roddy Cowie is with the School of Psychology, Queen University Belfast, BT9 5BN Belfast, U.K. (e-mail: roddycowie@googlemail.com).

Digital Object Identifier 10.1109/JPROC.2023.3315217

humans, and enabling artificial systems to imitate it narrows the gap between natural and artificial communication [6]. According to research summed up in the media equation [7], people approach interactions with artificial agents in the same way as interactions with other people, and hence, narrowing the gap, in respect of emotion as in other ways, eases the interaction.

It is well recognized that affective computing is one of the areas of AI, which poses ethical challenges [8]. This is not because it is “a bad thing.” Making interactions with machines more effective and less stressful is clearly a positive goal, and it has been central to the area since it began to take shape [4]. The difficulty is that working toward that goal, and others like it, raises the prospect of various negative consequences; and it is not obvious how to balance potential goods and harms.

Responding to the ethical tensions is not a minor issue. Developments in the area have the potential to bring very significant benefits, but it is also possible to see that they could have very negative outcomes. The issue is also far from simple. The central aim of this article is to convey the kind of challenge that the contrasting possibilities present.

A. Contributing Domains

A key part of the challenge is that the field brings together several conceptual domains. All of them present a double difficulty. On one hand, contemporary thinking indicates that they are highly complicated. On the other hand, much simpler pictures are implicit in familiar ways of thinking and talking about each. The challenge includes both recognizing that the simpler conceptions are not to be relied on and also bringing to bear the different kinds of more sophisticated understanding. Discussions of the relevant sources run through this article, but it is useful to give a brief outline of the individual domains at this stage.

The first domain involves ideas about affective aspects of human life. Ideas that come easily to mind portray them as relatively contained parts of life—for instance, “emotional episodes” where rationality is temporarily suspended or a distinctive kind of feeling that comes and goes. Contemporary understanding identifies those as aspects of a complex whole, involving systems with multiple components, whose actions run through large parts of human functioning [9]. The way technology interacts with those systems is more complicated than familiar simplifications suggest and is also more important.

The second domain involves the technology that interacts with the human systems. Discussions involving ethics are often akin to science fiction—they consider the benefits and risks that might flow from systems with abilities that we can imagine, but that are far removed from what we can implement. Thinking about the systems that we can implement is a very different matter. It has to take account of problems that turn out to be obstinately difficult and partial solutions with limitations that throw up new problems that would not have been foreseen. This adds another layer of complexity to be faced.

The third domain of ethical thought continues the pattern. Time-honored traditions suggest that moral conduct involves acting in accordance with a modest number of general, explicit principles. Assuming that we can rely on principles like that is simplified at several levels.

1) There have always been competing ideas about the appropriate principles.
2) Recent developments argue that the principles relevant to information technology are unlike any of the standard principles [10].
3) The emphasis on principles has been challenged. Particularists argue that moral understanding is grounded in intuitive responses to particular situations, where multiple factors interact in a specific way [11]. On that understanding, general principles are simplified derivatives: they attempt to capture as much as possible of the complex underlying pattern of responses in a compact way. Hence, it is to be expected that there will be ways of proceeding that are in accordance with the principles we are currently treating as guides, but that particular moral assessments identify as ethically unsatisfactory.

A fourth domain is linked to ethics but is worth separating. It involves the way we envisage ensuring that developments and applications are ethically appropriate. The most obvious question is who is required to take specified kinds of action. The simple assumption is that the responsibility rests on those who build and deliver the systems. Recently, influential discussions have argued that in a society where complex technologies are pervasive, the responsibility rests on the whole society and should be shared in co-responsibility between technology providers, deployers, and users [10].

On that understanding, it matters to form an overview of the ethical issues in affective computing that recognizes the relevant kinds of complexity. This article tries to move toward that kind of holistic overview. It is clearly not complete, but it offers a useful general understanding and may serve as a basis for a more extended treatment. It also responds to an obvious concern. It is standard practice to value simplicity and order in academic analyses. The danger is that if a domain is not intrinsically orderly, the standard emphasis on those virtues can favor accounts that fail to reflect important realities. It is reasonable to suspect that the ethics of affective computing may be that kind of domain. In that context, it is important to have discussions whose priority is to ensure that the complexities are acknowledged, whether or not that yields a satisfying order.

B. Connecting Domains

There are cases where emphasizing complexity makes it impossible to see a way forward. This is not the intention here. Part of the case is that the domains link in a constructive way. Roughly, affective technology has developed to a level where it throws up a great many issues that it is natural to see as morally significant. They often involve the way the technology interacts with features of human affective systems that are not brought out by familiar conceptions. From a particularist point of view, the intuition that issues like that are morally significant is to be taken seriously, whether or not it aligns with preconceived sets of moral principles. At least sometimes, the specific intuitions suggest more general principles whose relevance to the area might not be immediately obvious. Considering specific issues and principles leads to ideas about the kinds of action that might be taken to achieve ethically positive outcomes and avoid negative ones, and about whose responsibility it is to take them.

Broadly speaking, this is the approach which is followed in this article. It first considers background domains where current thinking indicates that matters are much more complex than familiar conceptions suggest, that is, ethics and human affective systems. It then considers research on affective computing with a view to identifying the ethical issues that it raises. This includes considering work that has already tried to identify ethical issues that are important for the area; and looking across major areas of activity, technical, and applications, to identify ethical issues that arise from them. On that basis, general themes, which seem particularly acute, are drawn out. Finally, this article considers relevant courses of action and whose responsibility it is to take them.

It is worth noting in advance that the outcome involves a very different emphasis from ethical discussions, including our own, which were written when the field was emerging. This illustrates why it matters to bring practical developments to the center of ethical discussions: they bring out issues whose importance may not be easy to recognize a priori.

II. CONCEPTUAL PRELIMINARIES

It matters to recognize that standard ways of understanding central concepts—most obviously “emotion” and “ethics”—are liable to have limiting effects. They make it easy to assume that any sensible person knows what the issues are and what is to be done. This is understandable. The concepts cover enormously complex aspects of human functioning. It is natural and adaptive to conceptualize phenomena like that in terms that deal as simply as possible with the aspects most relevant to the way we interact with them. However, when we are interacting with them in a new way, aspects that were previously not relevant can become important, and it matters to ask whether the familiar conceptions undervalue features that are significant in the new context.

A. Human Affective Systems

With regard to affect, the key issue is the intricacy and power of the systems that an affective agent aims to engage. The central term, “emotion,” reflects a historical choice: to highlight a distinctive kind of feeling rather than trying to specify what lies behind it [12]. However, psychology indicates that the feeling is an integral part of a complex, far-reaching pattern of activity, involving a distinctively evaluative kind of perception; selective evocation of memories; preparation to act; autonomic and biochemical adjustment; encoding into memory; and expressive behaviors [9]. Complementing that, neuroscience indicates that the processes involve functionally integrated systems spanning diverse cortical and subcortical regions [13]. Strikingly, the systems enable the perception of similar states in others by mirroring them—configuring themselves in ways that would give rise to the behaviors that others are displaying [14]. As well as being complex, the systems are notably sensitive: they derive strong responses from modest inputs. Tomkins called them amplifiers [15]. Affective computing engages with systems like that. Given their intricacy and power, we should expect the possible outcomes of engagement to be complex, involving various potential benefits and harms. This is the context for the ethical judgments being considered.

B. Ethics

The earlier comments on particularism indicate why ethical concepts need to be treated cautiously. For most of history, theorists assumed that their aim was to articulate principles belonging to a moral order independent of humanity, divine, or transcendental. People still talk as if there were objective moral principles, but it is hard to defend rationally. A well-known argument is that contemporary understanding of reality leaves no place for them. More recent arguments show that principles, which seem to be intuitively compelling, lead to unacceptable conclusions [16]. The most famous is the “repugnant conclusion” that highly plausible principles direct us to aim for a world where the greatest possible number of people live lives barely worth living. More formal treatments show that they imply contradictions.

