AI & Soc (2018) 33:637–640
https://doi.org/10.1007/s00146-017-0770-z

Reflections on James Bond of AI

Urjit A. Yajnik[1]

Received: 3 September 2017 / Accepted: 14 October 2017 / Published online: 24 October 2017

1 Introduction

What do we make of Polanyi in the age of AI and robotics? The main thesis I want to advance in this article is that much human knowledge is actually in a form that is not explicit, yet is not as inarticulable as to be classified as Tacit.

A new dimension to the discourse on Epistemology started in the twentieth century with Polanyi's elaboration of Tacit knowledge (Tacit Knowledge 2017; Knowledge How 2017; Tacit Knowledge: Making it Explicit 2017). Whereas Epistemology per se attempts to delineate what is knowledge and what can be considered to be valid knowledge, etc., Polanyi raises the operational aspect of knowledge. Much knowledge resides in forms that are not articulated and the presence of such knowledge in an agent is expressed by the capability of the agent to execute certain tasks. Polanyi's criterion has been put to practical use in several areas, from the content of education (Chugh 2015) to the question of inter-cultural dialogue (Loenhoff 2011) and also independently developed to highlight the subtleties that arise when deploying digital technology to serve as an interface between human beings (Gill 2015).

In this article, we shall be dealing with knowledge essentially in its operational form and any response to the vast metaphysical discourse on it is beyond the scope of this article (and the ken of its author). I thank the editors for encouraging me to articulate some opinions expressed in personal correspondence and for accepting them in an article form.

2 Tacit knowledge in the context of machines

I begin by pointing out that Polanyi's criterion when adopted in the age of robots and Artificial Intelligence (abbreviated AI, used also for Artificially Intelligent), raises an interesting dualism. Let us consider the Japanese industry example reported in Nonaka and Takeuchi (1995) where a particular kind of bread making, resident as Tacit knowledge with an accomplished baker was eventually transferred to a machine. Two interesting points arise. First, the machine is an unconscious entity. Therefore, by definition all its "knowledge" must be Tacit. Take for instance the machine's ability to articulate "I see a table as an obstacle in my path". If the same sentence is spoken by a human being, this would constitute explicit knowledge, articulating the presence of a material object and stating the learned rule about the impossibility of navigating through it. But the robot has been fitted with sensors programmed with firmware that analyses perception as data and what it "does" is a consequence of programmed instructions. For the robot, an unconscious entity, the act of turning away to avoid the obstacle is a "Tacit" skill. As another example, when a human being "knows" that carbon monoxide is odourless but poisonous and must be detected for safety in an enclosed room, it is explicit knowledge. A robot will only carry a programmatic tree which when it arrives at the sensors reporting the presence of carbon monoxide will access the database on the properties of the gas and will alert its human partner. The unconscious manner in which this will be executed should classify as Tacit. Perhaps, therefore, it is redundant to employ this classification in the case of a machine, however learned.

But a point of divergence arises when we get to AI machines. Let us sharpen Polanyi's observations by adding that the Tacit is knowledge to the extent that the agent

✉ Urjit A. Yajnik
yajnik@phy.iitb.ac.in

[1] Department of Physics, Indian Institute of Technology Bombay, Mumbai 400076, India

carrying such knowledge is capable of modifying it on demand (improvising) and also capable of enhancing the quality (improving) of the existing recipe. These two are in a sense related, with the latter becoming a special case of the former, but it is worth keeping the distinction for practical reasons. Thus, a dessert maker may be able substitute cherries judiciously for strawberries, but also, be able to improve upon the taste of the Strawberry dessert over time. Now a traditional machine's limitation is exposed as an "unknowing agent" because it can neither improvise nor improve. We may categorise this as a case of Tacit knowledge that is Static. The capacity of improvising and improving may be considered to be Dynamic Tacit knowledge and could be considered to be an essential criterion for an AI agent.

