conception of the mind as a computer. There is nothing in the least nostalgic about this, no lingering desire to affirm the incomparable specialness of the human mind. The new way of thinking about the mind arises instead from advances in human knowledge and technology, which refute the idea of a homogeneous cognitive domain and help us see that we are not the only possible type of epistemic agent.

CHAPTER 3

Mind, Emotions, and Artificial Empathy

[T]he Other is given to me as a concrete and obvious presence that I can in no way derive from myself and that can in no way be placed in doubt or made the object of a phenomenological reduction or of any other *epoché*. . . .

JEAN-PAUL SARTRE

The heterogeneity of the cognitive domain, or, as some may prefer to call it, the diversity of mind, raises the question of whether there is something distinctive about the human mind compared with other types of natural and artificial cognitive systems, and if so, what exactly this peculiarity might be. Philosophy of mind and cognitive science give a contradictory response. On the one hand, they refuse to pose the question, affirming as a matter of principle that we do not have any special quality; we are only cognitive machines, no different from any others. On the other hand, as we have just seen, they hold that we are archetypal epistemic agents, on the ground that human beings uniquely furnish the criterion for deciding what is really cognitive. And yet no sooner has this advantage been conferred upon us than it is taken away, because it is doomed to disappear, they say, once we know how to construct fully conscious artificial agents. In many respects, these agents will also be far more efficient than we are, particularly with regard to memory and the ability to carry out high-speed calculations.

We do not claim to be able to give a complete answer to the question of what the distinctive character of the human mind consists in. Our aim is less ambitious. We wish to show, first, that the Cartesian

conception of the mind, which places the knowing subject at the heart of the cognitive process, requires appealing to a social dimension that Descartes himself explicitly rejected. We argue that the Cartesian conception—which instituted what may be called *methodological solipsism,* a principle that by and large has been inherited by modern philosophy of mind, psychology, and cognitive science—rests on a *social cognitive dynamic* whose importance, and even reality, it fails to recognize.

This neglected social aspect is precisely what social robotics sets out to explore. It is also what characterizes substitutes in their capacity as technological devices of a particular type. Later in this chapter, and then throughout Chapter 4, we go on to consider social robotics as *a creative form of inquiry.* This is the second part of our purpose. Social robotics may be said to be *creative* because it looks to build robots of a new kind that are capable of interacting socially with human partners; and this attempt is a *form of inquiry* because it constitutes an investigation into the nature of human sociability. Social robotics is not limited to integrating intelligent machines into an environment that is supposed to be perfectly known and controlled. The machines it designs, on account of both their successes and their failures at social interaction, are also instruments for learning about how human beings relate to one another. Here the basic assumption is that affect is central to what makes us social beings. Robots, in order to have social skills, must be capable of artificial empathy and emotions. From this it follows that affect constitutes a fundamental dimension of mind. But what is affect? This question, which is at the heart of social robotics, will guide us throughout the course of the present chapter and Chapter 4.

Where Is the Mind?

The extended mind thesis seeks to liberate the human mind from the prison of the brain and, in a sense, from anthropocentric prejudice

by arguing that the mind can be realized on material supports that are external to the agent. In this respect, it does not dissent from the functionalist thesis of multiple realizability—namely, that cognition is essentially a computational process that can be implemented by very different means: a brain, for example, or a computer.[1] Yet it does not renounce another prejudice, the subjective or individualist perspective, in part because, unlike classical computationalism, it derives from a particular conception of embodied mind in which the question "Where is the mind?" occupies an important place. It puts Otto and, by extension, all individuals at the center of the cognitive process, since they are imagined to constitute its mooring in physical space, and it considers the subjective experience of knowledge as the prototypical form of cognitive activity. Andy Clark could perfectly well have taken the position, of course, that it is when we abandon subjective conscious experience as the criterion of mental reality and accept that unconscious processes are likewise mental processes that the extended mind thesis really makes its power felt.[2] But in advancing a "parity principle" as the ground for determining whether an external process is part of the mind, Clark committed himself to a different view: "If, as we confront a task, a part of the world functions as a process which, were it to go on in the head, we would have no hesitation in accepting as part of the cognitive process, then that part of the world is (for that time) part of the cognitive process."[3]