A broad alternative, of which particularism is a strong form, argues that morality derives from intuitive judgments based on particular situations, about what people should and should not do. Ethical theory aims to articulate principles that capture what is of value in judgments like that as explicitly and generally as possible. In brief, they are attempts to simplify an underlying reality that is complex and centrally concerned with balances. Hence, we should not be surprised that relying on them uncritically leads to some unacceptable conclusions.

This understanding has practical implications. This means that influential formulations need to be treated cautiously. Arguments often present certain principles as fundamental: we are invited to accept that if a course of action can be shown to agree with a particular principle, we are therefore ethically bound to support it; or if it violates one, we are bound to oppose it. The obvious positive example involves promoting happiness: more is said about that shortly. Obvious negative examples involve approaches centered on human rights, which condemn any violation of dignity, respect, freedom to choose, independence, or equality. Given that we know strongly held intuitions lead to contradictions, it is hard to defend the assumption that an argument from any one principle like that, or even a few, should be treated as decisive.

If we accept that balances are critical, it is natural to ask what defines a satisfying balance. A famous answer is the utilitarian principle of "the greatest happiness for the greatest number." This is credible when happiness is created indirectly, by providing other generally recognized goods (health, knowledge, and friendships), but it is very suspect in the context of a technology capable of inducing happiness without any of those. There are also notorious difficulties over tradeoffs between the numbers of happy people and their average happiness [16]. It is hard not to conclude that the desire for precision needs to be treated very cautiously. Some kind of balance is critical, but the criterion of balance is no clearer than the intuitions to be balanced.

The utilitarian formula points to another issue that is highly significant for affective computing. Its emphasis on happiness may be a simplification, but it reflects a strong tradition of linking ethics with assessments that are felt rather than logical. "Moral emotions"—such as guilt, shame, sympathy, and respect—are at the very least strong guides to moral judgment [17], and the feelings that an action creates in other people are very obviously significant for the way we evaluate it morally. Interventions in a domain like that can hardly avoid being morally charged.

An approach due to Floridi [10], information ethics, highlights two other levels of complexity. Its general argument is that information technology raises ethical questions that are not well handled by traditional formulations. They focus attention on individuals taking actions whose immediate consequences are the subject of evaluation. Floridi [10] argued that the social deployment of AI calls for a different perspective. Key arguments apply to affective computing as a subdiscipline of AI.

The first involves units of analysis. An AI system impinges on users when it is deployed, but reaching that point involves many actions that are ethically significant in their own right [18]. Global assessments can easily overlook key parts of the process or criteria relevant to evaluating them. For example, the collection of training data, evaluation during development, maintenance after deployment, and retirement all raise distinctive ethical issues. Conversely, actions that are ethically admirable in most respects can have consequences that are highly problematic, for instance, by deepening the "digital divide" between those who benefit from the new technologies and those who are left behind [19].

This complexity means that the development and deployment of an affective computing system involves many people. This raises a second level of concern, highlighted by Floridi [20]: "in distributed environments, it is increasingly common that a network of agents—some human, some artificial (e.g., a program), or some hybrid (e.g., a group of people working as a team thanks to a software platform)—may cause morally good or evil (henceforth loaded) actions through local interactions that are not, in themselves, morally loaded but neutral." Hence, "the risk is that 'everybody's problem' becomes nobody's responsibility" [21]. Information ethics theory aims to address that issue.

At a broad level, information ethics is a major influence on this article. It recognizes the need to consider components as well as whole systems and to address distributed responsibility. At the level of detailed prescriptions, it is less clear how well ideas oriented to AI in general transfer to systems concerned with affect. Clarification depends on drawing out the distinctive issues that affective computing raises. This is the central task in this article.

C. Responding to Complexity

It may seem unbalanced to dwell on so many conceptual difficulties. If the issues were minor, it would be: but they are not. It becomes obvious that there are large questions when we consider the prospect of machines taking on roles that would previously have been filled by an emotionally competent human being—companions for elderly people, teachers for children, playmates, friends for the lonely, confidantes, counselors, sales people, managers, police, and so on.

Addressing questions like that, it is worth making sure that we have not oversimplified the issues. This is reflected in the way this article moves forward. It aims to evoke a sense of the diverse and tangled issues that arise when we connect deeply sensitive aspects of humanity with highly complex technology. There may in the end be clear guidelines that allow us to navigate the area, but it is hard to see how we can evaluate them without an overview of the challenges to be dealt with.

III. STATE-OF-THE-ART

A. Ethical Background

It is useful to think of relevant ethical discussion in three phases. First, discussions of general principles date back to the ancient world. Second, the widening impact of computing prompted developments specific to it in the early 21st century, including a code for robotics developed by the U.K. Engineering and Physical Science Research Council (EPSRC) [22]; a major revision of the ACM's code of ethics [23]; and initial work on ethical issues in the emerging field of affective computing [24]. The third phase, currently in progress, addresses issues raised by the maturing field. This section considers the first two.

The comments on classical discussions in Section II suggest a broad orientation. Time-honored sources do not integrate into a single coherent system, but they do articulate strong feelings about what is desirable or undesirable. On that basis, it makes sense to respect as wide as possible a range of enduring intuitions but to beware of dogmatic insistence on conclusions drawn from any one principle. A useful pointer to the range of intuitions is a distinction between three basic approaches to the underlying principles of ethics—deontological, consequentialist, and virtue ethics. They stress fulfilling duties, acting to ensure the best outcome, and behaving in keeping with a good character.

There is not a clear relationship between that level and professional codes of ethics (most obviously bioethics, which governs medical practice) [25]. However, roughly, the codes tend to involve duties agreed between practitioners and society in general, reflecting shared judgments about what constitutes the best outcome. In terms of that picture, affective computing is at the stage of working to agree suitable codes. It matters to understand the more general principles because they underpin the negotiation.

The initial work that dealt specifically with affective computing was summarized in two handbooks [26], [27]. Döring et al. [28] proposed an overall framework known as principalism, which was developed in the context of bioethics by Beauchamp and Childers. It prescribes four duties toward people receiving professional services:

1) beneficence—to do them good;
2) nonmaleficence—to avoid doing them harm;
3) autonomy—to respect their right to make decisions about themselves even if they do not seem wise;
4) equity—to ensure that the benefits of expertise are fairly distributed.

Principles like that are an important touchstone. They provide the basis for asserting that particular choices are a matter of ethics, not personal taste. However, their effect on practice depends on clarifying how they connect to issues that arise in the area—enabling people to see where natural developments are liable to lead to benefits or inequities, what can be done about potential problems, and who should do it.

Other sources provided some key connections. The ACM code of ethics overlaps strongly with principalism but also highlights the duty to ensure that users have informed expectations, by providing "full disclosure of all pertinent system capabilities, limitations, and potential problems" [23]. The EPSRC code explicitly stipulated a duty to avoid deception, specifically in relation to vulnerable users [12]. Another source relates mainly to development rather than application. Psychology has codes and systems designed to protect participants in experiments. They transfer to the closely related issues that arise when subjects are recorded for affective databases [29].

Another expansion highlighted a feature of principalism. "First, do no harm" is a famous principle in bioethics. In keeping with that, there tends to be more written about harms to be avoided than goods to be gained. However, principalism implies a duty to think through the good we might do. An attempt to do that identified four areas [30]:

1) humanizing electronic communication;
2) broadening access;
3) additional resources to make judgments;
4) therapeutic applications.

This article also noted that affective computing may contribute to enjoyment, but questioned whether that can be counted as morally significant. This raises issues that are taken up later.

More recently, IEEE has put forward recommendations on a range of ethical issues affecting the field [24]: how affect varies across human cultures, the particular problems of artifacts designed for caring and private relationships, considerations of how intelligent artifacts may be used for "nudging," how systems can support human flourishing, and appropriate policy interventions for artifacts designed with inbuilt affective systems.

The sequence of developments reflects a concern to move beyond a priori, general formulations such as principalism, by drawing out ethical issues that arise in the particular area and proposing responses. The next part of the discussion extends that pattern.

