AI & SOCIETY (2019) 34:165–175
https://doi.org/10.1007/s00146-019-00884-0

EDITORIAL

From judgment to calculation: the phenomenology of embodied skill

Celebrating memories of Hubert Dreyfus and Joseph Weizenbaum

Karamjit S. Gill[1]

Published online: 18 March 2019

In celebrating the memories of our two founding advisory board members, Hubert Dreyfus and Joseph Weizenbaum, we celebrate their contributions to the vision of AI&Society—a humanistic vision of the interplay of art, science, technology and society, an alternative to the technocentric and deterministic narrative of artificial intelligence. During my visit to MIT in 1985, Joseph Weizenbaum not only inspired me to launch an AI journal on the theme of knowledge and society, his seminal book, Computer Power and Human Reason, especially his concern of the turn from of judgment to calculation, became and remains one of the central driving force of the Journal. Weizenbaum's interest in AI&Society was much more than mere academic curiosity; he proactively introduced the idea of the journal to the AI community in the USA, including Hubert Dreyfus, Terry Winograd and David Noble. What remains a fond memory for me is Joseph Weizenbaum walking with me to the MIT Media Lab, introducing me to Marvin Minsky who in turn introduced me to his research team including Seymour Pappert who also became a member of the Advisory Board the journal in the early days. Although many of the Journal Board members were conversant with Hubert Dreyfus' critique of artificial intelligence in his seminal book, '*What Computers Can't*, he makes his first personal contribution to AI&Society debate on the limits of cognitivist view of mind in his talk "Is Socrates to Blame For Cognitivism?" at the International Conference on Language, Culture and Artificial Intelligence, held in Stockholm in 1988. This was followed by the publication of his paper on "The Socratic and Platonic Basis of Cognitivism" in AI&Society (1988). Through his association with AI&Society, Hubert not only contributed to shaping the debates on Expert Systems in the 1980s, but he also inspired many authors to contribute articles to AI&Society to reflect on the ongoing AI debates whether they be on the Internet, neural networks, robotics, body or Jazz. The 1980s was also an era of academic soul searching in response to the political narrative in the UK on the lines that 'there is no such thing as society, only individuals'. Reflecting back upon this narrative of 34 years ago, the launch of 'AI For Society' conferences (Gill 1986) and the foundation of the 'AI&Society' journal were seen as the most appropriate academic response at that spur of the moment. It was also a time when some of the Journal founding editors, David Smith was leading research into IT and Education (ESRC), Richard Ennals (Imperial College, London) was managing the British Fifth Generation Programme, Alvey, Mike Cooley (1987) was directing the radical socially useful technology networks at the Greater London Enterprise Board (GLEB), Bo Corazon (Stockholm) was leading the European research network on Language, Culture and Artificial Project, Massimo Negrotti was leading an intellectually stimulating research into 'Culture of the Artificial' at Urbino, and Satinder Gill at Cambridge was making the tacit dimension of knowledge a core driving concept of AI&Society. Our editors of 1990s, Victoria Vesna (UCLA), Sha Xin Wei (Arizona) and Larry Stapleton (WIT) keep alive the humanistic spirit of research envisioned by Dreyfus and Weizenbaum. Whilst to many in the AI community, Hubert is remembered as an established author, an engaged teacher, an admired colleague, a creative collaborator, a committed friend, and a charming storyteller, he together with Weizenbaum remain inspirational guides and resource to many of our authors and readers, as illustrated by the tributes to them by our authors in this volume.

The deep concern of instrumental reason articulated by Weizenbaum (1976) continues its march in the guise of Big Data machine learning algorithms. We see an increasing manipulation of data to support and control human interactions, institutional and organizational structures. Moving

✉ Karamjit S. Gill
editoraisoc@yahoo.co.uk

[1] University of Brighton, Brighton, UK

beyond their (algorithms) role as computational artefacts, what concerns us is how these algorithms take account of the limits of our 'entrenched assumptions about agency, transparency, and normativity'. Reflecting on these issues, Gill (2017) draws our attention to data manipulation practices as problematic because they are inscrutable, automatic, and subsumed in the flow of daily practices. Beyond the issues of algorithmic transparency and openness, calculative practices have a serious impact on how domains of knowledge and expertise are produced, and how such domains of knowledge become internalized, affecting institutional governance. Moreover, these algorithms not only work within 'highly contested' online spaces of public discourse, they often perform with little visibility or accountability. This is an argument to move out of the 'black box' notion of the algorithm, and promote the idea of 'networked information algorithms' (NIAs); assemblages of institutionally situated code, practices, and norms with the power to create, sustain, and signify relationships among people and data through minimally observable, semi-autonomous action. If AI reflections are to move out of the 'black box' of instrumental reason, we need to learn from the performance practices of artists, where performance of data is seen not just in terms of its transformation into information, but also in terms of the interactivity between the artist and the audience. This interactivity itself becomes a tool for the continued evolution of an artist and a scientist and the amalgamation of their partnership. In the end, performance is about raising awareness of the interconnectivity of everything and everyone. Technology is or should be utilized to amplify the experience and/or the range of influence. As wearable sensors proliferate, we have access to rich information regarding human movement that gives us insights into our daily activities like never before. In a sensor-rich environment, it is desirable to build systems that are aware of human interactions within contextual and relational situatedness. Experiential scientists, craftspeople, medical practitioners and engineers transform raw data into information, then using their skills and experience transform information into knowledge, and through the application of their contextual knowledge and wisdom, make judgements about the accuracy, relevance and acceptability of data that are coming from many sources. In this transformation process, there is always a scope for human intervention at various levels of the data-to-action cycle, and that intervention, which reflects the many overlapping contexts, would bear witness to situated judgements. This is in contrast to an intervention based upon machine learning algorithmic calculations. In other words, the performance of data, in the hands of expert practitioners, is seen here in terms of an evolving judgement-making process culminating in action. This transformational process from data to action, encompassing feedback loops and human intervention, provides a human-centred perspective of judgement that is contrary to the computational model of 'calculation to judgement', in which data are used to compute judgement. We should, however, recognize that the computation model of judgement, turning judgement into an algorithm, is still a dominant focus of the data-driven AI. It may be tempting to argue that nothing has fundamentally changed in the data-action cycle except for the availability of an abundance of data (big data) and the exponential processing speed of computers. The fallacy of this argument then revolves around the idea that only if we have an abundance of data and the exponential processing speed of the computer, can we construct machine learning algorithms that can outstrip human cognition, to the extent that machines can better humans in processing a wider variety and larger number of data sets and working in different ways to those of humans in reaching analytical judgements. However, this calculation-centred view of judgement fails to recognize that human judgement is about the process of finding a coherence among often conflicting and yet creative possibilities that cannot be reduced to calculation. Moreover, human judgement resides in and reflects the dynamic and evolving nature of professional and social practices, enriching human experience, knowledge, skill and cognition. From this human-centred perspective, performance of data lies in the performance of practice of the 'data-action cycle'; in other words, the performance of inter-relations between data, information, knowledge, wisdom and action. This view seeks to understand the nature of the interface between the physical, cultural and our experiential worlds. The nature and practice of the interface here are fundamentally between, in-between, and across knowledges, experiences and practices of contextual domains, and not transactional in the sense of 'cause and effect' calculation (Gill 2015). This view shifts our attention from a purely technological fascination of machine learning to the evolving interaction of human systems and technology, thereby providing a symbiotic horizon of performing data. In the midst of the fascination with digital technology, we are cautioned to remember that performance of data in the hands of creative artists and scientists embodies social/cultural and spatial intelligence that conforms to the living.