The brain, or the head, remains the place where the mind resides and from which it extends and expands itself under certain conditions. The brain is the mind's port of registry, as it were, the harbor from which it never strays far enough to lose sight of the jetty. Why should what goes on in the brain enjoy this privilege, as the measure by which the cognitive (or noncognitive) character of processes that take place outside it is to be judged? What could be the reason for this primacy, if not that the cognitive experience of the intentional subject, translated in this case into the materialist language of the

brain, is taken to be, as with Rowlands's "mark of the cognitive," the ultimate arbiter of what is "really" cognitive?

The extended mind thesis is proposed as an answer to the question "Where is the mind?" Common sense generally replies "Inside me"—in my head, in my brain. The extended mind thesis adds: not always, for the mind sometimes extends beyond, seeps outside the skull and the skin that covers it. This does manage to qualify the spontaneous response of common sense, but only in a very limited way. An extended arm is able to distance the hand from the chest, for example, because the hand remains attached to the arm. The gesture by which the hand is distanced from the center of the body is possible, in other words, only because the hand continues to be part of the body. Similarly, if Otto's address book extends his mind by helping him to remember, it is because it permits *him,* Otto, to remember the address of the Museum of Modern Art. This is the necessary condition of the address book's being part of a genuinely cognitive process.

Although it is meant to seem paradoxical, in fact this argument does no more than restate the instinctive prejudice that experiencing the world from the first-person point of view reveals the fundamental structure of any cognitive process and therefore of mindfulness itself. Just as it does not recognize the existence of cognitive machines of another type than the human mind, the extended mind thesis, in agreement with the main part of cognitive psychology, presupposes that the subjective experience of knowing, of having one's mind "inside" oneself, constitutes the paradigm of all cognitive activity. Classical cognitivism, which regards the computer as the model of the mind, provides support for this attitude by offering a guarantee of scientific objectivity for what is merely a subjective prejudice.

The impression that the mind is inside of us, incontrovertibly obvious though it may seem in our intellectual culture, is in reality neither primary nor spontaneous. For *experiencing the world* from the first-person point of view is not equivalent to *experiencing the first-person point of view.* To experience the first-person point of view as a *point of view,* it is necessary to go beyond our direct perception of the world. It is necessary, in other words, to recognize that experience of the world is not the world itself. If it is true that the world is always immediately experienced *from* the first-person point of view, our immediate experience is not *of* the first-person point of view itself, but of being in the world, as one object among others, which presents us with opportunities and dangers in varying degrees. The "discovery" of the mind as something inside us, and, as a consequence, of the world also as being in a certain manner inside us, both representationally and intentionally, rather than our being immediately in the world, is therefore in no way primary or spontaneous. It is secondary and reflexive.

The Mind, Error, and the Other

Discovering that we perceive the world from a particular point of view requires that we have the experience that we are sometimes mistaken about the world and what it contains. Our own mind appears to us only when we discover that our apprehension of the world is not immune to error. A point of view is like a window through which we see the world. As long as the window is perfectly transparent, we do not notice that there is a window. Our cognitive failures, in making the window more or less opaque, give the mind substance by allowing us to appreciate the particularity and the limits of our point of view, which until then, absorbed by the immediate experience of being in the world, we were unable to see. This is a necessary condition—but not, as we will see, a sufficient condition—for our subjective apprehension of the world to interpose itself, in the form of a "mind," between ourselves and the world.

