AI & SOCIETY (2023) 38:973–974
https://doi.org/10.1007/s00146-022-01466-3

CURMUDGEON CORNER

Empathetic AI for ethics-in-the-small

Vivek Nallur[1] · **Graham Finlay**[1]

Received: 19 April 2021 / Accepted: 13 April 2022 / Published online: 19 May 2022

We contend that a more successful integration of AI with society would occur, not with AI deciding large questions of how society ought to be, but rather with empathetic interactions in the everyday tasks of individuals.

There has been much media attention on the society-wide effects of the rapid and unfettered deployment of AI (by AI, we mean any device/algorithm that uses AI techniques as part of its functioning, not necessarily just self-driving cars, robots, etc.). Most of these have focused on ethics-in-the-large, i.e., concepts like justice, fairness, bias—which can only be evaluated on a whole-society basis. We think that an equally important and immediate concern should be ethics-in-the-small, where technology has the potential to affect the quality of individual human lives. Consider, for example, mental-health apps that purport to offer support to individuals, or digital personal assistants that function as companions. Here, the notion of ethical behavior tends to be defined more by the individual's particular circumstances rather than general principles.

Instead of considering only large questions such as justice and bias while training AI, we should also be training AI to deal with daily human interaction. Nudges, such as those involving exercise/taking medicines/eating the right food for example, should have elements of empathy, as part of a general sensitivity to human emotions and goals. Nudging a depressed human to get some exercise may be well-meaning, but not ethically helpful or appropriate in this situation. This approach can include the insights of virtue and capability theory and help develop a notion of an AI-assisted understanding of the good life. None of the approaches employed for implementing ethical behavior in AI, surveyed in Nallur (2020), make *any* reference to affective states of human beings. This seems like a glaring oversight, especially while trying to create AI that is supposed to co-exist in society with human beings.

Unfortunately, AI-based systems have not attempted to model the human as an emotional being that may need empathy from time to time. Human beings tend to anthropomorphize pets, other animals, inanimate objects, and even abstract creations such as brands, teams, institutions, etc. This almost universal tendency to project humanness on clearly non-human entities indicates that the emotional projection is a primal need. This means that a well-designed AI-enabled system that is the object of anthropomorphization would be expected to understand the emotional projection by the human, reason about the human's mental landscape, and respond accordingly. This 'understanding response' is a clear indicator of the need for empathy in AI-enabled systems. Emotion recognition and understanding is an extremely nascent area, and there have been calls to ban emotion recognition in products, since the scientific foundations of such technologies are shaky. While we wholeheartedly agree with the need to be circumspect in the use of unvalidated technology, this points to the need for more research in understanding human emotions, not less. Without real emotion recognition, we are forced to rely on training AI solely on facial expressions, which could be severely misleading.

The urgency of this need for an 'understanding response' is evidenced by the burgeoning field of assisted living facilities, where robots co-exist with elderly patients on a long-term basis. Elderly patients are less able to communicate their needs or wants. In this scenario, it is imperative that the healthcare robot learns to anticipate not only the physical needs of the humans it is caring for, but also their emotional state of mind. Individuals' emotional states have been neglected by overly abstract theories of ethics. But there has been a new appreciation of the 'passions' and their role in moral life over recent decades that has made emotions at least complementary to rationality, in both ethical and prudential decision-making. Further, these emotions are not simply subjective experiences of individuals, but part of relationships between individuals and their wider community.

✉ Vivek Nallur
vivek.nallur@ucd.ie

Graham Finlay
graham.finlay@ucd.ie

[1] University College Dublin, Dublin, Ireland

Thus, feminist philosophers have emphasized the centrality of care relationships—which are more than simply responsibility for some person's welfare—and philosophers increasingly emphasize the importance of trust between individuals and within communities. Both care and trust relationships are central to the role of care robots and both require empathy to be created and maintained. Broadly speaking, empathy is the label we give to the processes involved in representing, understanding, and reacting to the internal, mental states of other human beings. While there is no consensus yet on whether the primary component of empathy is affective or cognitive, there is some evidence that both components exist. The affective component is commonly modeled using a simulation model, perhaps most easily described as "being in the others' shoes", where we intuitively experience what they feel, by simulating it in our minds. The cognitive component is modeled using a "theory of mind" approach that makes propositions about "the-other's" mental landscape. The metacognitive process of understanding that others have different beliefs and emotions and then performing some reasoning to infer their mental state is called mentalizing.

The potential for trust between an AI career and a human patient brings new meaning and urgency to the idea of 'trustworthy AI'. Because of their centrality to a distinctively human life, the capability theorist and theorist of the emotions, Martha Nussbaum, has identified relationships with others, including non-human animals as 'central capabilities' (Nussbaum 2013).Crucially, in Nussbaum's account of 'affiliation' empathy plays an important role. If AI careers or assistants can contribute to individuals enjoying this central capability, then they can greatly enhance individuals' well-being, but they *can only do so if they can uphold their part* of the relationship.

We believe that creating the capability to reason about—and demonstrate—affective concern is a neglected field of AI systems engineering and of AI ethics. Further research in this area contains considerable potential for greater understanding of human emotions and for concretely improving the lives of individual human beings.

References

Nallur V (2020) Landscape of machine implemented ethics. Sci Eng Ethics. https://doi.org/10.1007/s11948-020-00236-y

Nussbaum MC (2013) Creating capabilities: the human development approach. Harvard University Press, Cambridge