The single story of the universality of instrumental reason (Weizenbaum 1976) has so penetrated the culture of computation and machine learning that questions and challenges of human purpose are either ignored or misrepresented as if every aspect of real world can be formalized and represented in term of logical calculus. But what are the assumptions, Weizenbaum notes, behind the single computational story? (1) Computers have the power to acquire human knowledge beyond information, in other words, come to know what humans know, how and why they know; (2) all human knowledge is encodable in "information structures", and the acquisition of that knowledge is a function of the brain alone. This acquisition concept ignores that knowledge is

bodily and involves, in part kinesthetic. No organism that does not have a human body can know these things in the same way humans know them. Every symbolic representation of reality must lose some information that is essential for some human purpose; (3) some things people come to know only as a consequence of having been treated as human being by other human beings; (4) the kinds of knowledge that appear superficially to be communicable from one human being to another in language alone are in fact not altogether so communicable; (v) further, language in human use is not merely functional in the way that computer languages are functional. It does not identify things and words only with immediate goals to be achieved or with objects to be transformed. The human use of language manifests human memory. And that is a quiet different thing than the store of the computer, which has been anthropomorphized into "memory". The former gives rise to hopes and fears, for example. It is hard to see what it could mean to say that a computer hopes. These considerations touch not only on certain technical limitations of computers, but also on the central question of what it means to be a human being and what it means to be a computer. Weizenbaum (1976) points out that those who aspire to equating machine intelligence to human intelligence keep convincing themselves that by outplaying human go players, composing music or creating human such as social robots, machines have either already or soon going to outsmart human beings. This belief in machine intelligence sees no distinction between the functional machine and the knowing and imaginative human being. It seems that in this pursuit of machine intelligence, the validation of human intelligence has been reduced to the display of technological wonders, just as scientific knowledge has been reduced to wonders of data science.

Bo Goranzon in his introduction to Joseph Weizenbaum's seminal paper, 'The Last Dream' (this volume) speaks of fulfilling the last dream by creating an artificial intelligence, that is, by constructing an artefact that is to function like the human mind. Weizenbaum's intention is to question whether or not there are aspects of reality about which science cannot alone inform us. A closely related question is whether there is such a thing as dangerous knowledge and, if there is, whether such a thing as forbidden knowledge. In an article from 1950 entitled 'Computing Machinery and Intelligence' Turing stated that it was his conviction that computers should be able to imitate human behaviour perfectly and that this goal would be attained by the year 2000. This article presented a method of defining intelligence; the so-called *Turing test*. A person placed in one room and a computer in another. Both are able to communicate with the outside world but only through the medium of typewritten texts. Another person is placed in a third room and, after questioning both these intelligence, has to decide which of them is human. Turing maintained that if the interrogator fails in his task then one must ascribe intelligence to the computer. Descartes designed a tougher version of Turing test. The Cartesian test looks like this: before it can be judged to be intelligent, a machine must be capable of language actions and sensible actions independent of the programmer. Descartes arrived at a completely different conclusion to Turing's. The difference between man and animal or machine is that because man has a language, and the way he formulates his concepts. It is the intellectual position of the *Turing test* which Joseph Weizenbaum attempts to tackle in his now the classic *Eliza* example. He wanted to make people more aware of the limitations of computers by developing software, which simulates a psychotherapist's questions and responses. Weizenbaum found the reaction totally unexpected. The psychoanalysts directly affected by the application were enthusiastic. They saw the possibility of acquiring an instrument of shortening the waiting times for patients in psychiatric care. If the Eliza method proves beneficial, then it would provide a therapeutic tool, which could be made widely available to mental hospitals and psychiatric centres suffering a shortage of therapists. A computer system could deal with several hundred patients an hour. A human therapist can be viewed as an information-processor and decision-maker who is applying a set of decision rules which are closely linked to short-range and long-range goals. He is guided in these decisions by rough empirical rules telling him what is appropriate to say and not to say in certain contexts. Weizenbaum's application demonstrates that the Turing test is not a satisfactory one. It can be misleading. This gave Weizenbaum an insight into a fundamental problem; human beings are liable to attribute to new technology, in this case, a diagnostic programme in the field of medical care, more intelligence than it possesses. We lose our distance. We fail to realize what the limitations are. This is a feature of the emergence of a technological culture.