The link between error and the discovery of the mind nonetheless suggests why one finds a long philosophical tradition that, taking the

experience of being mistaken as a point of departure, associates the first-person point of view with the subject's privileged access to his own mind.[4] Already in Descartes, an illustrious representative of this tradition, one finds one of the fundamental arguments in favor of this privilege, which has been repeated in various forms up to the present day.[5] The thoughts that my mind contains represent the world, but these "representations" are not false in and of themselves. They can lead me astray only if I ascribe them to the world and if, through an act of judgment, I affirm that they correspond to what it contains, that they *represent* it as it is. In and of themselves, the contents of my consciousness, my mental states insofar as they are here, present in me, as immediate experience, as states of consciousness, cannot be false. I therefore possess a privileged epistemic access to myself, to my own mental states, that is free from error. This epistemic relationship to oneself is superior to any knowledge one may have of the external world, of which it constitutes the necessary condition.[6]

Since Descartes, this reflexive privilege has been accompanied by the view that knowledge of other minds can only be inferred. Whereas I have direct and infallible access to my own mind, I possess only indirect access to that of others. My knowledge of other minds is not only more uncertain than the knowledge I have of myself, it is inferior even to that which I can obtain about the external world. I can directly perceive a tree or a house, but I can only infer, on the basis of their behavior, the existence of other minds. I can never directly perceive the mind of another person, never directly observe his intentions or his emotions.[7] The mind remains concealed inside the fortress of the body, buried in a forest of behaviors, secure against any intrusion from outside—since it can be known only from within. It is therefore not surprising that the existence of other minds, the question of solipsism, should have long constituted, and should still constitute today, a fundamental philosophical problem.

The devaluation of the knowledge of the existence of others' minds, reduced to an oblique and contorted epistemic relationship, an indirect form of knowledge, has the fundamental consequence of shutting the subject up in his relationship to himself. It condemns the subject to seeing others only through the veil of his own mind, with which he alone is in direct contact, an internal theater that contains within it the whole world, but from which others are absent. One can only sense or suspect the presence of others. One can never meet them "in person," can do no more than imagine that these bodies that resemble one's own have a mental life "like mine." As a consequence, the mind is not a part of the world. The mind of another is at best only a theoretical entity; as for my mind, it is not an object in the world, but a stage on which the world can be seen. It is false to suppose that contemporary philosophy of mind and cognitive science have broken with either solipsism or idealism in this connection. Their efforts to embody the mind in the brain of an agent or in material mechanisms merely adopt and recast, as we have seen, this fundamental conceptual structure.

Artificial ethology offers a different image of the animal mind. The cognitive skills of an animal organism cannot be explained by its internal cognitive resources alone. They can be explained only if we take into account, as robotic modeling seeks to do, the complex relationship between its internal resources, its body, and the environment in which the animal acts. Such an approach is consistent with the radical embodiment thesis, according to which the cognitive abilities of a human organism emerge from the relationship between his nervous system and his body in interaction with the environment. The same thesis makes it possible to conceive of a diversity of types of cognitive system, whose characteristics will vary as a function of the particular features of both the terms of this relationship and the relationship itself. The mind may therefore be said to be

radically embodied: first, because it is inseparable from its particular embodiment, incapable of "escaping" it, of being embodied in a different form (as is supposed in the case of Otto's memory); second, because the mind in this view is not in the world in the way that an object is, but instead as a process. The growth of a plant or an animal, for instance, is not an object but a process, a web of events that take place in the world. In this connection, the term "embodiment" is liable to give rise to confusion, for in the Christian religious tradition it suggests the intrusion or incursion of a divine spirit into human affairs, whereas here the idea is that the mind emerges from purely physical arrangements.

Must we therefore reserve radical embodiment in this materialist sense for animal minds and hold, as Descartes did, that the human mind cannot be explained in so local a manner, through the mere "disposition" of the body's organs, because owing to its far greater universality it supersedes all such contingencies? Or must we say instead that the human mind's particular characteristics are due to its having emerged from a different environment than the one that produced, for example, lobsters and crickets? Descartes's own writings suggest a very different response to these questions than the one that is generally attributed to him and that philosophy of mind assiduously repeats while claiming all the while to reject dualism.

The Evil Demon

Descartes's discovery of the mind, and of the central place it occupies in the process of knowing, proceeded from what he called hyperbolic doubt, bearing upon the whole of the knowledge that until then he had taken to be true. It is important to recall that he considered hyperbolic doubt to be a means of achieving certitude in the sciences, not of discovering the origin of cognition in all its forms. If animals do not have a soul but yet are capable of cognitive behaviors, it follows that the human mind constitutes a particular type of cognitive system, which Descartes admittedly judged superior to all others, apart from divine understanding. His argument is therefore concerned with one type of cognition only, human knowledge, and its conclusions need not apply to other types, still less to all of them.

