This is an unedited version of Chapter 20 of Dreyfus, H. & Wrathall, M., eds.,
The Blackwell Companion to Phenomenology and Existentialism,
London: Blackwell, 2006: pp. 377-393

Phenomenology in artificial intelligence and cognitive science

Fifty years before the present volume appeared, artificial intelligence (AI) and cognitive science (Cogsci) emerged from a couple of small-scale academic encounters on the East Coast of the United States. Wedded together like Siamese twins, these nascent research programs appeared to rest on some general assumptions regarding the human mind, and closely connected methodological principles, which set them at such a distance from phenomenology that no contact between the two approaches seemed conceivable. Soon however contact was made, in the form of a head-on critique of the AI/Cogsci project mostly inspired by arguments from phenomenology. For a while, it seemed like nothing would come of it: AI/Cogsci bloomed while the small troop of critical phenomenologists kept objecting. Then AI and Cogsci went their separate ways. AI underwent a deep transformation and all but surrendered to the phenomenological critique. Cogsci meanwhile pursued the initial program with a far richer collection of problems, concepts and methods, and was for a long time quite unconcerned by suggestions and objections from phenomenology. The last decade and half has seen a remarkable reversal: on the one hand, a few cognitive scientists have been actively pursuing the goal of reconciliating Cogsci, whether empirically or foundationally, with some of the insights procured by phenomenology; on the other, many cognitive scientists and philosophers of mind who think of themselves as, respectively, mainstream and analytic, and have no or little acquaintance with, and often little sympathy for, phenomenology, have been actively pursuing research programs geared toward some of the key issues identified by phenomenological critics of early AI/Cogsci. It might seem then as if those critics were now vindicated. But while these new directions are undoubtedly promising, it is not yet clear that phenomenology and Cogsci can be truly reconciled. Some suspect that Cogsci must distort beyond recognition those phenomenological themes it means to weave into its fabric, while phenomenology may be losing touch with its roots by tuning onto the logic of Cogsci which is, after all, an empirical science. To the present writer, it is far too early for anything like a verdict, as the task of clarifying the issues and gaining a much deeper understanding of the issues on both sides has barely begun. But whatever emerges from this exploration will probably have deep consequences for both Cogsci and philosophy.

1. A phenomenologically inspired critique of early AI

In a book which appeared in 1972 and went through two further, augmented, editions (Dreyfus 1972/1979/1992), Hubert Dreyfus provided a massive set of interrelated arguments against AI. This seminal work, together with subsequent writings by Dreyfus and his followers, provides the backdrop to a large part of the discussion which has been developing since then. Dreyfus attempted to show that (i) contrary to heretofore unchallenged claims by its proponents, AI was not making significant progress towards creating intelligent artefacts; (ii) to a large extent, this lack of success was due to erroneous assumptions about the mind; (iii) AI was ignoring dimensions of the mind which are critical to intelligent, adaptive behavior.

Before the most central and enduring ideas are summarized, three remarks are in order. First, Dreyfus' approach is not aprioristic: he fights AI on its own turf, like a naturalistic philosopher of science who conceives of science and philosophy as continuous. The naturalistic spirit of his approach is also made manifest in his rejection of any form, overt or covert, of dualism: not for him, stirring appeals to consciousness, emotions, freedom, norms and values, or the Subject; not a thought directed against science or technology; no attempt to downplay the causal powers of the brain; no reliance on interpretive or other non-realistic views of mind or psychology. This straighforward naturalistic stance explains the otherwise unexpected efficacy of the Dreyfusian line of thought in penetrating the culturally hostile world of AI and Cogsci. It also shows the profound inadequacy of the analytic/continental distinction in accounting for the commerce of ideas in the case at hand: in fact, themes and theses originating in continental traditions are increasingly deployed and refined most ably by analytically-trained philosophers. What may be called Dreyfus' methodological naturalism, on the one hand, and the analytic co-opting of (aspects of) continental philosophy, on the other, raise interesting philosophical and meta-philosophical issues which cannot be pursued here.

Second, Dreyfus aims at a moving target, and a target which he himself helps to displace for the reasons just mentioned. The several components of his attack on AI form a cohesive 'long argument' (somewhat in the way Darwin viewed On the Origin of Species) in the context of early AI. The general structure of the argument is something like this: AI aims at explaining human intelligence by building intelligent machines, and to this end it takes on board a set of hypotheses (H) regarding the essential nature of the mind. But the assumptions in (H) are mistaken, and further, the mind exhibits a set of properties (P) which seem to account for, or be constitutive of, intelligence, and which AI ignores. The falsity of (H) and the reality and importance of (P) result in part from independent considerations, in part from an inference to the best explanation (of AI's persisting difficulties and of the patterns they follow). Unlike Darwin's object of study however, Dreyfus' has been rapidly changing: AI has developed new models, based on different assumptions, and it now sees it as an obligation to account for at least

some of the properties in (P); unsurprisingly, the clinical tableau presented by this new AI (actually, by the variety of new paradigms in AI) is different from that of early AI (also known as GOFAI: Good Old-Fashioned Artificial Intelligence, cp. Haugeland 1985). Thus it would seem that Dreyfus's long argument would need at least a thorough revision, or perhaps be archived among the minutes of successful pleas against reformed or extinct culprits. Sorting out those parts of the argument which can be safely shelved, those which need adjustment, and those which remain essentially valid as is, has in fact been a task for Dreyfus and for those who were convinced by the argument or at least took it seriously.

Third, however closely connected, the considerations gathered here under (ii) –the mistaken assumptions (H)– and (iii) –the ignored dimensions (P)– were distinct and met with unequal degrees of approval. Dreyfus' rejection of a view of the mind as an information-processor is at the heart of (ii), and to this day is deemed inconclusive by many thinkers, including some who are fully sympathetic to part (iii). The latter, on the other hand, revolves around such issues as the role of commensense, embodiment, engagement, context, which now figure among the core issues of the field. A crucial question therefore is the extent to which one can honor (iii) without siding with (ii).