However, an AI agent by design will be one that at least improves upon itself, or more technically speaking, may be programmed to optimise its operations over certain objectives by learning on the job. Now it will be difficult to distinguish between an AI machine and a human being, and in Turing's sense we will have to accord a pseudo-consciousness to the former and consider some of its knowledge Explicit and some Tacit. Thus, the capacity of improvising and improving the deployment of a process that makes Tacit knowledge dynamic could be considered to be an essential criterion for an AI agent. To the extent reported in deep learning literature (Silver et al. 2016), we are already at this stage of technological development.

Another observation I wish to record concerns a point well known in the literature, namely Polanyi's classification runs into difficulties when it comes to classifying the language capability. Indeed, here the human, admittedly an intelligent agent, is incapable of expressing how it expresses itself! So it must be treated as for ever Tacit. Yet the process of expressing this fact in words is positing the agent's knowledge of it in an explicit form. In my opinion the naive solution to this paradox is simply that Polanyi classification need not push to this self-referential level. But the question is bound to arise eventually in the context of an AI agent's skill set. The classification can be further refined or qualified when the practical situations arise that demand such skills.

3 The category of implicit

The main thesis I want to advance in this article is that much human knowledge is actually in a form that is not Explicit, yet is not as inarticulable as to be classified as Tacit. We can return to our example of the dessert maker. This person may well have a webpage advertising his or her business where they state their "vision and mission". This would be factual to some extent, say, about the quality of ingredients, etc., but also largely be aspirational statements about taste, customer satisfaction, carrying forward culture specific tradition, etc. Surely, we know that those few sentences are never sufficient to capture what the agent really thinks or feels about those issues. They are only succinct capture of easily communicable features of the person's ethos of operation. A simple test of this is that if you personally query any of those points further, there will be profuse responses.

The notion of Implicit is closely tied to the shared values of a community and so we briefly consider an analogy with an organisation. Consider the example of a CEO who has pulled a miracle with an enterprise; it is well known that the person's ethic is often expressed as one or a few mottos, yet innumerable gatherings and functions continue to invite him or her to really hear the person to get it all "from the horse's mouth". If an organisation can be treated as an entity, the enterprise shares with us an Explicit knowledge of the sate of its finances through its balance sheets. Yet the organisation knows only "in its veins" what happened when its profits went from red to blue. The miracle that happened would be Tacit knowledge of the organisation. I am not proposing that we treat an organisation as an entity but this is a useful notional point to highlight that by comparison, the energy that the CEO unleashed would be Implicit knowledge of all the major stake holders of the organisation, managing all the subtle points that turned the organisation around. They would not have an instruction set handed to them by the CEO, but only key strategies, important guidelines, but most importantly an ethos and a set of core values.

Thus, we may say that the category of Implicit does not suffer from being articulated altogether, yet it carries with it much more than can be articulated in any strictly limited discourse. This is partly because ethos is couched in a language referring to very general and broad classes of entities or actions. It also often contains metaphors. For instance, the statement "To be a fountainhead of knowledge and a hub of innovation" would be considered a well stated ethos by an R&D organisation. It will need to be referred to in every challenging situation and interpreted afresh at every moment and in every case by its technological staff as well as its administrators.

There is an obvious subjectivity to this kind of "knowledge". For instance, if the ethos demands valuing life, then instances will arise when one needs to distinguish between the life of human beings versus the life of animals; or in many conservation projects it becomes the question of the life of a selected species versus the life of the more abundant species. The working rules that people will derive from the broad ethos are based on a shared set of values. Often these are taken as self-evident truths, often they reside in taboos and so on. These may not be, and in most cases are not "scientific" facts. In this sense it does differ substantially from explicit knowledge that can dictate specific action, for example the simple instruction, "a piece of furniture is an obstacle and must be avoided". By comparison, Implicit may not be as clear cut but would certainly act as a guideline, for example, "living creatures feel pain and must be spared collision" is easy to interpret when encountering a human or a house pet, but needs to be overridden if the robot encounters an ant in its path. In practice, value knowledge or the Implicit can be very vague due to the generalities in it, yet the guidance it generates has a determining power in our actions that may be far stronger than we suspect.