B. Affective Computing: Recent Progress

Aspirational discussions of ethics can proceed without looking too closely at the technical state-of-the-art. This is not the case for a pragmatic discussion. The current work identifies a way of subdividing the overall task, and the different subtasks raise different ethical issues. Similarly, varied application areas have emerged, and they also raise different questions. The tradeoffs associated with a particular subtask or application depend on the technical difficulties that it poses. They involve the ethical merits of using current solutions, the ethical demerits, and the costs involved in improving the balance. This section offers a broad view of the state-of-the-art from that perspective.

1) Overall Breakdown of Tasks: The breakdown of tasks involves several layers. There is separation between recognition of human affective states and action designed to influence human states. Recognition involves acquiring inputs and mapping them onto a representation of relevant affective states. The mapping rules are generated by applying learning algorithms to a database containing labeled data. The actions that follow recognition may involve notification—informing a relevant actor (usually, but not always a human) or engaging with the person whose emotion was recognized. The bases for deciding how to engage are very varied.

The different areas raise different ethical considerations and balances.

2) Representation and Data: Representation is pivotal. Considerable effort has gone into developing appropriate representations of the relevant states. Although there is no consensus, most affective analyses are trained and evaluated based on two types of emotion models. Discrete models describe states using a set of classes, usually some variant on the "basic emotions" of fear, anger, surprise, sadness, happiness, and disgust. Dimensional models describe them in terms of continuous dimensions (commonly pleasure, arousal, dominance, and expectation).

There is no doubt that representations like that have merits, but straightforward versions do not capture aspects of affective states that there are reasons to think are significant. One is intentionality. Standard descriptions do not specify what an emotion is about, but philosophers argue that that is essential to understanding it [31]. Another is multiplicity: a substantial proportion of emotional episodes involve "mixed emotions" [32], [33]. A third is control: controlled anger and controlled joy are very different from the uncontrolled versions.

Those are technical points, but they also have an ethical dimension. It involves a principle centered on respect for other people: we should take pains to avoid basing actions toward them on an impoverished representation of their affective state. Using representations that we know exclude key issues sits uncomfortably with that.

In a real application, this is always in tension with the requirement for a representation simple enough to use. It also needs to be set in context. A representation may be inadequate, but often the alternative is that actions toward people will take no account of their affective state. This may involve a machine interacting with no reference to affective issues in the interaction or a human proceeding on default (and usually optimistic) assumptions about them. This illustrates a very characteristic feature of the area: the ethical challenge is to strike the best balance we can.

Representations reflect information derived from raw data. The issue of impoverishment affects data too. Early work was influenced by theories, which implied a privileged relationship between affect and a particular kind of signal—physiological (following James [34]) or facial (following Ekman and Friesen [35]). It has become clear that information is distributed across modalities, and relying on any subset risks simplification or outright error. In emotionally charged discussions, it is common for voice and face to reflect different parts of the speaker's affect—negative toward events being described, positive about being able to describe them [32], [33]. More dramatic examples come from sports. Competitors show facial expressions that viewers take to imply intense distress when they are presented in isolation. In reality, they are ambiguous: information from posture reveals that although some do signify distress, others signify triumph [36]. Context is also relevant there (whether a point has just been won or lost). It becomes crucial in settings where posture is less expressive, such as a game of chess [37]. Against that background, the restricted use of modalities also raises ethical issues. It means that basing judgments on information that we know may lead to a serious misinterpretation of the person's state.

Databases bring together all of the issues just considered and more. They contain recorded signals in particular modalities and descriptions of the emotions that humans perceive in them, couched in a specified representation. Recognition rules are then developed with the aim of reproducing the signal/representation relationships that occur in the database.

In the usual pattern, work on databases involves tradeoffs. Constructing a database involves a lot of effort. This is reflected in the fact that many studies return to the same databases [38] and a few corpora, such as IEMOCAP [39] or MSP-Improv [40], feature very frequently. The tradeoff question is whether the effort that would be required to overcome the limitations of existing databases is justified by the problems that the limitations pose. What is to be considered in this article is the moral aspect of the question. Most obviously, the limitations mean treating people who interact with affective systems as if their emotions were like those that are reflected in a database. This is ethically troubling when we know that there are likely to be major differences.

Four broad dimensions of difference suggest that effort is ethically required.

1) A very obvious difference is between naturally occurring emotions and acted simulations. It is much easier to acquire acted data, and this is reflected in the table of frequently cited databases given by Akçay and Oğuz [38]: 13 use acted data, as against seven which use natural.
2) There are obvious doubts about treating people who interact with affective systems as if their emotions could only be like actors' portrayals. This is partly because of a second dimension. Expressions of emotion are contextual. We would not expect children to show happiness during a mathematics lesson in the same way that they would during a football match. On that basis, it matters whether databases cover contexts like those where extracted rules will be applied.
3) A third, relatively high-profile issue involves group differences. It is no longer in doubt that culture affects both the way emotions are expressed [41] and the emotions underlying the expressions [42]. Hence, it is very problematic to attribute emotion to people from one culture by way of a database showing people from another. Blindness to age differences has also been noted in commercial face emotion recognition systems [43]. An area where differences are widely acknowledged is gender: procedures designed to take account of the differences are mentioned in the following.
4) Last, but not least, there are individual differences in expression of emotion, as elsewhere. An analysis [44] identified seven dimensions of variation. To reflect that kind of variation, the numbers of people contributing to a database should be large—at least 2^7 (that is the number needed to give representatives at different points on each of the seven dimensions). The table from Akçay and Oğuz [38] shows only four databases at that level.

To say that there are ethical issues in this area is not to condemn the research community. Theories that influenced early work in the field suggested that variation would be much narrower, and evidence to the contrary has emerged slowly. Responding to the evidence is a prime case of distributed responsibility. The development of very large databases requires the work to be funded, and given academic recognition. To be viable, it probably needs international collaboration. The research community can argue for those, but others have to bring them about.

Related issues extend beyond databases to the procedures used to extract rules from them. Choices there depend on the extent to which humans impose deliberate structure. Earlier approaches left a good deal of scope to explore techniques that might be appropriate to particular goals. The process began with extraction of theoretically motivated features. Then, for each modality, a kind of learning that seemed likely to be appropriate (such as a support vector machine or deep learning) was applied, aiming to assign theoretically motivated labels; and another stage of learning provided fusion. In contrast, recent work has emphasized “end-to-end” processing, typically involving deep learning, where relationships are driven by data, with minimal input from theory. This is reported to benefit overall accuracy. However, this is not the only possible measure of success, and it is not obvious how other measures can be addressed in a deep learning paradigm [45]. The ability to preserve ethically significant differences is one of the relevant issues. There are interesting pointers to the way that kind of concern might be handled. Recent research has combined end-to-end deep learning with pretrained models, wav2vec2 for acoustic and BERT (FlauBERT in French) for verbal content [46]. Another study indicates that overall performance benefits from coupling two learning processes, one aimed at discriminating emotions and the other at discriminating gender [47]. It is hard to predict whether similar approaches might apply to other issues mentioned earlier, such as intentionality, mixed emotions, and control.

3) Considering the Actions That Systems May Take: It is easy to think about recognition in technical rather than ethical terms. Ethical issues arise because they provide the basis for actions that will impinge in some way on humans. When we are considering the types of action that systems may take, the ethical aspect is generally more obvious: we can see the benefits that are to be anticipated. It makes sense to approach that area by considering the applications that are realistic prospects and prospective benefits.

A striking point is that a very substantial range of applications clearly stand to offer substantial benefits to substantial numbers of people if they can be done well. This points to an ethical conclusion: if the discipline is in a position to provide people with those benefits, then it has a moral duty to try. However, the qualification that they should be done well is significant. There are balances to be weighed because flawed efforts are liable to have ethically unacceptable effects.

One issue is the kind of problem considered in connection with recognition. Dealing with affective phenomena, it is crucial that actions are attuned to the person and circumstances. Recognition deals with one side, the person. Related problems arise with the circumstances. Learning techniques try to predict the empathic response that is appropriate given a preceding dialog [48]. They are increasingly sophisticated, but as with recognition, it is important to consider whether they have taken account of all the ethically significant dimensions of variation.