Let us now briefly review the main tenets of Dreyfus' analysis. Regarding the difficulties which beset GOFAI ('Promethean AI' in Jerry Fodor's phrase), little need to be said here: this variety of AI is essentially defunct (according to one of its founding fathers, Marvin Minsky, it has been 'brain dead' since the early 70s –by which he presumably means that although apparently alive and well, it had suffered internal theoretical injury from which it was not to recover).

By contrast, some of the basic assumptions underlying the defunct research program are shared by many working scientists and philosophers today; thus Dreyfus's objections retain most of their relevance. Intelligence (insufficiently distinguished, at the time, from 'mindedness', the property of having a mind or exhibiting the essential functions of the human mind) was hypothesized to be a property of information-processing systems such as suitably programmed computers. This implied that (i) mental processes operate on a uniform basis of discrete units of information; (ii) the units are carried by material vehicles whose causal powers are in principle independent from the entities about which they carry information; (iii) thinking results from, or rather, is nothing but, the performance of computations on symbolic 'representations' built up from elementary bits of context-independent information; (iv) the ability of a system to produce true, rational, or adaptive thinking is due to its possessing the requisite facts and a truth-preserving inference engine, *i.e.* a computational routine for drawing logical conclusions from the stored facts. An intelligent system placed in an environment is made up of three compartments or modules, one for the perceptive intake of transient information; another for the computation of the appropriate inferences about the state of the world and the required action; the third, for the motor control of the actions to be effected. The central module is thus

insulated from the world (a property sometimes referred to as the formality condition, or, indirectly, as methodological solipsism): it 'communicates' only via informational transducers, somewhat like a military command center sunk deep in the ground so as to escape any (directly) physical contact. It operates on a set of explicit propositions which together form a 'data base', a theory which serves as 'model' of, or 'represents', the (appropriate parts of the) world, and is limited to applying explicit formal rules to its 'data base'.

Dreyfus objected to this conception of the intelligent mind on a variety of grounds. He faulted it for being unmotivated; incoherent; wildly implausible from the most elementary phenomenological standpoint; vulnerable to collapse in real-world situations. At the same time, he claimed that AI was nothing but a zealous follower of the rationalistic tradition in Western philosophy. Uncovering its fatal flaws was therefore not as simple a task as might first appear, and however elementary, the phenomenogical stance deployed against AI's theoretical framework was rooted in the deepest sources of phenomenology in the historical sense.

Out of this vast set of considerations, three will be briefly explained. The first concerns the positing of context-free units of meaning, or primitives. Dreyfus co-invented with his colleague John Searle the amusing and now popular game of finding examples showing that even the simplest sentences, the most obvious pieces of behavior, could have radically different meanings, and call for radically different responses, according to the context in which they appear. The project of reducing this context-sensitivity to a mere matter of differences in collateral information (itself assumed to be context-independent) is, according to Dreyfus, doomed to fail, and the claim that somehow it must succed, if we are to provide a naturalistic account of intelligence, mere question-begging. Heidegger, says Dreyfus, was the first to pinpoint the fallacy of what he called the 'metaphysical assumption', which consists in viewing the background as simply extra information which merely needs to be made explicit.

Another problem Dreyfus raised concerns the idea of intelligent behavior as resulting from the application of formal rules to the information at hand. One objection has since then become familiar from discussions about rule-following initiated by Wittgenstein's skeptic argument: what rule must one deploy to determine which rule, or in which case a given rule, should be applied? (The regress had already been noticed by Lewis Carroll). But Dreyfus also undermined the reasons given by the rationalistic tradition up to and including AI, for thinking that rational behavior *must* result from rule application. First, he argued that from the fact that the trajectories of cognitive systems, like those of all physical systems exhibiting some regularity, are rule-*governed*, it is fallacious to infer that these trajectories causally follow from the system's (conscious or unconscious) *observance*, in the psychological or information-processing sense, of some rules. When I ride my bicycle, my trajectory 'obeys' a complicated systems of differential equations, but there is no reason to suppose that I am 'unconsciously' computing the solutions and using them to apply to the handlebar the angle appropriate to my negotiating the turn without falling. Second, H. and S. Dreyfus developed a model of skill

acquisition which, if accurate, undermines the argument from learning. Many skills are in fact taught initially by inculcating a set of context-free rules; isn't it obvious then that, as the learner's proficiency improves, she is 'automating' those rules and still applying them, albeit now 'mindlessly', whenever a particular move or gesture is requires? Isn't this in fact the paradigm of 'knowledge acquisition', and hence the very basis of the process by which a (presumably) essentially unintelligent newborn becomes in due course an intelligent adult? Now on the Dreyfus' model, based on a combination of phenomenological observations and results from experimental psychology, the rules given the beginner are not pushed down under the threshold of consciousness, ready to be unconsciously activated when needed; rather, they are discarded like training-wheels on toddlers' bicycles. Fluid, expert performance rests on entirely different principles, those of 'skillful coping'.

The third target which Dreyfus aims for is the assumption which AI was led to make explicit regarding commonsense, understood as what enables any normal human being to negotiate familiar and novel situations effortlessly, with rare and usually benign mistakes, and that even sophisticated and lightning-fast computers seem to lack, leading them to catastrophic breakdowns and wildly improper behavior. AI's assumption is that this commonsense is made up of a gigantic mass of propositional knowledge about the various realms of common activity, ranging from the taxonomy and behavior of middle-sized physical objects and substances to the basics of human interactions. In order for a computer to become truly intelligent, in the human sense of the word (and in particular useful outside its usual range of applications), it is obvious, according to AI, that it needs to have access to all the banal facts which a human being knows about the way the ordinary world works in ordinary circumstances. Dreyfus sees two seemingly unsurmountable, and to this day unsurmounted, problems with this proposal. The first is that there seems no end in sight for the task of collecting the 'facts' which together purportedly make up commonsense knowledge. The conjecture that with about 10 million items a computer should at last achieve commonsense, floated today by some die-hard 'factualists', is a wild guess. But even supposing that these millions of items were actually at hand, and conveniently stored in a data base, the second and even harder problem is that of relevance: how does a rational, rule-governed computer retrieve, among its millions of facts, the few which are required to solve the problem at hand? One answer which has been proposed in various guises and at various moments is to divide up the big bundle in smaller, more manageable bundles. But this won't do for two reasons at least. One is a matter of simple arithmetic: the bundles are individually more manageable only if they are smaller than the whole thing by several orders of magnitude, in which case the system will run into the problem of discovering in less than astronomical time which is the right bundle to exploit. The other reason is that human situations don't invariably involve just one domain. Interferences happen all the time, and intelligence at a quite ordinary level requires, indeed in large part consists in, dealing at least adequately, in all but perhaps the most extreme circumstances, with interferences. Flirting behavior in a restaurant ceases when