Massimo Negrotti in this volume provides an insight into the two concurrent mainstreams of AI of the past century, the symbolic and the neuronet, arising from philosophical traditions of rationalism and holistic schools. However, he surmises that we have observed little or no commitment of AI researchers to investigate and 'verify' such philosophical doctrines; rather, they have attempted to get results from the machine that could be similar to the ones we normally get on a daily basis from a human being. Negrotti pays tribute to the philosophical commitment of Dreyfus to provide an alternative vision to the rationalism of AI claims. It can be seen as a matter of a true struggle between two cultures, say humanities and technology that is still today in some measure alive. The main argument proposed by Dreyfus originated from his view on human perception and evaluation of things and situations, as something that comes from a plurality of sources, including human culture that cannot

be described by means of a symbolic strategy. The argument is that the core of human ability to understand and solve problems consists of its being able to act by means of intuition rather than by means of reasoning conceived as a conscious and formal calculation. As a consequence, a computer program cannot reproduce human mind because it lacks the human plural abilities. This conflict, Negrotti says, between two philosophical orientations arises from the polarization on the pole of intelligence that was clearly due to the apparent place that the concept of 'intelligence' has among the main concepts of philosophy and also in common sense vocabulary. The rise of technology of the 'artificial' that claimed to be able to reproduce such a human function could only arouse an intense series of issues, oppositions, needs of clarification, up to a sort of intellectual rebellion. It was, however, clear that this concept of the 'artificial' was assumed in a rather trivial way, that is to say, in the way we define it in everyday life: something manmade or, likewise, something, which is not natural. The reason for this is very simple: while conventional technology design, to be successful, must match nature exploiting its resources and respecting the constraints it poses, the technology of naturoids (artificial) must face an additional key problem: that of knowing (and then to modelling), in the best analytical way as possible, the natural thing or the process it wants to reproduce. To seek clarification, we turn to Dreyfus who focuses on the difference between formal logical calculation and intuition. His argument is that while the former occurs through explicit rules, the latter takes place without any reference to learned rules. Nevertheless, Dreyfus's five-stage skill acquisition model that allows an expert to make decisions neglecting the formal rules, does not explain in itself what intuition is. If intuition implies an implicit process of not only as a tacit process but even a sort of mysterious affair, then the field is open to any possible hypothesis. One can think, for example, that intuition can be a shortcut our brain follows applying the rules in the background and, supposedly, through a very high speed since it does not need to resort to more slow processes, analytical and language based, like a novice does retrieving the rules one by one. One can reasonably maintain that an expert follows *his own* rules without any conscious reference to the learned ones, but his personal rules should be nothing but a sort of synthesis substantially consistent with the learned ones. Otherwise, the failure of his decisions would be sure. This is why, as a matter of fact, no expert could be able to exhibit intuitive ability if he never learned the rules. In the field of 'musical intuition', for instance, Jean-Philippe Rameau in his *Le nouveau systeme de musiquetheorique* published in 1726, noted that 'While composing music is not the time to recall the rules which might hold our genius in bondage. We must have recourse to the rules only when our genius and our ears seem to deny what we are seeking'. Negrotti says that Dreyfus and other authors, in their criticism of AI claims, confronted researchers who really believed that the growing power of computers could open the door for capturing the roots of human thinking. In doing so, they exposed themselves to critical scrutiny. Nevertheless, having shown that AI is not the right road to follow for understanding human thinking should not prevent us to recognise AI's outcomes that are compatible with our nature and our needs.

Rafael Capurro in this volume recounts his 'A Long-Standing Encounter' with Hubert Dreyfus during his visit to the University of California, Berkeley in November of 1992. Capurro expected a US philosopher to be analytic and to reject Continental philosophy in general and Heideggerian phenomenology in particular. What he discovered was a philosopher in Dreyfus deeply conversant with phenomenological tradition through his book *What Computers Can't Do: The Limits of Artificial Intelligence* (Dreyfus 1978). Capurro says that both thinkers, Hubert Dreyfus and Joseph Weizenbaum helped him further develop his ideas on Information Ethics, especially Dreyfus' book, *Being-in-the-World: A Commentary on Heidegger's Being and Time*, in the wider context of Kierkegaard, Heidegger and Wittgenstein, the sources of morality and their relevance for Intercultural Information Ethics. Dreyfus presents a deeper understanding of the issues at stake. Capurro cites Hubert Dreyfus' concluding remarks in his book *On the Internet*: in sum, as long as we continue to affirm our bodies, the Net can be useful to us in spite of its tendency to offer the worst of a series of asymmetric trade-offs: economy over efficacy in education, the virtual over the real in our relation to things and people, and anonymity overcommitment that our culture has already fallen twice for the Platonic/Christian temptation to try to get rid of our vulnerable bodies, and has ended in nihilism. This time around, we must resist this temptation and affirm our bodies, not in spite of their finitude and vulnerability, but because, without our bodies, as Nietzsche saw, we would be literally nothing. Hubert Dreyfus' initial analysis in 1972 of "what computers can't do" finds an echo in present-day research into what algorithms can and can't do as well as in discussions of what they should and should not be allowed to do.

Peter Broedner, in Coping with Descartes' Error in Information System (this volume), reminds us of two fundamentally different thought traditions: the first "being-in-the-world"—a perspective of skillfully interacting and coping with it, and the second the representational perspective of conceptually mediated world recognition. While Dreyfus argues for the relevance of the first view, the second has dominantly influenced Western thinking since the days of Descartes' dualism separating mind from matter. Both these world views have been propagated for a long time in isolation and opposition with each other. Moreover, both correlate with two fundamentally different types of knowledge,