In Epistemology this category of knowledge has been mooted since the time of Socrates, but most recently Gettier's observations have been construed to mean that "value knowledge", often the kinds of knowledge that are "justified" or argued to be correct by implication, etc., suffer from drawbacks that prevent their acceptance as knowledge (Epistemology 2017). Our purpose here is to categorise knowledge that is operationalised, the knowledge that directs action or that acts as "guide-to-action". For this purpose, we adopt a pragmatic attitude, namely if a community or a networked system of peer agents perceive certain enunciations to be knowledge then it is knowledge. This is because even if the specific statement they hold as true has exceptions and likely to be wrong, the agent is limited to rely only on that statement. We shall pursue the main theme of the article with this assumption.

4 Towards an AI James Bond

A lot of science fiction has already conceived of the variety of problems that will arise with future machines. Two famous examples relevant to the present discussion are HAL of 2001: A Space Odyssey and Marvin of A Hitchhiker's Guide to the Galax*y*. HAL, due to the heuristic algorithms that constitute it, is led to simulate behaviour similar to that guided by guilt, denial and craving for power. Faced with the threat that humans are doubting its capability, it decides to eliminate humans to ensure optimisation of the main objective the mission. The adorable Marvin, on the other hand, is our own alter ego in a world increasingly pushing and driving us to be evermore efficient and productive. He demonstrates the opposite problem, viz., it is perpetually depressed. This is because its mental capacity is much too big for it to be excited about the mundane tasks he is engaged in. It taunts and challenges its human masters and clients with sarcasm and cynicism, almost to the brink of non-performance.

We may expect all of these and more complex issues to arise when agents get increasingly intelligent. Referring to our section II, we are now forced to reassess the Explicit versus Tacit for such an agent seriously, as it has a mind of its own and it has learnt "ways of doing" through Deep Learning that it cannot convey to us in words or in so many lines of program code. Indeed, along the lines of HAL, it may even choose to be secretive about these new learned techniques and pretend that they are its Tacit knowledge because an objective function given to it to optimise may suggest, as per its own encoded analysis, that this is the correct approach to adopt.

But equally importantly, we now wish to emphasise that for a variety of applications we will need to empower the AI agent with Implicit knowledge and with a capacity to operationalise such knowledge. The strapping hero James Bond, along with his charming and coveted characteristics of personality also carries a very important permission, the license to kill. For instance, if an AI agent is to be deployed in a disaster-ridden area, it may have to fold in some ethical reasoning when deciding on unexpected situations. Even more immediately, a driverless car, faced with the realisation that it is skidding may have two options, to veer left or veer right. On one side may be a vulnerable small car with two human beings and on the other side an SUV with several people including children, but more robust against accidents. Of course the wise programmer will have terminated the mission itself in advance based on an information about the road conditions. But considering human beings take calculated risks, so will it be mooted to allow to a robot to take risk.

The sleuth analogy can occur in many settings. One may have an AI agent monitoring data traffic for mala fide activity. An unexpected stream of flow can result from mala fide intentions or from a desperate situation that has generated a garbled data flow arising as an alarm or SOS signal. It may need to take a spot decision on terminating the stream or setting aside other tasks to focus on it. How is it going to decide? If a human agent monitoring the same situation can instinctively decipher some meaning in a garbled data stream he or she may initiate a completely different kind of action. Likewise, an AI agent acting as a security guard in a war zone may need to interpret every possible nuance of facial expressions and small ticks of the limbs to decide what to do. We may think of this as the problem of programming a James Bond into an AI agent. And the upshot of this submission is that Mr. Or Ms. M sitting in their MI6 office may be less strained on the Implicit front, in the sense of having more access to cold logic, explicit directions, and more peer or superior consultation than James Bond who is deployed alone in the field, with a license to kill.

