New Rationality and Value Alignment

The problem of value alignment is a central issue in the context of AI studies concerning the ethical dimension of the intelligent digital technology. The paper deals with the basic questions concerning the correspondence of the system of human values to what we would like or not like digital minds to be capable of. There is the suggestion that as humans cannot agree on a universal system of values in the positive sense, we might be able to agree on what has to be avoided. The paper argues that this might not be a reasonable path to follow. We still need to keep the positive approach in sight as well. It may be that there will never be a final solution of the value alignment problem. Rather we may be facing the era of endless adjustment of digital minds to the biological ones. The big aim is to keep humans in control of this adjustment. Philosophical analysis shows that the key concept in dealing with the issue of value alignment is value plurality. We cannot take neither a relativism nor the subjective view as the former means that the values are depending on the cultural tradition and the latter means that they depend on individual preferences. Value pluralism, however, does not necessarily contradict objectivity. We can accept pluralism but still take the position of objective ethics. This idea is discussed in the paper. The main idea of the paper, however, concerns rationality. The foundational idea here is that as in the human mind the rational and emotional sides are interconnected, the same should be achieved in the case of AI. So far, AI has been developed too much along the lines of traditional rationality where logical correctness has to be adhered to at any cost. However, as human mind is not fully rational in the traditional sense, the same approach could work in the case of AI. The latter could rather be developed along the lines of the non-traditional approach to rationality initiated by Nicholas Maxwell. The core of Maxwell's new rationality is the idea that in the case of taking decisions in the situation where there is no real problem in sight, it is more rational to adhere to spontaneous instincts rather than deliberating. There is seemingly no difference from the point of view of AI that can calculate with very high speed. However, in the case of ethical issues, the higher speed of decision taking may not count.