being characterized as implicit and embodied or experiential knowledge (e.g. "knowing how" versus "knowing that" or the "tacit dimension"), as opposed to explicit and conceptual or propositional knowledge. Both views can claim to be based on sound evidence, and it seems impossible to refute one or the other. The long-lasting controversy between both perspectives is, however, by far not an idle dispute among academics but rather of particular relevance to social practices dealing with complex computing machinery. It forms the epistemological background for designing, evaluating, and appropriating this specific semiotic type of machinery operating signs as distinct from energy or material transforming machinery operating with forces. Brodner argues that the basic mistake made by the protagonists of the representational world view, from Descartes' *cogito* to the present cognitivist and AI communities, consists in regarding conceptual cognition as exclusive access to the world for humans. With this stance, they ignore or even deny the biological roots of human cognition, the existential fact of bodily being-in-the-world with its immediate intuitive perception in situated action. Prior to conceptual cognition, successful continued action in and interaction with the surrounding physical and socio-cultural world, based on collective intentionality, holistic perception and immediate experience, are at the core of sense-making and human intelligence. It expresses itself as tacit "knowing how" and skillful acting which allocates pre-representational meaning to things, to dealing with them, and to interactions with others, thus constituting a social practice. Computer systems, even those being enabled to adapt to environmental conditions by sensor data, attain their functionality solely through conveniently designed algorithms from outside, based on propositional knowledge about their field of application. They, therefore, are lacking own intentionality and self-determined activity as indispensable material basis for perception, sense-making, and experience. Attributing "artificial intelligence" to computer systems distracts the awareness from the fact that all intelligence is on the side of the programmer who provides the computer system with functions that, as its objectifications, solely mimic or simulate intelligent behaviour in a generally limited way. Broadner concludes that more recent efforts on the part of AI and robotics for "embodying" their systems in order to broaden the range of automatic behaviour do not really change the picture. They all amount to the implementation of purely physical, mainly mechanical or electrical devices enabling sensor-controlled automatic movement in a physical space. This reductionist view of "embodiment" does not with any respect transcend the border to a living body deliberately and autonomously acting in the physical and socio-cultural world around it from which meaning arises.

Liberati and Nagataki in this volume remember Hubert Dreyfus in relation to their deep concern of human vulnerability in the pervasive introduction of robots into our everyday life, raising the issue of peaceful co-existence of the human and the machine. Drawing upon Merleau-Ponty, they describe 'peaceful coexistence' as mutual understanding between oneself and others: a relationship holding on a 'common ground' of consciousness, or the intersubjective world of perception. To them, Dreyfus deeply connects intelligence and body based on a phenomenological viewpoint, in the sense that an intelligence must be embodied into a body in order to function. According to his suggestion, any AI designed to be human-like is doom to failure if there is no tight bound with a human-like body. For the authors, the introduction of vulnerability into robots is not a mere neutral introduction which turns robots into something more than mere tools. This introduction touches directly the constitution of the human subjects too. The robots, through their new vulnerabilities, shape the ethical choices the human beings are called for to make. In addition, thanks to their different gaze, they make the human subjects naked and vulnerable in different ways. The introduction of vulnerability into robots can be seen as a way to elevate them from mere tools and, at the same time, as a way to modify who we are.

Harry Collins in his memorial on Dreyfus, in this volume, remembers him as an argumentative friend during their intense encounters at conferences. Collins says that their arguments were always enjoyable but they never got anywhere in terms of changing the other's mind. In retrospect, Collins puts the enduring argument between them: it was about the relative role, in the acquisition of expertise and understanding, of physical practice, on the one hand, and spoken discourse on the other. For Dreyfus, it was all practice, for Collins, language is central even to the mastering of a practice, and language is a property of society not the individual body. For Dreyfus, there are two kinds of knowledge stuff: mathematics and physics on the one hand, which is computerisable, and most of the rest of our way of being in the world, which is not. It is that distinction which sociology of scientific knowledge had dissolved and replaced with two ways of 'attending' to the world, there being no deeper difference of principle. Becoming a social being within an existing society can be accomplished by coming to share the spoken language alone even though, as it happens, for most people it involves sharing a subset of practices too. Collins says that the concepts that are acquired through language alone, if this is how it comes about, will have been initially developed in the collectivity, or form of life, in concert with the associated practices. It is crucial to get away from the old binary division between practice as a means of acquiring tacit knowledge and talk as a set of formulae with the symbols having the same kind of role as they do in mathematics; speech is a highly tacit-knowledge laden activity: fluent language-speaking is a practice! In more recent years, the role of language has been worked out

in more detail under the heading of 'interactional expertise'. The tension was about *why* computers could not do things: to repeat Dreyfus' 'bottom line' was the body, to Collins it was embedding in society. This difference has born more fruit as far as Collins is concerned in his book, *Artifictional Intelligence* (2018), as a critique of deep learning computers which is based on questioning their degree of true embedding in society. Another thing that puzzled Collins about Dreyfus was his idolisation of Martin Heidegger, even to the extent of having a mocked-up photograph of him sitting in his car with Heidegger on the cover of the 2000 Dreyfus Festschrift. Collins says that Dreyfus told him that Heidegger was the greatest philosopher of the Twentieth Century (with Wittgenstein at number two) and that was all there was to it. If Dreyfus' ideas were great, then he was great irrespective of the fact that the man himself acted completely outside the norm of universalism in the vilest way. According to Collins it was, of course, Dreyfus's universalism that also made it that he was as ready to speak in a completely engaged and sincere way with someone like him even though, year after year, Collins could never convince Dreyfus that he was in any way right about our differences.

Steve Torrance and Frank Schumann in their article on 'The Spur of the Moment: What jazz improvisation tells cognitive science' (this volume) provide a rich source of insights into improvisation that go beyond Dreyfus's notion of skilled coping, for example, through the central enactivist notion of 'sense-making'. In their suggestion of improvisation as an extension of enactivist theory, they see expert improvisers, in music and in life, as walking on a path of open-ended expansion of their mindful experiential relation with their doing. At the heart of an improviser's expertise (and of day-to-day living), lies a form of 'higher-level inner sense-making' that spontaneously creates novel forms of agentive goal-directedness in the moment. Torrance and Schuman argue that their account thus supplants Dreyfus's idea of the ego-less absorbed expert by that of an improviser enacting spontaneous expressions of a self, in music or in life. For them, improvisation is ubiquitous in life, it deserves to occupy a more central role in cognitive science. The case of jazz improvisation acts as a rich model domain from which to explore the nature of improvisation and expertise more generally. In exploring the activity of the jazz improviser against the theoretical backdrop of Dreyfus's account of expertise and of enactivist and 4E accounts of cognition and action, they draw our attention to improvisation in the relation to absorbed coping. They then challenge Dreyfus' claim that expert performance in-the-moment, in particular if improvised, is mindless or egoless. They further conclude that the Dreyfusian account has to be carefully rewritten. They suggest that expert improvisers do not enter a stage of absorbed mindless coping, but are instead on a path of open-ended expansion (and sometimes transformation) of their mindful experiential relation with their doing. For instance, a player's engagement with the piano's keys may unify her with the piano's sound, until she becomes habituated to perceive and act with direct, non-deliberative, intentional agency on musical features such as melodies, the form of the flow, the affective connotations of the playing, and so on. However, with jazz improvisation the abilities of players span a wide spectrum of performance levels. Thus, jazz presents ample opportunities to study embodied improvisation at multiple levels of proficiency, yielding insights into how such skills develop.