A rather different issue is appropriate targeting. It is easy to judge systems by their success in creating positive feelings. However, this is dangerous unless we also take care that the feelings are in keeping with reality. For example, we should not build artificial driving instructors that make people feel better about their driving but do not improve their skills. The philosophers’ emphasis on intentionality helps to express the point. The ethically appropriate target is to establish feelings that are both positive and well grounded—that is, the feelings reflect the reality of what they are about.

Those points draw out intuitions that are relevant to a substantial range of actions that systems may take. Other, specific issues arise in connection with individual applications.

4) General Facilitation: A very broad, basic goal was mentioned early on: making it likelier that people’s interactions with machines will be effective and unstressful. Given that those interactions pervade contemporary life, the general ethical case for pursuing the goal seems obvious. There are specific contexts, considered in the following, where the case seems strong. Key difficulties are perhaps clearest in the general form, involving digital assistants, or chatbots.

Chatbots illustrate the issue of feelings about something that are well grounded as well as positive. Luger and Sellen [49] gave a memorable description of using one: “like having a really bad PA.” Initial interactions, using humor, did generate positive feelings toward the system. However, the feelings were not well grounded: they led people to expect human-like competence which, in reality, the system did not have. Competence is not the only kind of human attribute that may be attributed. For example, a recent study described informants who consider a chatbot, Alexa, as “girlfriend, mistress or wife, and compare Alexa with their real girlfriend or wife” [50].

The point here is not to pass an ethical judgment on outcomes like that. It is to illustrate the kind of ethical question that arises when systems arouse feelings that are positive but not well grounded in reality. As elsewhere, the question involves balance. Ethics presses us to avoid arousing feelings that are not grounded in reality. However, it also presses us to provide help for people who struggle to navigate the web, given its importance in contemporary life [51]. Striking a balance is not straightforward.

Chatbots highlight an issue that calls for clarification. They invite overestimates of their similarity to humans. On the principle that feelings should be in keeping with reality, this is ethically problematic, and it matters to understand what provokes it. Generality seems relevant. A chatbot interacts with a user across a broad range of tasks, as a human might. This invites a sense of broad similarity to a human, which seems less likely with systems whose operation is confined to a specific context. However, if that issue is reduced, systems oriented to specific applications also raise specific ethical concerns.

5) Issues Specific to Particular Applications: Particularist analysis indicates that an appropriate structure needs to distinguish applications; bring out the ethical issues that arise, positive and negative; and reflect their interplays. As a working approach, applications can be divided first by the kind of action they take toward the user—conversing, passing information about one person to another person or system, or more direct intervention. This grouping brings out different kinds of ethical issues. The issues include associated ethical positives—generally reflecting the purpose for which the system was intended. Ethical negatives, which are generally not intended, emerge in the attempt to achieve the positives.

6) Systems That Pass Information: There is a range of cases where the system does not take actions directed to the user, but it provides information or controls a mechanical adjustment. They include a variety of application areas where there are obvious ethical reasons to invest effort. The differences between them draw attention to a less obvious ethical dimension.

Recognizing patterns of emotion contributes to a range of medical diagnoses—mainly, but not all, psychiatric. In that area, human clinical judgment is not necessarily secure; and where technologies could be developed to supplement it, it is hard to doubt that they should be. A recent review by clinicians identified multiple areas where that is a realistic prospect: the autism spectrum, depression, schizophrenia, amyotrophic lateral sclerosis, locked-in syndrome, and forensic psychiatry [52]. A nonpsychiatric application in medicine is pain detection [53]. Clinical value and limitations need to be properly assessed and understood, but that holds for any aid to diagnosis.

Nonclinical, but with similarities, are applications concerned with detecting comfort and stress. A range of techniques have been used to detect stress, with input from wearable devices being a notable source [54]. Contexts are also varied. For example, one task is recognizing when the temperature in a vehicle is causing discomfort to passengers [55]. Another is recognizing when arrangements in a workplace are allowing people to operate in a state of "flow," which is both subjectively positive and effective [56]. Often, these are areas where older methods could be used (thermometers or questionnaires), but the new techniques are less simplistic or troublesome.

Beyond those are methods concerned, broadly speaking, with assessing what people like. Again, wearable technology features in assessing responses to films [57] or advertisements [58]. There are also techniques that draw patterns from inputs to social media. They range from identifying emotionally charged comments about a specified topic (e.g., a product) [59] to formulating what makes particular content inspiring [60].

The range from medical diagnosis to market research brings out another broad ethical issue, which might not come to mind immediately. In brief, it is equality. Equality is a troublesome concept, but there is no doubt that it carries ethical power [61]. In that context, it is notable that medical applications help people with a poor quality of life to reach a level that most people take for granted. In contrast, the obvious assumption is that making advertisements more effective will benefit a group who are already more powerful and wealthy than most and increase their advantage. This seems to be part of an intuition that the first has a positive ethical appeal, which the second does not.

7) Systems That Make Emotive Interventions: The archetypal applications of affective computing involve systems that both assess users' affective state and also go on to take actions designed to affect it.

This kind of system tends to evoke generalized ethical alarm. It is to the effect that we are creating systems that have very powerful human abilities—enabling them to sway people in a way that bypasses reason but that are not subject to the controls that apply in humans—notably conscience and social pressure. However, as elsewhere, it is useful to approach the issue by way of specific applications, where the ethical positives are apparent.

Intelligent tutoring is one key area. It seems clear that intelligent tutoring systems can improve learning, compared to other readily available options [62]. We would expect intuitively that some kinds of emotional rapport would enhance the process. The case is not clear-cut [63], but a broad review gives grounds for optimism. Echoing an earlier point, the most promising applications target specific skills, such as mathematics or handwriting [64].

Certain kinds of therapy are closely related. There is evidence that robots with appropriate affective skills can help children on the autism spectrum to acquire social skills [65]. There is also interest in affectively engaging games as a therapeutic tool [66]. Since there are wide differences between individuals, it is important that systems are attuned to the characteristics of the individual, and emotion recognition techniques have addressed that issue [67]. Chatbots have been used to deliver therapy for depression, and adding empathic features is expected to enhance them [68]. Affective computing also features in therapies for depression and generalized anxiety disorder, by directing the therapist rather than the patient. Sentiment analysis allows the therapist to recognize the kind of interaction that succeeds in influencing the patient's emotions [69].

In tutoring and therapy, the overwhelming ethical issue is validation. This links back to an issue raised in the context of databases. The affective computing community needs external support to conduct research capable of providing solid evidence on validity or otherwise. Providing support is an ethical matter.

Another substantial body of work involves shaping feelings not toward humans but toward machines. The goal is stated in various ways, but one that clarifies the ethical positive is to establish feelings that allow people to work effectively with the machines. Two steps toward the goal are widely recognized. Effective cooperation depends on enabling trust, and trust follows from attribution of human-like characteristics [70]. Affective routes are central to evoking anthropomorphic impressions. They include personality-driven expressions of emotion [71] and appropriately chosen forms of humor [72].

The issues in that area are profoundly difficult. The question is whether an ethically desirable end—enabling people to work with machines—justifies means that are obviously problematic. Anthropomorphism is problematic because it goes against the principle that feelings should be in keeping with reality. Trust is a notoriously dangerous aspect of humanity. It is a feature of the way we are that trust can be won by routes that are very far from justifying it. Creating new ways to win it is ethically disturbing. The problem is that it seems to be key to the way we interact with complex agents, and if we need those interactions, winning trust is a necessary step.

8) Happiness and Reality: A tension between two ethical principles was introduced earlier. One is that we should aim to promote positive affect if we can. The other is that we should take care that the feelings are in keeping with reality. The tension has been reflected at various points since. However, there are major applications where it is central.

Social isolation is a major issue. It is often discussed in connection with groups who have limited mobility or access to communications, notably people who are elderly or cognitively impaired, but surveys show that it is very widely distributed [73]. A long-standing proposal is that artificial companions could provide a remedy [74]. Versions that are now generally available include empathic and emotional abilities [75].