While recognizing that one is sometimes mistaken is a necessary condition of becoming aware of one's point of view on the world as a particular point of view, it is not sufficient for this point of view to be transformed into a "mind"—that is, into a representation of the world that interposes itself between oneself and the world. This is because, in principle, cognitive errors can be detected and rectified. So far as this can be done, at least most of the time, one's point of view is scarcely apt to acquire the reality, the "ontological weight" needed for it to become something to wonder about.

Thus Aristotle sees a tower that from far away seems to him round. On coming nearer, however, he realizes that it is in fact square.[8] He therefore concludes that the apparent shape of the tower depends on its position in the world and that in changing his own position in relation to it he can discover its true shape. The difficulty facing Descartes in his search for absolute certainty is of another kind. Considering this example, he would conclude instead that he is liable to be mistaken about the shape of the tower at any moment, wherever he may stand in relation to it. Unlike Aristotle, Descartes does not seek to know whether he is mistaken at a specific time, but under what conditions it is possible to be mistaken at all, and what he must do in order *to be able to avoid error at any time.* The problem, as Descartes formulates it, is this: how can one guard against hyperbolic doubt, against the mad suspicion that perhaps there is no tower at all, no world within which such things as towers exist? The philosophical discovery of the self as mind requires exactly this radical sort of uncertainty. But the philosophical fable that leads to this discovery is much more revealing, and much less misleading, than is generally

supposed—at least if one is prepared to regard it as a datum, a symptom needing itself to be analyzed, rather than as a true account.

According to Descartes, "I think, therefore I am" offers a certitude that no other thought, belief, or assertion can claim and establishes the priority of the epistemic relationship of a person to himself, to his mind, by comparison with any epistemic relationship to the external world. Now, if "I think, therefore I am" offers greater certitude than, for example, "I scratch my nose, therefore I am," it is because it involves a very particular kind of thought: doubt—the sort of doubt that bears upon everything one thinks one knows. The epistemic priority of the mind is the reverse, the flip side of the claim that any other knowledge is more uncertain and imperfect. It asserts that even when an individual doubts everything, he remains assured of his own existence.

Yet doubting everything is not an easy thing to do. Doubting elementary truths of mathematics, fearing that you are mistaken when you perform a very simple operation—counting the sides of a triangle, for example—is not obvious at all, especially when you wish to convince your readers that you are not so deranged, as Descartes puts it in the *Meditations on First Philosophy* (1641 / 1647), as "those mad people whose brains are so impaired by the strong vapour of black bile that they confidently claim to be kings when they are paupers, that they are dressed up in purple when they are naked, that they have an earthenware head, or that they are a totally hollowed-out shell or are made of glass."[9] To accomplish such a feat, Descartes finds himself obliged to introduce a second character in his fable, usually called the "evil demon" (sometimes "evil genius"). The fact that this demon is an imaginary creature must not distract us from the essential point, that it is a logical necessity.

Let us look once again at the example of the tower and cast it in the form of hyperbolic doubt. As Aristotle approaches the tower, a subtle holographic projection makes it appear square to him, when

in fact it is round, as it first appeared to him from a distance and contrary to how it appears to him now that he draws nearer to it. In other words, whereas Aristotle believed that he had corrected his initial error, hyperbolic doubt invites us to suppose that he cannot tell the difference between being mistaken and being no longer mistaken. Consequently, no matter what Aristotle does, whether he stretches out his arm to touch it, whether he walks around it, whether he studies its shadow or observes its reflection in a nearby pond, the tower always appears square to him—when in fact it is round! An evil demon—a cunning computer engineer, as we might think of it today, or a wily film director—does everything in his power to deceive him. Now, what can it really mean in this case to say that the tower "is in fact round" or that someone "is mistaken" when he counts the number of sides of a triangle, when the subject is assumed never to have had the experience of being wrong? Aristotle, by hypothesis, can never conclude that the tower is round since this fact is hidden from him by the demon's artifices; and Descartes, for obvious reasons, will never manage to convince himself that a triangle may have more (or less) than three sides. This situation is utterly different from the one in which Aristotle, approaching the tower, discovers that it is in fact square, whereas from a distance it appeared to him to be round.