Mark Coeckelbergh, in 'Skillful coping with and through technologies' (this volume), argues that Dreyfus's account of skillful coping can be developed into a general view about handling technology which gives due attention to know-how/implicit knowledge and embodiment. Coeckelbergh sees Dreyfus's account of skillful coping as a contribution to thinking about technology, to the extent that it enables us to say more about the kind of knowledge and experience involved in the use of everyday technologies such as hammers, cars, and doorknobs. On knowledge and trust, he notes that the assumption made by some authors that trust necessarily relies on explicit knowledge and the suggestion that Dreyfus's model of skill acquisition necessarily leaves out the role of the coach are problematic. Trust does not require that everything is made explicit; on the contrary, otherwise there is no need for trust. Trust is not necessarily the outcome of a rational process but, according to a 'social-phenomenological' view, is already embedded in the social and in relations. Coeckelbergh makes us aware of how Dreyfus's account of skillful coping relates to virtue ethics, albeit a virtue ethics of a particular kind: one that does not necessarily involve reasoning and judgment, but rather a knowing-how to best respond to the world and to others in particular situations. Compared to the often shallow contemporary debates about AI, robotics, or trans-humanism, for example, Dreyfus's thinking is a relief. His philosophical arguments against the idea that AI can give us a human-like general intelligence are still highly relevant today, when absurd Platonic-Cartesian ideas such as mind uploading or projects trying to artificially recreate the human brain are gaining more traction than they deserve. It is also stressed that the general direction Dreyfus took in his thinking on skill is not only important for philosophy of technology, but also stands as one of the conceptual building blocks *all* philosophers have at their disposal today to cope with a ghost that continues to haunt both philosophy and AI: the 'crazy' idea that knowledge and thinking can be entirely formalized and divorced from lived experience and active coping in the world.

Whilst many AI futurists, enchanted by the creation of disembodied machines to mimic higher mental functions, embrace the idea of "spiritual machines" and "mind

children", Min-Sun Kim in 'Disembodied Evolution' in this volume turns to Hubert Dreyfus's manifesto ("What Computers Can't Do") for an alternative embodied view of human relationships with the machine. This view on the inherent inability of disembodied machines to mimic higher mental functions draws on Michael Foucault's notion of a utopian ideal of the relation of the body to technology. He defines Utopia in his short essay, 'Utopian Body as, ...a place outside all places but it is a place where I will have a body without a body that will be beautiful, limpid, transparent, luminous, speedy, colossal in its power, infinite in duration. Untethered, invisible, protected-always transfigured. It may well be that the first utopia, the one most deeply rooted in the hearts of men, is precisely the utopia of an incorporeal body'.

Citing Kurzweil, Min-Sun Kim says that according to the disembodied version of evolution, it will be a being that, if it ceases to be human at all, will not be abandoned as a redundant shell. The brain will finally be free to travel among the stars. Through the mechanically embodied existence we may find salvation from the limits of bounded experience and human finitude.

Nadin in this volume reminds us of the intellectual challenge of Hubert Dreyfus and Joseph Weizenbaum to the emergence of symbolic processing as the assumed embodiment of artificial intelligence, in the broader context of algorithmic computation. He takes us back to the historically memorable moment of the Dartmouth Conference, the birth of the new religion of the machine and disembodied AI and consequent tensions among the AI community. Hubert Dreyfus, respectable philosopher and admirable teacher, ascertained that the analogy brain–computer hardware and mind–computer software is a misleading assumption. The same holds true for the assumed discrete computation driven by algorithms on symbolic representations. Weizenbaum did not exclude the possibility of AI, but claimed that with larger and larger programs, more and more entangled, it becomes very difficult (if not impossible) to distinguish between desired outcomes and possible extremely consequential malfunctioning. Weizenbaum specifically associated decision-making with computation, but argued that the choice is a human capability, not within the possibilities of digital processing. Computers have no wisdom or compassion, which in his view (passionately ascertained) is part of human intelligence, together with emotions. Nadin says that both Dreyfus and Weizenbaum asserted that the brain is not a computer and intelligence is more than solving problems based on rules. In contrasting human reason to the power of computers, their arguments were about principles—and many 'insecure personages' feel intimidated by principles. The reason to remember Dreyfus and Weizenbaum is not to adjudicate victory for somebody, or even something. The architecture of thought deserves attention, not the incidental misreading (such as the impossibility of speech recognition that Dreyfus asserted, or Weizenbaum's anthropomorphic take on the degree of intelligence that the organism embodied by the computer might achieve). The major issue is not whether Dreyfus and Weizenbaum were right, but rather what the consequences of machine reductionism are. Instead of focusing on the new "records"—chess or Go, image identification, or "learning to learn", etc.—we can benefit from reconsidering their accomplishments that have inspired the scientific community.

Ignacy Sitnicki, in this volume, reflects upon Hubert Dreyfus's position concerning substantial difficulties in making strong AI effectively feasible. This position is grounded in the view that general intelligence is embodied and closely tied with human autonomous biological structure and consciousness. On the other hand, human intelligence in principle cannot be fully an object of symbolic representation primacy because in character it is also intuitive and instinctive, surpasses formal logic and information processing. Dreyfus rejected the possibility of making AI symbolic architecture equal to general human intelligence. In arguing that possibly a strong AI may be feasible in the future, a reservation should be made. If AI means the transfer of human intelligence onto a certain artificial autonomous system, it will not probably be the same general human intelligence, but rather a simulated form of general intelligence on a certain level of complexity. How similar is this intelligence—that is the question. Also, the next question may be raised as well: if similarity means that AI will never be entirely equal to human intelligence, but just nearly equal or that AI may surpass it—what may happen to humankind–AI relations if AI emerges as a dominant intelligent agent?