There is evidence that companions of that kind can affect feelings positively, by alleviating loneliness [76]. Nevertheless, the area has provoked extended ethical discussion, involving issues such as deception, informed consent, monitoring and tracking, and perpetuating real social isolation [77]. The underlying theme is a mismatch between positive feelings and negative reality.

A last substantial application raises a related issue. Affective computing has become an established part of computer gaming [78]. It is the essence of gaming that it arouses desirable feelings by directing users away from reality. In terms of "the greatest happiness," very little would seem to be more ethically appropriate than enabling people to immerse themselves in rewarding games. However, in the usual pattern, reality involves balances. Happiness in the short term needs to be balanced with the risk of addiction in the long term: more is said about that in Section IV. A subtler issue is the principle that individuals should be enabled to fulfill their potential. Few would think that a life happily immersed in games, insulated from the challenges of reality, met that requirement. In another recurring theme, it is also far from obvious whose responsibility it is to ensure that appropriate balances are kept.

IV. ACUTE ETHICAL CONCERNS

Section III offered a broad picture of affective computing as it currently is. This includes acknowledging that ethical issues are pervasive, and they regularly involve tradeoffs: it is the norm that developments aimed at ethically admirable goals will also pose ethical problems. As a result, research is continually faced with judgments about balances. This is central, but not the whole picture. There are conceivable developments that are very clearly to be resisted on ethical grounds. This section turns to those areas.

Reflecting points made by Floridi et al. [10], [19], [20], many acute issues arise, not from technology as such, but from its interaction with social structure. Particularly striking are scenarios involving selective empowerment. The problem is that those whose power enables them to acquire control over the technology may thereby acquire new power and disempower others.

A prominent issue in that category, highlighted by McStay and others [79], is emotional surveillance. Examples are widespread. Preparing for the 2012 London Olympics, British security services conducted sentiment analysis on thousands of social media platforms, looking for changes that might signal preparation for protest or terrorism. Examples of visual monitoring include children's emotions in a classroom and women's faces in a city for signs of possible harassment [80]. Those involve double asymmetries—the fact that authorities presume to amass deeply sensitive information about ordinary people, and the power that systems and information like that give them if they choose to use them. Both are redoubled by the disrespect implicit in accepting information generated by systems that are known to be inaccurate.

On a similar level is emotional manipulation of beliefs and evaluations—in brief, indoctrination. Techniques that were described earlier enable agents to shape people's understanding by appealing to feelings and bypassing rational evaluation. This may involve establishing trust in a persuasive agent or selecting forms of expression that associate beliefs or evaluations with satisfying feelings. Recently, there has been much discussion of related techniques described as "nudges" [81], [82]. It reflects ancient concerns. Socrates memorably condemned speakers who misled the public by presenting ideas that were insubstantial, or toxic, in an alluring way: he called them "pastry chefs" [83]. Again, the threat is that those who control our artificial pastry chefs will acquire far greater power than their ancient counterparts.

At the other end of the scale is the risk of denying the least privileged their due as human beings. This is clear in concerns about artificial companions. They offer a way to give the appearance of caring for disadvantaged groups, where in reality they are being denied the human contact that they are ethically due. Similar issues apply in medical and therapeutic applications: they offer a way to give patients less than the full human attention that they are ethically entitled to.

Power differentials also affect issues in recognition. When systems are attuned to the way emotion is expressed by some groups and not others, it would be very surprising if the attunement did not favor majority and/or dominant groups.

In all of those areas, there is unmistakable potential for profoundly unethical use of affective technology. The issue is whether people's attitudes to each other will draw them toward those uses or to uses where the balance is clearly positive. This is not in the control of the affective computing community. What it can do is to alert people to questions about their attitudes to each other, which take on a new significance in the context of the new technology.

Two other acute concerns are at least somewhat separate.

The first is addiction. It is not in doubt that there are issues in that area. In the standard psychiatric manual, DSM-5, Internet gaming disorder (IGD) is recognized as a clinical condition, and social media addiction (SMA) is identified as a "condition for further study" [84]. Studies of IGD report prevalence rates of around 10% among gamers [85]. This clearly raises questions about the ethics of creating progressively more emotionally engaging games and social media experiences.

Clearly, the issue needs to be seen in perspective. It is a feature of humanity that people are susceptible to addiction: they are drawn to affectively satisfying experiences despite the costs. Affective computing does not create that problem, and it is not clear how it affects the risk that people will be drawn into addiction of some kind. The immediate ethical obligation is to understand how the overall cost of addiction is altered by the new ways of providing affectively satisfying experiences.

The last acute concern to consider is deception. Deliberate manipulation of beliefs has been considered already, but it is argued that a deeper kind of deception is central to the area [86]. It arises when people are drawn to regard an artificial agent as humanlike. The point was made earlier that there are marked tensions in that area. On one side, a great deal of writing presents it as a goal that systems should be accepted as humanlike, and evaluations use it as a measure of success. On the other hand, it is argued that evoking anthropomorphic impressions is deeply unethical. A robot does not have emotions in the sense that humans do, and techniques that lead people to think it does are unethical. More fundamentally, techniques that blur the difference between robot and human draw people into a false understanding of their own humanity. Several kinds of responses are worth considering.

An appealingly simple response is to say that builders have an obligation to tell users clearly that their products are artifacts, not humans. The problem there arises from the intricacies of the relevant human systems. When people decode emotion-related signals, it involves complex nonconscious processes: activating the systems within themselves that would lead them to give similar signals and attributing similar internal activity to the source [14]. A factual disclaimer does not stop that process.

A second response is that we should avoid deception by giving robots their "own emotions" [87]. This raises two obvious objections. One is that emotion is not a thing to give robots. It would be unethical in the extreme to equip them with human-like tendencies to suspend reason and take unrestrained action. The second is that the idea is science fiction verging on fantasy: what constitutes human emotion is a feeling and this is something robots cannot possibly have. Both points carry weight, but they do not close the issue. They identify features of human emotion that should not or cannot be incorporated into robots. However, human emotions involve multiple features, and it is possible to envisage incorporating others into robots in ways that at least mitigate the problem of deception—for instance, ensuring that signals associated with friendliness correlate with attaching a high priority to goals involving the recipient's welfare. It is conceivable that adjustments along those lines could mitigate the problem of deception without disposing the robot to ethically unacceptable action. It is a matter of detailed work to establish how successful that kind of development could actually be.

A third response is that if affective agents are to become part of society, education needs to equip people with appropriate ways to understand them. On one side, this involves finding and teaching ways to override processes whose natural tendency is to credit affective agents with too many human characteristics; on the other side, this involves ways to establish reactions that are more appropriate. Educational elements probably need to be coupled with ensuring that agents give usable signals of the ways they differ from humans.

This kind of education is not simply a cost to be paid for the privileges that advanced education can bring. Understanding how humans can resemble machines in diverse ways, and yet be fundamentally different, can help to clarify what it means to be human. This kind of clarification is itself ethically significant.

V. TOWARD AN ETHICAL ECOSYSTEM

Few dispute that there are ethical issues surrounding affective computing. The challenge is first, how to understand them; and second, what kind of action to take. The discussion so far has focused on the first. This provides a context for considering proposals on the second. There is a difficult balance to be struck between benefits and risks, including the costs of mitigating risks.

As a rough guide, it is useful to separate three broad areas of action. One that is well recognized involves the development of codes to regulate the development and use of affective systems. A second is ensuring that the expert communities concerned are suitably informed about the full range of issues relevant to ethical judgments. The third is ensuring that the general public is suitably informed about affective systems that are likely to impact their lives.

A. Regulation

Regulatory codes already exist (see Section III-A), but they need to be updated as the implications of research become clearer. Major efforts are currently underway for an AI Act developed by the EU. As pointed out in Article 19 of the proposed AI Act, emotion recognition is a highly invasive form of surveillance that "involves the massive collection of sensitive personal data in an invisible and unaccountable manner, enabling the tracking, monitoring, and profiling of individuals, often in real time."