the courted one chokes on a fishbone, or when a racist remark is made by a customer sitting at a neighboring table; etc., to any degree of embedding. Life, not just knowledge or belief fixation or the interpretation of utterances, is 'Quinean' in Fodor's sense: at any moment, anything might turn out to be relevant in any situation. (One form of the relevance problem has gained fame in the AI literature under the label of the *frame problem).*

But then, *how* do humans achieve intelligence? This is the question which the constructive part of Dreyfus' program attempts to answer, by bringing out the features or dimensions of the mind which AI has been oblivious to, and which may precisely hold the key to intelligence. Dreyfus proposes an account of intelligent behavior which sits somewhere between description and explanation. According to the perspective one adopts, one will tend to think of what follows as part of the *explananda* or part of the *explanantia.* The explanatory function of the account consists in redescribing the phenomena so as to make apparent, first, that, once solved, a large part of the mystery of intellligence dissolves, and second, that an explanation of these phenomena will in all likelihood call on principles radically different from those propounded by AI. But although, as will shortly be seen, Dreyfus has exploited his own suggestion by proposing that connectionist networks go some way towards providing the desired solution, thus showing how the initial probem –the elucidation of the material basis of human intelligence– might be solved, the most enduring contribution he makes in this part of his work consists in drawing attention to what he regards as *fundamental* abilities of the human mind and which have been either completely ignored by the classical rationalist tradition, or categorized as sophisticated and derived from more basic powers.

'Abilities' is not quite the right word for what Dreyfus points us to. Rather, he invites us to think of the mind not as a set of abstract functions of an autonomous organ, the brain, but rather as a complex of emerging properties of something considerably more inclusive. This enlargement goes through three phases. First, what is traditionally attributed to the solitary, cogitative or computational mind, Dreyfus, drawing on Merleau-Ponty, assigns to the entire body. The body is no puppet manipulated by the brain; rather, the body relies on its various organs, with their characteristic shapes and subject to their proprietary constraints (knees don't bend backwards, arms are less than a mile long, eyes don't pop out of their sockets to check what's behind our backs, etc.) to generate appropriate responses to the situation at hand. This 'skillful coping', however elaborate and complex it would appear as a performance of a computational-representational device, is for humans (and presumably for other animal species) a basic, primitive ability, regardless of the way it becomes part of their repertoire. In brief, the mind is *embodied.*

The second stage of the enlargement brings in the physical environment. Combining insights from Merleau-Ponty and the American psychologist J.J. Gibson (1904-1979), Dreyfus suggests that objects and relations in visual space are not identified from a neutral perspective on the basis of their computationally salient features. Instead, they are grasped as 'affordances'

(Gibson), as 'in order to's' (Heidegger), they are holistically perceived as potentials for action. The mind is not initially disconnected from the physical environment; it leans on it from the very beginning and is engaged in an uninterrupted interaction, somewhat in the way a fish moves about in the water ('*immersion* in daily activity' is a phrase which comes up frequently in this discussion). Physical space is not a homogenous repository of objects, it is a structured realm of possible trajectories, tools and doings, and it is in this strong sense that the mind must be regarded as *physically embedded or situated.*

In the last stage, the intrinsically social nature of the mind is brought to light. Human activity is responsive to social practices and to the particular individuals we are, directly or indirectly, connected with. Foremost among social practices is language, but the most ordinary tools and pieces of equipment are permeated with socially determined uses and purposes: as Heidegger stresses, equipment invariably is equipment-for. Paths are both to-be-treaded-on-by-me and historical traces of social activity projecting backward and forward. The mind is now seen as *socially and culturally embedded or situated.*

So finally, according to Dreyfus, we are beginning to comprehend the full implications of Merleau-Ponty's somewhat cryptic claim that "[t]he life of consciousness –cognitive life, the life of desire or perceptual life– is subtended by an 'intentional arc' which projects round about us our past, our future, our human setting, our physical, ideological and moral situation." (Merleau-Ponty 1962: 136; quoted by Dreyfus in Wrathall and Kelly 1996).

But *how* does the embodied, embedded and multiply situated mind actually accomplish the remarkable feats it is credited with? This is not a question to which Dreyfus or any of his followers claim to have an answer: it is the job of science to uncover the material basis of these capacities. However, there is a proposal which goes some way towards bridging the gap. Intelligent behavior, the proper comportment in a given situation, may result from an ability to match the situation to one sufficiently and relevantly similar among a stored repertory of previously encountered situations. Much more needs to be said about this, but space permits no more than to stress the priority given in this conjecture to active perception and pattern matching. Provided this intermediary level of description is phenomenologically and conceptually secured, the question then arises of how to connect it with a causally more basic level. One possibility is to move directly to the neural level and search for the processes in the brain responsible for the abilities in question. Another is to construct intermediate models, systems which are in turn realizable in the neural tissue: such is strategy of (a certain brand of) connectionism, as we will see presently.