Hongladarom in this volume introduces us to Hubert Dreyfus's paper on "Anonymity versus Commitment: The Dangers of Education on the Internet" and its relevance to the field of philosophy and technology in education. In it, Dreyfus argues that the Internet has brought about a situation where the users lose their individuality and identity, and instead become faceless members of a large group where nobody knows anybody personally, known as "the Public". Drawing from the idea of Søren Kierkegaard, Dreyfus argues that the Public, mediated by the Internet, functions as an environment which suppresses awareness of one's own unique condition and individuality, a condition which is necessary for commitment. Translated to educational theory, this means that the effect of the Internet is such that it tends to promote people who think the same way, following wherever the crowd leads, rather than self-sufficient persons who strike out on their own. Thus, for Dreyfus the Internet would tend to promote students who shop around, sampling this or that course offering only to see, as consumers or spectators, what these courses look like rather than identifying themselves with one particular course of study and getting

to see these courses from the inside. Hongladarom notes, however, that Dreyfus in 2009 conceded many points where the Internet could actually provide benefits. However, his core idea, that unconditional commitment is required for genuine education, still remains. For Dreyfus, the Internet does not encourage commitment because it is anonymized, offering the individual the faceless Public where she does not have to be committed to anything at all. We see the relevance of Dreyfus argument in the era of 'post-truth' when he contends that the effect of the Internet is such that it could undermine unconditional commitment, but unconditional commitment would come to no avail at all if what is being committed to cannot be ascertained to be truthful.

David Casacuberta in using Dreyfus' legacy to understand justice in algorithm-based processes (this volume) draws on Dreyfus's phenomenological account of ethical expertise to establish the abilities that an algorithm should include to make ethical inferences. We need such criteria to avoid designing intelligent systems that appear to be prima facie fair but do not meet the standards of (human) experience-based moral competence. He says that finding a way to include some ethical expertise in computers is not only a relevant issue for future AI developments, such as autonomous driving or general artificial intelligence, but a pressing concern to revise and improve current software that helps humans make decisions. In citing chess or car driving as examples of expert behaviour, Dreyfus considers ethical comportment also to expertise and expect it to show a similar developmental pattern as reflected in the model. The beginner ethical expert would learn some principles and maxims and use them regardless of the context. This enables the learner to gradually move towards the highest stage in which rules and principles are left behind and ethical answers are intuitive, holistic and spontaneous. So a beginner would never lie as the maxim tells us, but an expert would lie to tell the truth according to the situation. Here Dreyfus takes an openly Aristotelian view: 'what is best is not evident except to the good man'. And what is good is learned in practice; one becomes an expert by being exposed and by experiencing the same types of situations as those endured by experts. Also, one must also be able to sense satisfaction or regret at the outcomes of one's action, a sense of regret or satisfaction that must be somehow shared or endorsed by other experts. Dreyfus' phenomenology of moral expertise has a particular emphasis placed on the linkage between knowledge and context. Dreyfus defends a rational, discursive base in our beginning as moral agents, when our ethical knowledge is linked to discursive thinking. Ethical rules are internalized in humans, and they arise depending on the context, and the relationship that the subject has with it. Moreover, according to Dreyfus, in familiar situations, however, problematic, the expert contextual intuitive response is superior to judgment based on detached, abstract rules and principles. According to Dreyfus, ethical frameworks of autonomous cars that rely on a formalization of the trolley problem are not enough, as they are purely declarative systems, rather than based on pre-reflective rules that enable us to make sense of the surroundings. What is needed is a framework in which there is a big qualitative difference between losing a human life and damage to a car, and not just a rule-based system in which the price of repairing a Tesla might weight more than the insurance to pay for hitting a small child. Thus, a pure machine learning approach would not work either. A neural network can be used to spot correlations and find patterns that are relevant to decide whether a person is going to return a credit card or whether he is going to re-offend. However, such a protocol does not assure us that the system is going to be fair. Casacuberta says that we are not saying that the algorithms are biased per se, we just want to point out that existing biases in society, e.g. racial ones, might end up producing unintended unfair results. If a discrimination is already happening in a society, the system will just include it in its pattern recognition system. An artificial expert in ethics that meets Dreyfus requirements needs a pre-reflective system that has some autopoiesis, the ability to make sense of the surroundings and generate context-based judgements of the ethical implications of a situation, that are not just rule based. Casacuberta concludes that to have algorithms capable of making fair decisions, better algorithms and learning environments are needed, so they are able to generate process akin to general intelligence, and not just pattern matching from past examples. Moreover, in a sense, they need to know the difference between what it is now, and what it should be, which can never be deduced from past examples. This needs human experts to scrutinize such deductions. Dreyfus has pointed out that in addition to experts scrutinizing the biases of algorithms, we also need common citizens that master the nondeclarative, pre-reflective ultimate nature of ethical judgements, and give that extra dimension to assure that we are having not only accurate software but also a fair one.