The EU has much to gain by shaping a framework for the use of AI that is associated with trust and respect for human rights of individuals and the rule of law. The AI Act focuses on identifying applications deemed to pose risk and therefore to need regulation. Different levels attract different actions. Areas that involve affect are prominent. Applications involving unacceptable risk are prohibited. They include manipulative "subliminal techniques," those that exploit specific vulnerable groups (with physical or mental disabilities), or are used by public authorities for social scoring purposes. High-risk applications are tightly regulated. They include education and vocational training, and worker management.

The other types of proposal involve normalization. Several projects aim to propose standards for affective computing. They include the IEEE P7014 project on emulated empathy, ISO propositions on the representation/annotation of emotions, and European CEN-CENELEC Joint Technical Committee 21 "Artificial Intelligence." A notable feature is risk analysis for data and algorithms, including the issue of emotional manipulation in nudging systems.

In many cases, it is impossible to know in advance the consequences of massive, intense, and repetitive use of these technologies. Building a multidisciplinary monitoring ecosystem to avoid addiction, isolation, and manipulation/nudging is a societal and political challenge. There is a significant cost to this balancing act.

B. Awareness in Expert Communities

Research has underlined the intricacy, power, and pervasiveness of affective phenomena in human life (see Section II-A). Their pervasiveness means that ethical issues arise when people are required to interact with systems that take no account of that aspect of human life or reflect simplified conceptions of it.

Addressing those issues requires that the people who design, evaluate, and deploy artificial systems should understand how complex emotions are and how they pervade our choices and relationships with others.

On one side, sophisticated awareness is fundamental to describing and monitoring the potential ethical problems in technologies that engage with human affect. This includes recognizing that there are many gray areas in all types of applications. Many industrial companies are already moving in that direction by developing internal ethical guidelines for good practices.

The other side is recognizing how many areas stand to benefit from developments that take appropriate account of the way affect runs through human life. This kind of understanding has gained ground in some areas, for instance, in education or health. However, the potential is much wider.

As an example, human language is an essential element in shaping, even determining, cultural characteristics, human perceptions, and even entire worldviews. Affect is integral to it: it reflects actual and expected feeling about the things it describes, the things being said, and the participants saying them. However, the affective language of conversational agents captures very little of that. It reflects a linguistic universe without deep lying evaluations of things talked about or people talked to. In one sense, tomorrow they could express themselves better than many of us... and seem "more intelligent." In another sense, their expression remains shallow. Working out how to address that anomaly is hugely challenging, not least because it is charged with ethical significance.

Not least, it matters that the community is equipped to respond when dramatic questions about the enterprise of AI are given a high public profile. Claims about the threat that AI poses appear regularly in the media, and it matters that the expert community can rebut them effectively. Among other things, a convincing response should show that people active in the field are fully aware of the way developments in their area may impact humanity as a whole, for better or worse. This includes reassuring the public that the community's ethical outlook is sensitive to intuitive concerns and not restricted to observing a set of formal guidelines.

There are active moves to establish informed ethical engagement in the affective computing community. Significant examples featured in the major conference in the area, Affective Computing and Intelligent Interaction (ACII 2023), held while this article was nearing completion. It included a workshop on "Moral Imagination in Affective Computing" [88] and a tutorial on the European Community's AI Act [89], which sets out formal responses to a wide range of issues mentioned here.

C. Development of Public Awareness

The challenge of ensuring appropriate kinds of understanding is not confined to experts. It is increasingly the case that the way we socialize with machines, and perhaps with humans, is being modified by the emergence of interactive and adaptive systems using emotions. The change calls for education of citizens on digital interactions, including digital ethics. This takes time and must be reinforced from childhood at school, especially for affective machines. The issues cover a wide range of ethical topics: censorship of content, discrimination and unfairness, manipulation, marketing to children, and changing our collective behaviors. These areas inspire critical questions centered on the ethics, goals, and deployment of innovative products that can change our lives and society. Awareness of the questions is key to ensuring that society learns how to use the new technology in a way that avoids the risks and gains the benefits.

VI. CONCLUSION

Affective computing brings the ability to build systems that are very unlike anything that has existed before and potentially touch many areas of life. Recognizing the depth and breadth of its impact suggests a deceptively simple conclusion: that developing the area should be understood as an ethical matter. Whatever is done should be done with awareness that it may bring significant goods to significant numbers of people or inflict significant wrongs on them. This depends on taking pains to understand the kinds of good and the kinds of wrong that may be done and the combinations that are likely to arise from developments that we can envisage. It carries an obligation to share that understanding with authorities who are in a position to legislate against developments that advantage some at an unacceptable cost to others and with people in general whose ability to benefit depends on understanding what they are dealing with.

What we have done in this article is to encourage a broadly particularist outlook on those ethical issues. On that understanding, the basis for moral judgment involves understanding the situations that are to be faced, with their various complexities. If so, the first obligation is for the affective computing community to develop that kind of understanding. Explicit principles are not a replacement for it. On the contrary, it needs to underpin the way we formulate and apply them.

This is plainly an idealistic proposition. However, it is a proposition that arises naturally from the material that has been summarized. It makes sense to set it out and invite people to consider it. ■

REFERENCES

[1] R. Cowie, "Describing the forms of emotional colouring that pervade everyday life," in *The Oxford Handbook of Philosophy of Emotion The Oxford Handbook of Philosophy of Emotion*, P. Goldie, Ed. Oxford, U.K.: Oxford Univ. Press, 2009.

[2] A. Damasio, *Descartes' Error*. New York, NY, USA: Random House, 1994.

[3] R. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 2000.

[4] R. Cowie et al., "Emotion recognition in human–computer interaction," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32–80, Jan. 2001.

[5] E. Douglas-Cowie et al., "The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.* Berlin, Germany: Springer, 2007, pp. 488–500.

[6] L. Devillers, "Human–robot interactions and affective computing: The ethical implications," in *Robotics, AI, and Humanity*. Cham, Switzerland: Springer, 2021, pp. 205–211.

[7] B. Reeves and C. Nass, *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. New York, NY, USA: Cambridge Univ. Press/ICSLI, 1996.

[8] L. Devillers, F. Fogelman-Soulié, and R. Baeza-Yates, "AI & human values," in *Reflections on Artificial Intelligence for Humanity*. Cham, Switzerland: Springer, 2021, pp. 76–89.

[9] R. Cowie, "The enduring basis of emotional episodes: Towards a capacious overview," in *Proc. Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2015, pp. 98–104.

[10] L. Floridi, *The Ethics of Information*. Oxford, U.K.: Oxford Univ. Press, 2013.

[11] J. Dancy, "Moral particularism," in *The Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), E. N. Zalta Ed. Stanford School of Humanities and Sciences: Stanford, CA, USA, Sep. 2017. [Online]. Available: https://plato.stanford.edu/archives/win2017/entries/moral-particularism/

[12] T. Dixon, "'Emotion': The history of a keyword in crisis," *Emotion Rev.*, vol. 4, no. 4, pp. 338–344, 2012.

[13] L. Pessoa, "A network model of the emotional brain," *Trends Cogn. Sci.*, vol. 21, no. 5, pp. 357–371, May 2017.

[14] B. Wicker et al., "Both of us disgusted in My insula: The common neural basis of seeing and feeling disgust," *Neuron*, vol. 40, no. 3, pp. 655–664, Oct. 2003.

[15] S. Tomkins, "Affect theory," in *Approaches to Emotion*, K. Scherer and P. Ekman, Eds. Mahwah, NJ, USA: Lawrence Erlbaum Associates, 1984, pp. 163–195.

[16] C. Cowie, "A new argument for moral error theory," *Nous*, vol. 56, no. 2, pp. 276–294, 2022.

[17] J. Neu, "An ethics of emotion?" in *The Oxford Handbook of Philosophy of Emotion*, P. Goldie, Ed. Oxford, U.K.: Oxford Univ. Press, 2009.

[18] L. Floridi, M. Holweg, M. Taddeo, J. A. Silva, J. Mökander, and Y. Wen, "capAI—A procedure for conducting conformity assessment of AI systems in line with the EU Artificial Intelligence Act," Mar. 2022. [Online]. Available: https://ssrn.com/abstract=4064091. Accessed: Sep. 26, 2023, doi: 10.2139/ssrn.4064091.