In the end, the traditional intellectualist view of the mind, culminating in classical AI, is seen to be, as Merleau-Ponty says, not so much utterly false as abstract. In fact, the claim is that it gets things exactly backwards: first comes the engaged body, tuned to its environment through constant adjustments involving perception-action arcs, skillfully coping with situations in a world which it inhabits and shares with others; second come a series of disengaging

procedures which gradually make space for a detached intellect reflecting on context-independent facts in order to discover, by deliberate and conscious search, a solution to a given problem and to implement this solution through an appropriate sequence of actions. Far from being the more basic mode, reflective problem-solving is an advanced elaboration, requiring the deployment of sophisticated cognitive tools and techniques such as a logically-reformed language use, record-keeping, writing, calculating, etc. But further, the traditional account of such cogitative processes is no more than a rational reconstruction, a 'model' of the competence deployed, not a phenomenologically true description of the performance, nor a scientific-psychological account of it. In the accomplishment of abstract intellectual tasks, our skills for coping are brought to bear, whether in the analytical set-up of the problem to be solved, in the application of the rules most likely to uncover candidate solutions, or in the final decision to select and apply one of them: it takes a distinctive know-how to put knowings-that to good use (or to any use at all, for that matter).

2. A methodological interlude. Threesomes

It may seem at this juncture that the issues are fairly clearly delineated, and that it only remains for AI/Cogsci to draw the lesson by making the required adjustments in its assumptions and the scope of its empirical investigations (or else show that the phenomenological accounts are, wholly or partly, mistaken). In fact, as hinted in the introduction, this is not at all the way things are going. The problem situation is vastly more complicated. The main reason is that what, in the context described above, appeared as forming a unity, has come apart and turned out to be a plurality.

Dreyfus correctly saw in GOFAI a technological venture aiming for intelligent computing systems, as well as a research programme within scientific psychology, and to boot a set of foundational assumptions amounting to a doctrine in the philosophy of mind. But while even then AI was only notionally or programmatically, in the eyes of its more ambitious and articulate proponents, technology, science and philosophy rolled into one, it gradually transpired, in the 1970s and 1980s, that there were in fact three distinct areas with admittedly active exchanges between them as well as border regions. This is perhaps not always clearly perceived outside the concerned areas, due in part to a variable and misleading terminology, in part to an objective historical evolution.

Terminology first. 'Cognitive science' took hold gradually, beginning in the mid 1970s, and stabilized to its present, still somewhat uncertain, definition, only a decade later. Initially, it was barely distinguishable from cognitive psychology, itself a very recent creation (1967). Cognitive psychology was understood then as the new, computational-informational approach in the psychology of cognitive processes, themselves construed in a restricted sense as those

which subtend the formation and treatment of knowledge (and more generally, belief). Cognitive psychology was thus more than just a part of psychology (on par with such branches of scientific psychology as clinical or social or differential psycholgy): it was a research program, or, in Kuhn's sense, a 'paradigm' or a 'disciplinary matrix' within psychology, and the Siamese twin of (early) AI. Actually, by the time the locution had been coined, cognitive psychology had begun to separate from AI, and re-identified some of its roots in previous traditions within psychology (from Vygotsky to Piaget, and including in fact parts of behaviorism). Similarly, cognitive science in the beginning was restricted to a (slightly broader) paradigm, what might be called an 'interdisciplinary matrix': it referred to the study of mental processes conducted in the framework of the computational-representational theory of mind, also known, thanks to John Haugeland, as 'cognitivism'. Although psychology occupied the center of this new program, it involved to an important extent linguistics, AI, the brain sciences (not yet dubbed 'neuroscience'), philosophy, and some tidbits from the social sciences. For the purposes of the present chapter, let us call this program Cogsci-1, so as to clearly distinguish it from the construal most current today, labelled here Cogsci-2, which is like Cogsci-1 but with no commitment to cognitivism. Whether it is conceptually unproblematic to proceed in this way (appealing implicitly to the examples of, say, physics, biology or geology, which don't come, on the face of it, with strings attached to any particular 'physicalism', 'biologism' or 'geologism') is a genuine issue which cannot be discussed here. *De facto*, many practitioners of Cogsci-2 regard cognitivism as no more than a school, a set of assumptions to be accepted or rejected piecemeal *ad libitum*, or again as a cohesive paradigm within their discipline, but in no way a condition of its existence.

Historical changes now. AI has undergone a profound overhaul. First, it has shed its 'Promethean' ambition, or rather, it has ceased making it an official goal. The new frontier within AI now concerns 'intelligent' cognitive prostheses, aids for the human agent, and falls under the wider label of applied cognitive science. Part of the old AI is indistinguishable from software engineering; the remainder is divided between applied logic and natural language processing and is a province of Cogsci. The Promethean inspiration has moved to Artificial Life and more recently to Artificial Consciousness, with very limited effect on Cogsci. Finally, AI has taken on board connectionism (see Rumelhart *et al*, 1986, Smolensky & Legendre 2005) and tries to federate all the new modelling techniques useful in Cogsci; unfortunately for the new AI, the most powerful methods are wielded by physicists, who tend to deal directly with the neuroscientists, with whom they share the intuition that modelling the brain is a safer bet than simulating the mind.

Cogsci-2 is opportunistic, as mature sciences are wont to be. It has broadened its scope to include not only just about every respectable topic in scientific psychology, including 'hot' cognition (emotions, motivation, etc,), consciousness, animal cognition and the origins of mind, but also entirely novel themes.

Finally, philosophy of mind has expanded enormously; it follows Cogsci opportunistically, but also pursues an agenda of its own, which makes room for every conceivable ontological option, including dualism. There is no trace left of its former role of handmaiden to cognitivism.

Thus, instead of one doctrine uniting the efforts of philosophers and scientists (somewhat like the mechanical philosophy around Descartes' and Galileo's time), the camp which was challenged by Dreyfus and his followers from the mid 1960s to the late 1980s has split up in three disciplines, each of which has diversified and expanded beyond recognition. Inspiration, penetration or critique by phenomenology will thus take on different forms according to the particular disciplinary or doctrinal province of the cognitive 'galaxy' one is aiming for.