Jeff White in his paper, Dreyfus on the "Fringe" in this volume, reflects on the evolution of Dreyfus argument on symbolic processing. This has been developed over half a century, towards an emphasis on the difference between situated human cognition and the "minded" sense of symbolic reasoning. Dreyfus allows that any problem can in principle be solved by a digital computer "provided the data and the rules of transformation are explicit" but this requires that the problem be structured for the machine beforehand in order to avoid an infinite counting out of possibly salient dimensions and to an "infinite regress" of higher and higher levels of operations required to judge preceding results. Here again on the fringes of consciousness, Dreyfus suggests that "only the uniquely human form of information processing which uses the indeterminate sense of dissatisfaction" avoids

this regress. At the same time, the human ability to "retain this infinity of facts on the fringes of consciousness, being "a uniquely human form of information processing not amenable to mechanical search techniques", allows human beings access to the open-ended information characteristic of everyday experience" without requiring that this inexhaustible list be made explicit. Without fringe consciousness and insight into intended meaning, for example, language programs must explicitly and exhaustively define terms relative to other terms, but even this requires that the language user be able to tolerate ambiguity in the use of terms even as their definitions are refined. Without insight into fringe consciousness and tolerance for ambiguity until intentions are resolved, computation continues perhaps ad infinitum as explicit meanings are fixed and language—as well any other form of action—becomes practically impossible. Meanwhile, the ambiguity-tolerant human being avoids such endless loops as he/she is able to, in the context of gameplay. The crux of this argument rests in the nature of fringe consciousness, that it is situated, and that natural language is context dependent not due to the nature of language, or to the laws of logic, but due to the nature of the language user and logic programmer, him/herself. The problematic which Dreyfus zeroed-in on more than five decades ago, and with which he wrestled for the rest of his life, ultimately comes down to convincing enthusiasts of AI to redirect their efforts in the short term, until understanding of the human condition meets with technologies up to expressing this understanding in an artificial medium. Those determined otherwise he likened to "alchemists" who, having refined "quicksilver" from dirt, worked centuries from the same methods hoping for gold. For 52 further years after this earliest assay, Dreyfus consistently found his contemporaries vainly working to transmute one type of substance to something essentially different, only to muddle on in blind optimism without the insight to restructure the problem while ignoring his message. "Similarly, the person who is hypnotized by the moon and is inching up those last branches toward the top of the tree would consider it reactionary of someone to shake the tree and yell 'Come down!'" Regardless, his mission continued, and at this point we may ask if we are in similar position, today? There is no sense in engineering artificial compliments to human intelligence when this natural intelligence is bent only on mutually assured self-destruction by way of the medium. So in the end, how might Hubert Dreyfus advise current researchers in AI? We should employ technologies to enhance access to fringe consciousness, so that we can mine culture and tradition for meaningful directions forward. We should do what we do best and allow machines to do what they do best. We may also benefit from Dreyfus' observation that the salvation of what he called "Heideggerian AI" may come from the deeper appreciation of Heideggerian philosophy, specifically that of authenticity as applied not only to model architectures, but to the researchers themselves who must first understand their authentic condition before articulating similar in artefacts.

Cathrine Hasse in 'Post Human Learning' (this volume) argues that the human in the Dreyfus's stage model is not a rational, symbol-processing machine-like creature. It is not rational in the sense as rationality as detached, brought to bear on practical predicaments from a standpoint other than one of immersion in them. It is a phenomenal sensing human, that is, a being-in-the-world. It is a human that initially learns rule-based, like a computer, but soon turns into an embodied learner, whereby learning increasingly stems from situated experiences. Hubert Dreyfus was fascinated by human knowledge and perception as emerging from body–world relations. Dreyfus used MIT work on artificial intelligence to nail down what a human was not: a human was not learning, knowing and perceiving like a machine. Whilst artificial intelligence machines operated according to symbolic rules to find solutions to purely formal tasks—human intelligence, Dreyfus argued, is embodied and situated. The human in the Dreyfus model, unlike machines, can exist in a vague world such as the world described by phenomenology as well as post-phenomenology. Humans can discriminate between a rising and falling hum of a passing factory while driving and a car engine that slowly revs up and down. In such ways, humans differ from machines in terms of their abilities to distinguish the essential from the inessential features of a particular instance of a pattern; use cues which remain on the fringes of consciousness; take account of the context; and perceive. Moving to the debate on the posthuman learning perspective, we see technology as not just a mediator, but that takes part in the co-creation of a collectively shared socio-cultural material world. Even seemingly abstract symbols are in this view materials entangled in phenomena that include the prior learning in the body of the driver as well as a whole community of drivers. With his emphasis on learning as cultural, collective word-meaning in a material world, Vygotsky goes much deeper than perceiving learning as the'useful changes in the workings of our minds'. Learning is not just a change of 'mind', but of a somewhat collective bodily becoming of a material world. This acknowledgement makes it possible to sustain Hubert Dreyfus's claim that AI will never succeed in making truly thinking machines.

Sjoukje van der Meulen and Max Bruinsma in Man as 'Aggregate of Data' (this volume) make the point that although Dreyfus' categorical distinctions between man and machine are still relevant today, their relation has become more complex in our increasingly data-driven society. We humans are continuously immersed within a technological universe. In the spirit of Dreyfus' systematic critique in *What Computers Can't do*', the authors argue that within the ever-expanding data-sphere of the twenty-first century, a new

concept of man as 'aggregate of data' has emerged, which further erodes and undermines the categorical distinction between man and machine. This raises political and ethical questions beyond the confines of technology and artificial intelligence. Moreover, this seemingly never-ending debate on what computers should (or should not) do provokes the philosophical necessity to once again define the concept of what it is to be 'human'. The root cause of the error that most AI technophiles make, according to Dreyfus, is a lack of understanding of the fundamentally different nature of man and machine, which leads to serious moral issues concerning humanity, and is thus the reason why he insisted until his death in April 2017 that we should reflect on the concept of man more deeply and philosophically in the current age. The authors argue that something essential has changed about our concept of man in the information age: a new type or 'part' of man, which functions and acts at the nexus of his physical and immaterial manifestations—deeply affecting both—needs to be taken into consideration: 'Man as aggregate of data'. They further argue that in today's information and network society, the individual is also compiled of (data) components which characterize him and which are in a sense interchangeable for other components. These are his data, information of all sorts, which together result in a more or less unique profile. The nagging question is whether, in today's network society, one can still conceive of an authentic Self in the classical sense of 'knowing thyself'. How can this new type of individual or 'dividual' in Deleuze's terms, that is, primarily known (and knows himself) as aggregate of data and functions, still be called authentic? And what happens to the autonomous being who does not wish to subordinate himself to what others, including computers and robots, expect him to do? In other words, is a hybrid human–data aggregate still a self-aware individual with the capacity of taking responsibility for his own thoughts and actions or are his mind and behaviour more and more conditioned and controlled by algorithms and software codes. Remarkably, say the authors, the idea of the Quantified Self hinges on data (immaterial information, that is) generated by our physical bodies interacting with our techno-material environment. Far from forgetting that our being in the world is still very corporeal, the Quantified Self integrates our 'flesh' within what Stephen Humphreys and others have called 'datasphere', which comprises every aspect of our lives, both private and Public. In the 'datasphere', our bodies virtually dissolve with any other representation of our existence, physical or abstract. This is not to say that we are becoming less carnal, but that the old dichotomy of body and mind is being radically redefined—bodies too become amendable in a conceptually different way than the ancient *mens sana in corpore sanum* suggested. The concept of man as aggregate of data, in short, does not make him less physical—it shifts the focus from the "flesh" to the ways in which our bodily movements and actions, including our most private ones, "take on materiality as they become artefacts of the datasphere".