[19] L. Floridi, "Information ethics: An environmental approach to the digital divide," *Philosophy Contemp. World*, vol. 9, no. 1, pp. 39–45, Spring/Summer 2001.

[20] L. Floridi, "Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions," *Phil. Trans. R. Soc. A*, vol. 374, Dec. 2016, Art. no. 20160112, doi: 10.1098/rsta.2016.0112.

[21] L. May, *The Morality of Groups: Collective Responsibility, Group-Based Harm, and Corporate Rights*. Notre Dame, IN, USA: Univ. Notre Dame Press, 1987.

[22] M. A. Boden et al., "*Principles of Robotics*. Swindon, U.K.: The United Kingdom's Engineering and Physical Sciences Research Council (EPSRC), 2011.

[23] D. W. Gotterbarn et al. (2018). *ACM Code of Ethics and Professional Conduct*. Accessed: Sep. 26, 2023. [Online]. Available: https://dora.dmu.ac.uk/server/api/core/bitstreams/1e5b3cb8-2d77-4ab4-885b-d5996b74605f/content

[24] *The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems: Affective Computing*, IEEE Standards Association, Piscataway, NJ, USA, 2018.

[25] J. Flynn, "Theory and bioethics," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed. Stanford School of Humanities and Sciences: Stanford, CA, USA, 2021. [Online]. Available: https://plato.stanford.edu/archives/spr2021/entries/theory-bioethics/

[26] P. Petta, C. Pelachaud, and R. Cowie, *Emotion-Oriented Systems: The HUMAINE Handbook*. Berlin, Germany: Springer, 2011.

[27] R. A. Calvo, S. D'Mello, J. M. Gratch, and A. Kappas, *The Oxford Handbook of Affective Computing* (Oxford Library of Psychology). Oxford, U.K.: Oxford Univ. Press, 2015.

[28] S. Döring, P. Goldie, and S. McGuinness, "Principalism: A method for the ethics of emotion-oriented machines," in *Emotion-Oriented Systems*. Berlin, Germany: Springer, 2011, pp. 713–724.

[29] I. Sneddon, P. Goldie, and P. Petta, "Ethics in emotion-oriented systems: The challenges for an ethics committee," in *Emotion-Oriented Systems*. Berlin, Germany: Springer, 2011, pp. 753–767.

[30] R. Cowie, "The good our field can hope to do, the harm it should avoid," *IEEE Trans. Affect. Comput.*, vol. 3, no. 4, pp. 410–423, 4th Quart., 2012.

[31] P. Goldie, *The Emotions: A Philosophical Exploration*. Oxford, U.K.: Oxford Univ. Press, 2000.

[32] E. Douglas-Cowie et al., "Multimodal databases of everyday emotion: Facing up to complexity," in *Proc. Interspeech*, Sep. 2005, pp. 813–816.

[33] L. Devillers, L. Vidrascu, and L. Lamel, "Emotion detection in real-life spoken dialogs recorded in call center," *J. Neural Netw.*, vol. 18, no. 4, pp. 407–422, 2005.

[34] W. James, "What is an emotion?" *Mind*, vol. 34, pp. 188–205, Apr. 1884.

[35] P. Ekman and W. V. Friesen, "The repertoire of nonverbal behavior: Categories, origins, usage, and coding," *Semiotica*, vol. 1, no. 1, pp. 49–98, 1969.

[36] H. Aviezer, N. Ensenberg, and R. R. Hassin, "The inherently contextualized nature of facial emotion

perception," *Current Opinion Psychol.*, vol. 17, pp. 47–54, Oct. 2017.

[37] G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva, and P. W. Mcowan, "Context-sensitive affect recognition for a robotic game companion," *ACM Trans. Interact. Intell. Syst.*, vol. 4, no. 2, pp. 1–25, Jul. 2014.

[38] M. B Akçay and K Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Commun.*, vol.116, pp. 56–76, 2020.

[39] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.

[40] C. Busso et al., "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 67–80, Jan./Mar. 2017.

[41] R. E. Jack, O. G. B. Garrod, H. Yu, R. Caldara, and P. G. Schyns, "Facial expressions of emotion are not culturally universal," *Proc. Nat. Acad. Sci. USA*, vol. 109, no. 19, pp. 7241–7244, May 2012.

[42] N. Lim, "Cultural differences in emotion: Differences in emotional arousal level between the East and the West," *Integrative Med. Res.*, vol. 5, no. 2, pp. 105–109, Jun. 2016.

[43] E. Kim, D. Bryant, D. Srikanth, and A. Howard, "Age bias in emotion detection: An analysis of facial emotion recognition performance on young, middle age and older adults," in *Proc. AIES*, Jul. 2021, pp. 638–644.

[44] L. K. Barr, J. H. Kahn, and W. J. Schneider, "Individual differences in emotion expression: Hierarchical structure and relations with psychological distress," *J. Social Clin. Psychol.*, vol. 27, no. 10, pp. 1045–1077, Dec. 2008.

[45] J. Wagner et al., "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," 2022, *arXiv:2203.07378*.

[46] T. Deschamps-Berger, L. Lamel, and L. Devillers, "Exploring attention mechanisms for multimodal emotion recognition in an emergency call center corpus," in *Proc. ICASSP*, Jun. 2023, pp. 1–5.

[47] S. Ting-Wei, "End-to-end speech emotion recognition with gender information," *IEEE Access*, vol. 8, pp. 152423–152438, 2020.

[48] R. Goel, S. Susan, S. Vashisht, and A. Dhanda, "Emotion-aware transformer encoder for empathetic dialogue generation," in *Proc. 9th Int. Conf. Affect. Comput. Intell. Interact. Workshops Demos (ACIIW)*, Sep. 2021, pp. 1–6.

[49] E. Luger and A. Sellen, "'Like having a really bad PA': The Gulf between user expectation and experience of conversational agents," in *Proc. CHI Conf. Human Factors Comput. Syst.*, May 2016, pp. 5286–5297.

[50] Y. Gao, Z. Pan, H. Wang, and G. Chen, "Alexa, my love: Analyzing reviews of Amazon Echo," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov. (Smart-World/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, Oct. 2018, pp. 372–380.

[51] A. Baki and J. E. Kim. (2021). *How to Help People Navigate the Internet, Voice-First*. [Online]. Available: https://blog.google/technology/next-billion-users/voiceusers-playbook/

[52] K. Grabowski et al., "Emotional expression in psychiatric conditions: New technology for clinicians," *Psychiatry Clin. Neurosciences*, vol. 73, no. 2, pp. 50–62, Feb. 2019.

[53] Z. Chen, R. Ansari, and D. Wilkie, "Automated pain detection from facial expressions using FACS: A review," 2018, *arXiv:1811.07988*.

[54] T. A. Roldán-Rojo, E. Rendón-Veléz, and S. Carrizosa, "Stressors and algorithms used for stress detection: A review," in *Proc. 9th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep./Oct. 2021, pp. 1–8.

[55] T. Olugbade et al., "Toward intelligent car comfort sensing: New dataset and analysis of annotated physiological metrics," in *Proc. 9th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep./Oct. 2021, pp. 1–8.

[56] E. Di Lascio, S. Gashi, M. E. Debus, and S. Santini, "Automatic recognition of flow during work activities using context and physiological signals," in *Proc. 9th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep./Oct. 2021, pp. 1–8.

[57] (2016). *Hollywood is Tracking Heart Pounding Movie Scenes with Wearable Tech*. [Online]. Available: https://www.wareable.com/wearable-tech/heart-racing-bear-scenesthe-revenant-2186

[58] (2012). *Millward Brown, Affectiva: A New Way to Test Emotional Responses to Ads*. [Online]. Available: https://www.bizcommunity.com/Article/224/19/70834.html

[59] (2021). *Mood and Genre filters for 'Liked Songs'*. The Spotify Community. [Online]. Available: https://community.spotify.com/t5/Community-Blog/Mood-and-Genre-filters-for-Liked-Songs/ba-p/5160106

[60] O. Ignat, Y.-L. Boureau, J. A. Yu, and A. Halevy, "Detecting inspiring content on social media," in *Proc. 9th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2021, pp. 1–8.