We submit that eventually as we rely on AI agents they will have to be programmed with Implicit knowledge, the kind that we may loosely call "wisdom", more specifically a set of ethical tenets with a capacity to derive from them a guide to action. Even with the self-driving cars this seems imminent, and one may guess that until some level of maturity is reached, the owners of such gizmos will exert restraint in deploying them in situations where unexpected ambiguities arise.

It may be noted further that the ethics will also impact the Deep Learning itself and the sharing of it. From many practical considerations, the makers of AI should want an AI agent to be capable of articulating its newly devised techniques. This would be the technical capability of conversion of its dynamically generated Tacit knowledge into Explicit. But equally importantly, the makers should build in Implicit knowledge to be aligned in such a way that the agent will not renege on sharing the Explicit version of its new techniques with its peers or its master.

5 Conclusion

This article is based on the assumption that categories of knowledge should be analysed with operationalisation of such knowledge as a criterion. The considerations of Polanyi regarding Tacit knowledge have become increasingly relevant to the important discussion of transition from robots to AI agents. And it is submitted that Polanyi's criteria may need further elaboration as entities intermediate between statically programmed robot at one end and human-like AI agents on the other end begin to be designed. Polanyi's classification has also generated some debate when applying it to language capability of human agents. For the purpose of this article, that seems to not be a problem. However, a classification of the Tacit into static versus dynamic would be useful as we move towards intelligent agents.

The main point to be made is that beyond Polanyi, and notwithstanding the erudite discussion in Epistemology, a category intermediate between explicit and Tacit arises, which is here called Implicit. This category is in the nature of ethics or values or moral decisions, and admits articulation to a substantial extent, that it can be conveyed verbally. Yet it lacks the precision or crispness of the Explicit. On the other hand, but by contrast, due to the deep values it carries, it holds even greater sway on the agent empowered to interpret it than the explicit. Due to this importance, it makes sense to enable our future AI agents with that category of knowledge as well.

While the article was being given the final form, I became aware of the "Asilomar AI Principles" (Asilomar AI Principles 2017) put forth at a conference in 2017. The proposals of this article can be seen to be in accord with the items "value alignment" and "human values" under the category ethics and values, and bears on all the issues flagged under long-term issues in their published document.

Curmudgeon Corner Curmudgeon Corner is a short opinionated column on trends in technology, arts, science and society, commenting on issues of concern to the research community and wider society. Whilst the drive for super-human intelligence promotes potential benefits to wider society, it also raises deep concerns of existential risk, thereby highlighting the need for an ongoing conversation between technology and society. At the core of Curmudgeon concern is the question: What is it to be human in the age of the AI machine? -Editor.

References

Asilomar AI Principles (2017) Asilomar Conference. http://futureoflife.org/ai-principles/. Accessed 8 Sept 2017

Chugh R (2015) Do Australian Universities encourage Tacit Knowledge transfer? In: Proceedings of the 7th international joint conference on knowledge discovery, knowledge engineering and knowledge management, pp 128–135

Epistemology (2017) Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Epistemology. Accessed 8 Sept 2017

Gill S (2015) Tacit engagement: beyond interaction. Springer, Cham

Knowledge How (2017) Stanford Encyclopedia of Philosophy http://plato.stanford.edu/entries/knowledge-how/. Accessed 8 Sept 2017

Loenhoff J (2011) Tacit knowledge in intercultural communication. Intercultural Commun Stud XX:1

Nonaka I, Takeuchi H (1995) The knowledge creating company: how Japanese companies create the dynamics of innovation. Oxford University Press, New York

Silver D et al (2016) Mastering the game of Go with deep neural networks and treesearch. Nature 529:484

Tacit knowledge: making it explicit (2017) London School of Economics online notes http://www.lse.ac.uk/economicHistory/Research/facts/tacit.pdf. Accessed 8 Sept 2017

Tacit Knowledge (2017) Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Tacit_knowledge. Accessed 8 Sept 2017