On the side of phenomenology, in contrast with the Dreyfus line strongly moored in the works and intentions of the original movement, there are at present, besides extensions and variations of this authentic inspiration, two diluted varieties at work in the field of cognition. Phenomenology-1 is no more than the consideration of consciousness, qualia and the first-person reports of introspection, the structure of intentional states, etc.; it demands no connection at all with philosophical phenomenology; in fact, one can claim to be in that sense a phenomenologically-inclined philosopher of cognition or cognitive scentist without having read a line of Husserl, Heidegger or Merleau-Ponty. Phenomenology-2 consists in attempts to domesticate typically phenomenological themes in the cognitive culture; for example, embodiment, or concern, or shared intention, or equipment, with some of their attendent properties, might be designated as requiring sustained attention, without the need being felt to attend to these phenomena in the style and with the tools of phenomenology. Phenomenology-3, finally, is the approach illustrated originally by Dreyfus, and pursued with increasing intensity by the current which he created and by some more recently formed schools.

To a first approximation, the three shades of phenomenological interventions correspond to three kinds of effects on Cogsci. Phenomenology-1 is a *heuristic* device for Cogsci: it merely suggests the inclusion of new phenomena to the agenda. Phenomenology-2 brings in *constraints*: insofar as it suggests not only some phenomenon, but a requirement that its central aspects be taken into account, a phenomenological-2 contribution will typically impose on the cognitive-scientific accounts involving the phenomenon the obligation to take care of its phenomenological properties. Finally, the sense of a phenomenological-3 intervention is to propose, or impose, ontological or metaphysical options. Admittedly, the boundaries are fuzzy and permeable; still, there are clear-cut differences between central cases. But the massive influx of phenomenology-1 and -2 may be in the process of inducing in Cogsci changes of a magnitude such that the effects are no less than what one would expect from a successful phenomenological-3 therapy. Perhaps, if this turns out to be the case, one should conclude that cognitive scientists will have rediscovered on their own some of the guiding intuitions of the

great phenomenologists. On the other hand, such an outcome raises the question of the internal consistency of Cogsci thus revised: will it not require serious readjustments, the reexamination of some of its results and the abandonment of some of its classical tenets? This would be reminiscent of what physics underwent during the 19th century, when it had finally to shed the remainders of the mechanistic philosophy which had kept developing during the 18th century side by side with Newtonianism. But this 'new Cogsci' will not likely resemble anything like a pure natural science, and may be something like a morphed interpolation of brain science and phenomenology, within which what we now still think of as scientific psychology will have undergone a complete transformation. This remains however quite speculative, and the remainder of the chapter will be devoted to ongoing, documented developments.

3. A sample of phenomenologically-inspired interventions in cognitive science

Phenomenology-1

A list of themes which have recently made it to the top of Cogsci's agenda would include consciousness, emotions, culture and distributed cognition, and social cognition. On all of these, save occasionally the first, there is no visible trace of an influence from phenomenological writings, vocabulary or style of inquiry.

Consciousness in this context is approached mostly in functional or operational terms, with heavy emphasis on neuroscientific studies. There are many 'theories' and 'models' of consciousness on the market, few if any of which include anything like a careful phenomenological examination of the phenomena. On the other hand, there is an abundant literature on 'phenomenal' consciousness, the 'explanatory gap' which seems to separate it from any conceivable scientific account, and the 'hard problem' this raises. [CROSS REF TO APPROPRIATE CHAPTERS].

Emotions has also been an explosive topic, with contributions from conceptually- as well as empirically-minded philosophers, evolutionary biologists, anthropologists and neuroscientists. The important insight procured is that emotions should not be seen as a set of phenomena separate from cognitive processes or faculties, but an integral part of the mind, so that even the seemingly plausible divide between 'cold' and 'hot' cognition should be abandoned.

Individual cognitive processes are increasingly seen as highly dependent, or even derivative, on populational phenomena. Cognition is thought to be in part, or for some authors in essence, a distributed process, involving entire populations of individual minds (or perhaps one should say people, or organisms) and things, whether culturally-enrolled natural objects and processes or artefacts. The radical view here is that all cognition is intrinsically social (cp. Hutchins 1995); the moderate view is that there is an important family of processes which constitute an integrated realm of socially-supported cognition.

Social cognition also refers to the bases, in individual minds, of the perception and understanding of people as agents; in more traditional language, one wants to understand what in the mind of individuals makes them capable of supporting intersubjectivity. The field began with the realization that social interactions among apes, or other creatures without language such as infants, heavily depend on the ability of an individual A to attribute to a conspecific B views, *i.e.* beliefs, desires and intentions, of its own, in particular, of holding about a given state of affairs beliefs which differ from A's. Without such a 'theory of mind', it is said, the social life of individuals is quite limited (as in the case of apes), or impaired (as in the case of autistic children). Recently, the basis and resources of this 'naïve psychology' (as this ability or set of abilities is also known) have been further explored. Of particular interest is the discovery of 'mirror neurons' in macaque premotor cortex (Rizzolatti *et al*,), and the postulation of similar 'mirror systems' in humans, which are seen by some as providing the neurophysiological basis for the identification of a conspecific's intentions, and possibly of one own's intentions, no longer considered to be known on the basis of an incorrigible first-person report or intuition. Another line of inquiry is pursued by Michael Tomasello, whose conjecture is that the basis of intersubjectivity is the much more powerful faculty of identifying, and participating in, *shared* intentions –an instance of the general trend of seeing action as more basic for cognition than knowledge or belief.

Phenomenology-2

Action is indeed a key idea, perhaps the single most important factor in the renewal of Cogsci. It seems that sheer reflection has oriented many analytic philosophers, and subsequently or in parallel some cognitive scientists, toward the rejection of the modular view propounded by GOFAI and cognitivism, according to which action is nothing but the result of a planned sequence of motor episodes accomplished by mindless effectors under instructions from Central Control. Actionist views, whose roots plunge in the past of physiology and medicine, but were also adumbrated by Piaget and Merleau-Ponty, among others, are now occupying center stage. What justifies their inclusion under 'phenomenology-2' is the fact that, while acknowledging no direct indebtedness to phenomenology in the historical sense (with some exceptions as we'll see), they start with detailed, unprejudiced examinations of the structure of action, and only then ask how Cogsci can accommodate the phenomenological data thus acquired.