[61] S. Gosepath, "Equality," in *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition), E. N. Zalta Ed. Stanford School of Humanities and Sciences: Stanford, CA, USA, 2021. [Online]. Available: https://plato.stanford.edu/archives/sum2021/entries/equality/

[62] J. A. Kulik and J. D. Fletcher, "Effectiveness of intelligent tutoring systems: A meta-analytic review," *Rev. Educ. Res.*, vol. 86, no. 1, pp. 42–78, Mar. 2016.

[63] M. Obaid et al., "Endowing a robotic tutor with empathic qualities: Design and pilot evaluation," *Int. J. Humanoid Robot.*, vol. 15, no. 6, Dec. 2018, Art. no. 1850025.

[64] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, and F. Tanaka, "Social robots for education: A review," *Sci. Robot.*, vol. 3, no. 21, 2018, Art. no. eaat5954, doi: 10.1126/scirobotics.aat5954.

[65] B. Scassellati et al., "Improving social skills in children with ASD using a long-term, in-home social robot," *Sci. Robot.*, vol. 3, no. 21, 2018, Art. no. eaat7544.

[66] B. Scassellati, M. Matarić, and H. Admoni, "Robots for use in autism research," *Annu. Rev. Biomed. Eng.*, vol. 14, nos. 3–4, pp. 275–294, 2012.

[67] Z. Shi, T. R. Groechel, S. Jain, K. Chima, O. Rudovic, and M. J. Matarić, "Toward personalized affect-aware socially assistive robot tutors for long-term interventions with children with autism," *ACM Trans. Human–Robot Interact.*, vol. 11, no. 4, pp. 1–28, Dec. 2022.

[68] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial," *JMIR Mental Health*, vol. 4, no. 2, p. e19, 2017.

[69] C. Jee and W. D. Heaven, "The therapists using AI to make therapy better," MIT Technol. Rev., Dec. 2021. Accessed: Sep. 26, 2023. [Online]. Available: https://www.technologyreview.com/2021/12/06/1041345/ai-nlp-mental-health-better-therapists-psychology-cbt/

[70] A. Waytz, J. Heafner, and N. Epley, "The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle," *J. Exp. Social Psychol.*, vol. 52, pp. 113–117, May 2014.

[71] X. Ma, E. Yang, and P. Fung, "Exploring perceived emotional intelligence of personality-driven virtual agents in handling user challenges," in *Proc. World Wide Web Conf.* San Francisco, CA, USA: Association for Computing Machinery, 2019, pp. 1222–1233.

[72] H. Ritschel, I. Aslan, D. Sedlbauer, and E. André, "Irony man: Augmenting a social robot with the ability to use irony in multimodal communication with humans," in *Proc. 18th Int. Conf. Auton. Agents MultiAgent Syst.*, 2019, pp. 86–94.

[73] Statista. (2022). *Loneliness among Adults Worldwide by Country 2021*. [Online]. Available: https://www.statista.com/statistics/1222815/loneliness-among-adults-by-country/

[74] Y. Wilks, "Artificial companions," in *Proc. Int. Workshop Mach. Learn. Multimodal Interact.* Berlin, Germany: Springer, 2004, pp. 36–45.

[75] Replika. *The AI Companion Who Cares*. Accessed: Sep. 26, 2023. [Online]. Available: https://replika.ai/

[76] D. Siemon et al., "Why do we turn to virtual companions? A text mining analysis of Replika reviews," in *Proc. 28th Americas Conf. Inf. Syst.*, Minneapolis, MN, USA, 2022, Paper no. 1288. [Online]. Available: https://aisel.aisnet.org/amcis2022/sig_hci/sig_hci/10/

[77] E. Portacolone, J. Halpern, J. Luxenberg, K. L. Harrison, and K. E. Covinsky, "Ethical issues raised by the introduction of artificial companions to older adults with cognitive impairment: A call for interdisciplinary collaborations," *J. Alzheimer's Disease*, vol. 76, no. 2, pp. 445–455, Jul. 2020.

[78] G. N. Yannakakis and A. Paiva, "Emotion in games," in *Oxford Handbook on Affective Computing*. Oxford, U.K.: Oxford Univ. Press, 2015, pp. 459–471.

[79] A. McStay and L. Urquhart, "Emotional artificial intelligence: Guidelines for ethical uses," Emotional.AL.org, 2019.

[80] R. Scarff, "Emotional artificial intelligence, emotional surveillance, and the right to freedom of thought," EasyChair, Manchester, U.K., White Paper no. 5371, 2021.

[81] R. Thaler and C. Sunstein, *Nudges*. Paris, France: Vuibert, 2017.

[82] A. Vugts, M. van den Hoven, E. De Vet, and M. Ver-Weij, "How autonomy is understood in discussions on the ethics of nudging," *Behavioural Public Policy*, vol. 4, no. 1, pp. 108–123, 2020.

[83] Plato, *Gorgias*, J. M. Cooper Ed. Indianapolis, IN, USA: Hackett Publishing, 1997.

[84] *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed. Washington, DC, USA: American Psychiatric Association, 2013.

[85] P. K. H. Chew and C. M. H. Wong, "Internet gaming disorder in the DSM-5: Personality and individual differences," *J. Technol. Behav. Sci.*, vol. 7, no. 4, pp. 516–523, Jul. 2022.

[86] A. Sharkey and N. Sharkey, "We need to talk about deception in social robotics!" *Ethics Inf. Technol.*, vol. 23, no. 4, pp. 309–316, Sep. 2021.

[87] K. LaGrandeur, "Emotion, artificial intelligence, and ethics," in *Beyond Artificial Intelligence*. Cham, Switzerland: Springer, 2015, pp. 97–109.

[88] (Sep. 10, 2023). *Moral Imagination in Affective Computing*. MIT Media Lab, Cambridge, MA, USA. [Online]. Available: https://sites.google.com/view/MoralImagination-ACII2023

[89] *The Potential Impact of the AI Act on Affective Computing Research and Development*. Accessed: Sep. 26, 2023. [Online]. Available: https://acii-conf.net/2023/program/tutorials/the-potential-impact-of-the-ai-act-on-affective-computing-research-and-development/

ABOUT THE AUTHORS

Laurence Devillers (Member, IEEE) is currently a Full Professor of artificial intelligence (AI) at Sorbonne University, Orsay, France, where she heads the research team on “Affective and social dimensions in Spoken interactions with (ro)bots: Technological and ethical issues” at CNRS-LISN. Since 2020, she has been leading the interdisciplinary Chair (including economists, linguists, and computer scientists) on AI and digital nudge HUMAAINE: HUman-MAchine Affective INteraction & Ethics at CNRS. She is responsible for the JTC21/CEN_CENELEC WG4 on Foundational and Societal Impact of AI that includes AI-Enhanced Nudging, Trustworthiness AI, and “Green” AI. She wrote large-audience books: *Les robots émotionnels* (Ed. L’Obs., 2020) and *Des Robots et des Hommes: mythes, fantasmes et réalité* (Ed. Plon, 2017). Her topics of research are human–machine coevolution: from the modeling of emotions and human–robot dialog to the ethical impacts on society and the risks and benefits of AI.

Ms. Devillers is a member of the National Comity Pilot on Ethics of Numeric (CNPEN). She is also the President of the Foundation Blaise Pascal on Cultural Mediation on Mathematics and Computer Science.

Roddy Cowie (Member, IEEE) is currently an Emeritus Professor of psychology with Queen’s University, Belfast, U.K. He used computational methods to study a range of complex perceptual phenomena—perceiving pictures, the experience of deafness, what speech conveys about the speaker, and, in a series of EC projects, the perception of emotion, where he has developed methods of measuring perceived emotion and inducing emotionally colored interactions. Key outputs include pioneering papers on emotion recognition in human–computer interaction (2001), the emotional states that are expressed by speech (2003), and ordinal representations of emotion (2018), as well as special editions of *Speech Communication* (2003) and *Neural Networks* (2005), and the *HUMAINE Handbook on Emotion-Oriented Systems* (2011).