Simon Penny in the Enactive Performative perspectives on Cognition and the Arts (this volume) recalls his discovery of Dreyfus's '*What Computers Can't Do'* as he saw unfolding of the crisis in AI often called the *common sense problem*. Dreyfus seemed to have put his finger so adroitly on many of the misgivings of a philosophical neophyte, that his writings confirmed that the felt disquiet was not simply that of a technical newbie overawed by a triumphant and sophisticated technology. Dreyfus' phenomenological account provided a theoretical framework for these concerns gave them the structure of an argument and an inquiry into the historical and critical study of AI and cognitive science. Penny notes that the early 1990s collapse of the GOFAI paradigm around the common sense/framing/symbol grounding problem had left a younger generation of AI researchers scrambling. In this context, two developments arose, which are, to Penny, inextricably linked—Artificial Life and *post-cognitivist* theories of cognition—embodied, distributed, enactive, situated and the rest. Penny says that we are at a historical moment when the traditional valorisation of abstraction is being brought into question by leaders in cognitive science and other fields. This has great relevance for the explication of cultural practices. Hubert Dreyfus led this charge, reminding computer scientists that the special qualities of human intelligence are a result of having a history of human embodiment, and that such a history of embodiment builds the brains, minds and intelligence we have. His arguments were not so much rejected as found simply incomprehensible by an AI community locked into a paradigm in which dualism was axiomatic. The very acceptance of the notion that there exist separate and complementary entities we can refer to as *mind* and *body* (in Descartes' terms, the *res cogitans and res extensa*) prevents us from understanding holistically, the intelligence inherent in the behaviours of whole persons integrated into environments, structured and unstructured, articulated by tools, procedures and interpersonal interactions. The conventional internalist conceptions of cognition can say little which is useful about the kinds of *sensorimotor integration* which are fundamental to action in the world. To transcend the limit of constrained cognition, we turn to practices of the arts as they epitomize and refine these sensorimotor intelligence to a high degree. The reluctance of *conventional cognitive science to* embrace arts has hampered useful discussion of cognition and the arts for much of the last century. We need to recognize that along with the big Cartesian bogeyman come related ideas concerning the nature of thought as "reasoning" on "representations". We are naturalized to such ideas, even if on particular occasions we adopt postures which are contrary to them, and even if this dualism sits uncomfortably with

the neurological materialist idea that cognition occurs in the brain. Penny further remarks on the uncritical reflection and adherence of many researchers to the orthodox idea that cognition occurs largely or exclusively in the brain and that the brain is a kind of computer. This assumption is nothing but the result of proliferation of AI functionalism which is perpetuated in popular culture long after their demise in AI theory itself. For Penny, this renewed currency of dualist, functionalist ideas is in large part due to the infiltration of digital computing into diverse aspects of human culture. Computing, as our paradigmatic technology, became the main source of metaphors for human cognition. Yet in day-to-day life we are presented with very different experiences of cognition as it is lived. Strangely, we seem to be content with a philosophical explanation that is at odds with our lived experience. Yet, cognition includes experience; it is being *in the world*. There is no cognition except for current experiences in the world or reference to past experiences in the world. You cannot reason about Plato's cave without having had situated, embodied experiences of windows and shadows which make the metaphor meaningful. But cognition is also *doing* in the world. Many of our behaviours we call cognitive in the narrow and conventional sense are facilitated by, or cannot occur without, physical action in association with artefacts and tools. In this sense, cognition is not only embodied but also *embedded* and *enactive*. From this base, one could lay out a new account of cultural action. This would, Penny asserts, entail a reconceptualization of conscious/non-conscious thought/action; a reconceptualization of "nature and nurture" through the idea of cultural bootstrapping of latent capacity/neural exploitation; and a reconceptualization of cognition as embodied, enactive and integrated with the material and cultural world. Penny opines that the current revolution in cognitive science provides a basis for a paradigm shift which will allow new ways of speaking about embodied, materially engaged action. Such an approach holds the potential to level the (academic) playing field that has for so long been tilted in terms of the abstract and the symbolic. It has the potential to provide an entirely new register in which to speak about what we might call *cultural cognition*—embodied art and cultural practices in a new way that gives full recognition to the materially, socially and spatially situated intelligence involved in human cultural activities, both 'high' and 'low'.

AI&Society warmly welcomes reflective contributions to the debates on Calculation-Judgement—Calculation and the phenomenology of embodied skill, exploring their relevance and impact on AI narratives in the pursuit of seeking harmonious interactivity of art, science, technology and society.

References

Collins H (2018) Artifictional intelligence. Polity Press, Cambridge
Cooley MJ (1987) Architect or bee? Hogarth Press, London
Dreyfus HL (1978) What computers can't do: the limits of artificial intelligence. HarperCollins; Revised, Subsequent edition (1 Jun. 1978)
Dreyfus HL (1988) The Socratic and platonic basis of cognitivism, AI&Society. Springer, Berlin
Gill KS (ed) (1986) Artificial intelligence for society. Wiley, Chichester
Gill SP (2015) Tacit engagement: beyond interaction. Springer, Berlin
Gill S (2017) Uncommon voices of AI, AI&Society. Springer, Berlin. https://doi.org/10.1007/s00146-017-0755-y
Weizenbaum J (1976) Computer power and human reason: from judgment to calculation. W. H. Freeman, Francisco