Some examples of currrent action-based research programs are theories of 'active perception' (O'Regan & Noë); the 'new robotics' (Brooks); theories of motor control (Jeannerod & Jacob). They all lead directly to deep and controversial questions about four related topics of central philosophical significance, which space unfortunately does not permit me to discuss here. One is whether we should continue to think of representations as playing a role in, or (as cognitivism holds), quite simply defining, cognition. Anti-representationalism has for many

years been a rallying cry of nay-sayers and rebels of all stripes, and constitute a common ground for thinkers coming from phenomenology, and insiders to AI and Cogsci. Unfortunately, nobody could agree, for a long time, on exactly what this amounted to. Somewhat like antipsychologism at the turn of the 19th century, of which every proponent (there were many) kept accusing the others of being insufficiently firm in their conviction, of still being in the grips of the fatal mistake, antirepresentationalism has been something of a flag the possession of which rival chapels have been fighting over. Recent developments in empirical Cogsci afford a much better chance to treat the issue in a less ideological manner. The second, very closely related, general issue, bears on the notions of non-conceptual content and of thought without language. The third question, also in the immediate vicinity, is that of the body. *Embodied* cognition is another program, research topic, and ralllying cry for many. Analytic philosophers and mainstream cognitive scientists by the hundreds are working at spelling out the implications of the idea that cognition is a property *of* a body *for* a body, and brush shoulders with 'professional' phenomenologists as well as roboticians and 'ALifers' (specialists of Artifical Life). 'Virtual reality', a tool for investigating perceptual and motor capacities in unnatural environments, is also a means to ask questions about 'presence', something to do with a property of objects and persons over and above any of their attributes. The fourth issue concerns autonomy, and the related themes of self-organization and selfhood. On all of these issues, empirical and philosophical inquiries feed one another and in some estimates are profoundly changing the state of the play.

We have just listed, under the headings of phenomenology 1 and 2, a rather overwhelming array of new themes, directions and proposals. It then becomes an issue whether these sometimes connected, sometimes disparate ways of breaking away from the cognitivist tradition can be integrated into a common vision. This is a challenge which has not gone unmet. There are on offer quite a number of proposals for an integrated account, or at least for programs aiming at providing such an account. For the sake of exposition, they can be grouped under three headings : philosophy, science, models.

Philosophical treatments proceed by identifying a master theme to which all or most of the proposals found acceptable can be related. The most popular theme is externalism: philosophers have been arguing for a long time, and from various types of consideration, in favor of (sometimes restricted) versions of externalism (also labelled 'anti-individualism'), contrasted with 'solipsistic' conceptions of the mind. The new perspectives reported in this chapter plead in favor of a generalized form of externalism. One example of a global philosophical treatment along the externalist line is McClamrock (1995). This work sits on the borderline between type 2 and type 3 phenomenological interventions: developed almost in its entirety in purely cognitive-scientific analytic style, it concludes with a brief section in which the historical phenomenological sources are quoted. It also shares with phenomenology a sensitivity to the interplay between ontology (how the world is cut up in regions) and

naturalistic epistemology (how the embedded subject becomes acquainted with, by makes himself at home in, the world). This show of genuine phenomenology is typical of the synthetic attempts described here: (historical) phenomenology can hardly be totally ignored by someone trying to bring all these proto-phenomenological attempts under one roof. Another synthesis, this one inspired by philosophy of science as well as a generally ecological perspective, it that of Sunny Auyang.

The best-rounded synthesis is that by Andy Clark (1997), a philosopher well-versed in Cogsci and not so interested in foundational issues. Clark provides a sketch of what Cogsci might look like if enough of the promises made are kept, and enough overall consistency is maintained. One theme which he weaves in, and which space does not allow to develop in this chapter, is the contribution of neuroscience. It raises foundational problems which are left wide open by Clark, and to which there are today no agreed-upon solutions. Another synthesis straddling philosophy and Cogsci/AI, and taking as its unifying theme the emotions, is DeLancey (2002).

On the AI-modelling side, there are three broad directions of integrative research. The oldest and best-established, connectionism (or neural-net modelling, sometimes also called neurocomputation), grew out of a tradition (see Anderson & Rosenfeld 1989) with no direct links with phenomenology, and in response to phenomenological-2 considerations, but has been greeted by Dreyfus, among others, as a partial fulfilment of the constraints brought to light by (genuine or strict) phenomenological inspection. Neural nets are essentially plastic perception machines which learn by exposure to specific cases, and do not rely on rigid rules; they exhibit a number of properties, such as context-sensitivity and graceful degradation (their performance does not degrade catastrophically, like classical AI programs, when the circumstances of the problem at hand begin to drift away from the normal conditions, those of the learning phase). They are not 'mentalistic' in that they do not rely on the classical notion of representation, they are devoid of anything like a central control unit, they proceed in a massively parallel fashion, and they exhibit a degree of self-organization. However, Dreyfus notes that their disembodiment (they are after all nothing other than programs implementable on a regular computer) stands in the way of their ever becoming true analogs of human intelligence.

This is exactly the objection the 'new robotics' program associated to Rodney Brooks and his collaborators at the MIT AI Lab means to crush in the egg, by starting with a body (the robot) with primitive sensors and effectors, and inviting it to learn and pick up novel, emerging cognitive capacities by interacting (in 'flesh') with the world. Unsurprisingly, Brooks advertises a radical antirepresentationalism; with other ecologically-oriented thinkers such as O'Regan and Noë, he credits Dreyfus with the insight he is working out: the (real) world is its own 'best model'. There is no argument there against anything which has a claim on being a *serious* representationalist view, but the emphasis on real-time, in-flesh interaction has both biological and phenomenological plausibility.