What can "be mistaken" mean here, in a situation where one never has, and indeed never can have, the experience of being mistaken because it has been ruled out in advance? What can "be mistaken" mean in a situation where one cannot even conceive, but can only *imagine* the possibility of error? According to the Cartesian fable, it can mean only one thing, namely, that "round" is the way in which the tower appears to another epistemic agent, and that "more or less than three" is the number of sides of a triangle perceived by another, more powerful epistemic agent who may or may not attempt to fool Aristotle or Descartes, to prevent him from discovering how things truly are. The essential thing is not whether this agent seeks to prevent

the subject from sharing his point of view, but that subjective error can only be conceived or imagined in relation to another agent having a different point of view. This agent alone makes possible the hyperbolic doubt on which the certainty of one's own existence depends, the assumed epistemic privilege of first-person knowledge of the mind. Hyperbolic doubt requires the existence of another cognitive agent who can take me as an object of knowledge and declare that I am mistaken.

Descartes's hyperbolic doubt and the metaphysical discovery of the mind to which it gives rise require the presence of someone who is able to judge that the subject, Descartes, is mistaken. What permits, though evidently it does not justify, the claim that knowledge of oneself and of one's own mind enjoys epistemic priority over one's knowledge of the world, particularly knowledge of others, is the active presence of a second epistemic agent who exercises his cognitive abilities in the same world in which Descartes exercises his own. For the argument to be successful, for Descartes to be convinced of the certainty of his own existence, Descartes himself must be taken as an object, must be operated on, so to speak, by another epistemic agent who will have "devoted all his energies" to deceiving him. The epistemic priority of first-person knowledge is an illusion, for both the discovery of the mind and the impression that it occupies the center of the cognitive process are made possible by, and flow from, the presence and the action of another epistemic agent who acts upon the subject.

The environment from which the characteristics of the human mind emerge—characteristics that, since Descartes, have been supposed to define its superiority by comparison with other cognitive systems—is a social environment. The cardinal virtue of the Cartesian metaphysical fable is that it implies just this, that *the subjective structure of first-person experience of the mind does not reflect the structure of the cognitive process from which it emerges.* To be sure, Descartes himself

drew an altogether different conclusion. Ever since the publication of the *Meditations,* the epistemic priority of knowledge of oneself and the mind has generally been accepted as something obvious. A close reading of what Descartes actually says makes it clear, however, that the discovery of the mind is the result of being fooled—of being fooled, or otherwise led into error, by another epistemic agent. The social environment inhabited by the Cartesian subject is therefore not emotionally neutral. It is shot through with affect, riddled with the anxiety of a subject who imagines not only that he has been led into error by another, but that the error is one that he is not even capable of conceiving.

Methodological solipsism and the priority of self-knowledge over knowledge of others therefore have no sound philosophical basis. There is no reason to take at face value the spontaneous and commonsense answer to the question "Where is the mind?"—neither in its usual version, nor its metaphysical version, nor in the vague and underdetermined version proposed by the extended mind thesis. The mind is neither in the brain, nor in the head, nor outside the agent (much less in Otto's address book), *but in the relations that obtain between epistemic agents.* This is why the question "Where is the mind?"—just like the question "How far does the mind extend?"—scarcely makes any sense at all.

Emotive and Empathic Robots

The branch of robotics that explores the social environment from which mind emerges, a domain largely neglected by philosophy of mind and cognitive science, assigns a central place to the study of affect. Compared with other disciplines, the mixed empirical / theoretical style of research practiced by social robotics gives it a distinctive flavor. On the one hand, it seeks to devise solutions to engineering problems, so that robots will have this or that functional property or