Finally, we come almost full round by returning to AI. AI was under intense pressure to reform: it wasn't working and Dreyfus has given reasons why it shouldn't. Unlike science and unlike philosophy, a technology cannot survive defeat for very long. Many researchers realized that the mind they were trying to emulate was too impoverished for the models it inspired them to be of any use in the 'real world'. As a result, there were a number of attempts to build models with 'consciousness', 'emotions', 'involvement' etc., and this is still going on. But the more interesting current is the one initiated by Terry Winograd and Fernando Flores (1986), who took their inspiration directly from Dreyfus' phenomenological account and to some lesser extent from Searle's theory of speech acts, which has a strong actionist component. The 'existential AI' advocated by Winograd and Flores is realistic enough to forego the pretension of building self-sufficient intelligent artefacts, and is content to aim for 'intelligent', *i.e.* adaptive, useful, interfacing tools for working communities. In this endeavor, they put phenomenological descriptions of engagement, thrownness, commitment, equipment, etc, to interesting use: a suitably programmed computer is first and foremost a piece of equipment, and an understanding of tool-use is a prerequisite, it would seem, for the manufacture of useable tools ('usability' has indeed become a central concern of computer and communication technologies). There is however the shadow of a possible paradox here: if the lesson from existential phenomenology is that theory is not the royal route to practices (and surely kayaks, hammers, pubs and even guns did not result from better theories), perhaps theorizing about the right sort of software is not the right way to get it. There is a mirror-image to this apparent paradox: when von Neumann was told that there was something wrong with his machine, he rhetorically asked to know *exactly* what was missing (the obvious implication being that with this information in hand, the machine could be fixed). It is far from obvious that there is a way out of these dual conundrums.

Phenomenology-3

As we near the end of this chapter, we come to what some would regard as the philosophical core of its topic: what can (genuine, historical, strict) phenomenology contribute *today* to Cogsci, and in virtue of which of its features? As a very partial answer, complementing the first part of the chapter, some directions of inquiry pursued within a European school of 'naturalized phenomenology' in the spririt of Husserl rather than Heidegger will be briefly presented (cp. Petitot *et al.*1999).

The first is formal ontology. This project of Husserl's [REF TO APPROPRIATE CHAPTERS] is being revived by scholars of Austrian philosophy before and beyond Husserl, and incorporates themes from ecological vision (Gibson) as well as Gestalt psychology. Mereology is a central part of the project (see Smith 1982), but it turns out that there are many domains, such as places (Casati & Varzi 1999), sounds, holes, shadows, objects, ... which are in

need of a formal characterization. The immediate applications concern semantics and perception, and the possible relation between them.

Another project takes up directly from Husserl, and attempts to enroll the tools of contemporary mathematical physics to show that Husserl's objections to the naturalisation of eidetic contents were based on an outdated stage of scientific development, and that physics has now the means to detect the objective morphological structures in the natural processes subtending perception which alone can be put in correspondence with the eidetic contents. The optic flux is not an amorphous sheaf of energy, it possesses enough structure to allow the visual system to 'interpret' the 'sense data', and today's mathematical physics can provide an objective account of this interpretative process.

However, the most popular route from Husserl and Merleau-Ponty to Cogsci, within this school of thought, goes through the analysis and examination of the various modes and levels of intentionality. Here are a couple of examples of what Gallagher (1997) calls, after Varela et al. (1991), 'mutual illumination' between Cogsci and phenomenology. Sean Kelly examines Merleau-Ponty's account of 'motor intentionality' and the body's tendency towards maximum grip as an experiential equilibrium, and shows how it supports, and is in turn supported by, the account of visual perception by cognitive neuro-scientists such as Milner and Goodale, and such dynamic brain models as Walter Freeman's, which involve chaotic attractors. The second example is the parallel which R. McClamrock draws between Husserl's conception of the relation between noesis and noema and Cogsci's intuitions about multiple realizability (several physical processes filling the same cognitive function) and context-dependence (one physical process filling different cognitive functions). Another example is Tim van Gelder's account of time consciousness based on second-hand summaries of Husserl's work (*in* Petitot *et al.*1999). Such attempts raise a couple of questions about the possibility and meaning of 'naturalizing' Husserl against his own expressed intentions. One is whether such an enterprise is coherent, and whether it really is possible to peel apart the side of Husserl's thought which can be put to use in Cogsci from his anti-naturalistic arguments. The second concerns the chances of success of a naturalized version of Husserl. On this last point, opinions differ: Petitot et al. disagree with Dreyfus's cognitivist-representationalist interpretation of Husserl. On the first question, there are other disagreements: here Petitot and Dreyfus find themselves in the same camp, together with (to some extent and under proper reading) Daubert, Merleau-Ponty, Aron Gurwitsch, Ervin Straus or Roger Chambon, facing opposition by what remains to this day a vast majority of Husserl scholars.

However these issues concern phenomenology more than they do Cogsci, whose sole interest is to take clues anywhere it can find them to make progress in its quest for scientific advances. And this leads to a final question, which concerns the weight of phenomenological reports as prima facie, defeasible evidence to which Cogsci is bound, pending countermanding instructions from psychology or neuroscience. Considering the plain fact that every cognitive

scientist starts with some rudimentary form of phenomenological account, is it not a case of misplaced methodological perfectionism to bar more refined accounts from the scientific process?

DANIEL ANDLER

References and Further Reading

Anderson, J. & Rosenfeld, E., eds. (1989). Neurocomputing: Foundations of Research. Cambridge, MA: MIT Press.

Andler, D., ed. (2004). Introduction aux sciences cognitives. Paris: Gallimard.

Andler, D. (2000). The normativity of context. Philosophical Studies, 100, 273-303.

Auyang, S.Y. (2001). Mind in Everyday Life and Cognitive Science. Cambridge, MA: MIT Press.

Baars, B.J., Banks, W.P., Newman, J.B., eds. (2003). Essential Sources in the Scientific Study of Consciousness. Cambridge, MA : MIT Press.

Bechtel, W. & Graham, G., eds. (1998). A Companion to Cognitive Science. Oxford: Blackwell.

Bermudez, J.L (2003). Thinking Without Words. Oxford: Oxford University Press.

Bermudez, J.L, Marcel, A., Eilan, N., eds., (1995). The Body and the Self. Cambridge, MA: MIT Press.

Berthoz, A.(2000). The Brain's Sense of Movement. Cambridge, MA: Harvard University Press (Original work published 1997).

Block

Brooks, R. (2002). Flesh and Machines. How Robots Will Change Us. New York:Vintage Books.

Carruthers, P. & Smith, P.K., eds. (1995). Theories of Theories of Mind. Cambridge: Cambridge University Press.

Casati, R. & Varzi, A. (1999). Parts and Places: The Structures of Spatial Representation. Cambridge, MA: MIT Press.

Chambon, R. (1974). Le Monde comme perception et réalité. Paris: Vrin.

Clark, A. (1997). Being There. Putting Brain, Body, and World Together Again. Cambridge, MA: MIT Press.

Davidson, R.J., Scherer, K.R., Goldsmith, H.H., eds. (2003). Handbook of Affective Sciences. Oxford: Oxford University Press.

DeLancey, C. (2002). Passionate Engines. What Emotions Reveal about the Mind and Artificial Intelligence. Oxford: Oxford University Press.

Dennett, D.C. (1991). Consciousness Explained. Boston: Little Brown.

Dreyfus, H.L. (1972). What Computers Can't Do. A Critique of Artificial Reason. New York: Harper and Row. Revised edition (1979). Augmented edition (1992), What Computers Still Can't Do. Cambridge, MA: MIT Press.

Dreyfus, H.L., ed. (1982). Husserl, Intentionality and Cognitive Science. Cambridge, MA: MIT Press.

Dreyfus, H.L. & Dreyfus, S.E. (1986). Mind over Machine. The Power of Human Intuition and Expertise in the Era of the Computer. Glencoe, IL: The Free Press.

Fodor, J.A. (1983). The Modularity of Mind. Cambridge, MA: MIT Press.

Freeman, W.J. (1999). How Brains Make Up Their Minds. London: Weidenfeld & Nicolson.

Gallagher, S. (1997). Mutual enlightenment: Recent phenomenology in cognitive science. Journal of Consciousness Studies, 3, 195-214.

Gibson. J.J. (1979). The Ecological Approach to Visual Perception. Boston: Houghton-Mifflin.

Griffiths, P. (1997). What Emotions Really Are. Chicago: University of Chicago Press.

Gurwitsch, A. (1966). Studies in Phenomenology and Psychology. Evanston, IL: Northwestern University Press.

Haugeland, J. (1985). Artificial Intelligence. The Very Idea. Cambridge, MA: MIT Press.

Haugeland, J., ed. (1981). Mind Design. Cambridge, MA: MIT Press.

Hutchins, E. (1995). Cognition in the Wild. Cambridge, MA: MIT Press.

Jacob, P., Jeannerod, M. (2003). Ways of Seeing: The Scope and Limits of Visual Cognition. Oxford: Oxford University Press.

Levine

McClamrock, R. (1995). Existential Cognition. Chicago and London: Chicago University Press.

Mele, A. (1992). Springs of Action: Understanding Intentional Behavior. Oxford: Oxford University Press.

Merleau-Ponty, M. (1962). The Phenomenology of perception. (C. Smith, Trans.). London: Routledge and Kegan Paul (Original work published 1945).

Nagel

Neisser, U., Fivush, R., Hirst, W., eds. (1999). Ecological Approaches to Cognition: Essays in Honor of Ulric Neisser. Mahwah, NJ : Lawrence Erlbaum.

O'Regan, J.K. & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. Behavioral and Brain Sciences, 24(5), 939-1011.

Petitot, J., Varela, F.J., Pachoud, B., Roy, J.-M., eds. (1999). Naturalizing Phenomenology: Issues in Contemporary Phenomenology and Cognitive Science. Stanford: Stanford University Press.

Port, R.F., & van Gelder, T., ed. (1995). Mind as Motion: Explorations in the Dynamics of Cognition. Cambridge, MA: MIT Press.

Pylyshyn, Z. (1984). Computation and Cognition. Toward a Foundation for Cognitive Science. Cambridge, MA: MIT Press.

Rizzolatti, G., Fogassi, L. & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. Nature Neuroscience Reviews, 2, 661-670.

Rumelhart, D, McClelland, J. & the PDP Research Group (1986). *Parallel Distributed Processing. The Microstructure of Cognition.* 2 vols. Cambridge, MA: MIT Press.

Searle, J. (1983), Intentionality : An Essay in the Philosophy of Mind. Cambridge: Cambridge University Press.

Smith, B., ed. (1982). Parts and Moments. Studies in Logic and Formal Ontology. Munich: Philosophia.

Smolensky, P. & Legendre, G. (2005). The Harmonic Mind, Cambridge, MA : MIT Press.

Tomasello, M. (1999). The Cultural Origins of Human Cognition. Cambridge, MA: Harvard University Press.

Varela, F., Thompson, E. & Rosch, E. (1991). The Embodied Mind. Cognitive Science and Human Experience. Cambridge, MA: MIT Press.

Weiskrantz. L. (1988). Thought Without Language. Cambridge: Cambridge University Press.

Winograd, T., Flores, F. (1986). Understanding Computers and Cognition. A New Foundation for Design. Norwood, NJ: Ablex (repr. 1987 Reading, MA: Addison-Wesley).

Wittgenstein, L. (1953). Philosophical Investigations. London: Macmillan.

Wrathall, M. & Kelly, S., eds (1996). Existential Phenomenology and Cognitive Science, an issue of The Electronic Journal of Analytic Philosophy. http://ejap.louisiana.edu/EJAP/1996.spring

Wrathall, M. & Malpas, J., eds. (2000). Heidegger, Coping, and Cognitive Science, Essays in Honor of Hubert L. Dreyfus. Cambridge, MA: MIT Press.