Machines Can't Think (Or Make Decisions)!

"Is it possible for a machine to think?...the trouble which is expressed in this question is not really that we don't yet know a machine which could do the job. The question is not analogous to that which someone might have asked a hundred years ago: `Can a machine liquify gas? The trouble is rather that the sentence, `A machine thinks (perceives, wishes)' seems somehow nonsensical. It is a though we had asked `has the number 3 a colour?'"

Ludwig Wittgenstein (Blue Book)

In his 1978 textbook pioneer computer scientist Richard Bellman proposed the following definition of Artificial Intelligence; "The automation of activities that we associate with human thinking, activities such as decision-making, problem solving, learning ...". This and several similar definitions that followed are interesting in that they define AI indirectly in terms of the type of problems it can solve, rather than directly by characterising the type of `thing' it is. In other words, the term AI simply refers to those algorithms that can solve problems that humans deem as needing `thinking' to solve, because they find them difficult and requiring significant conscious cognitive effort. From this perspective there need not be any objective characteristics of AI algorithms that distinguish them from non-AI algorithms. There is nothing about the type of optimisation used in neural networks that singles them out as being AI. They are AI because they can classify objects in images or translate text from English to Spanish. On the other hand, one might perhaps ask if there are objective criterion for characterising the set of problems that humans deem to require thinking? I believe that there are reasons to suppose that this is not the case.

Whether a problem requires "thinking" to solve it, seems to have little to do with the difficulty of the problem itself but more to do with the human propensity to be able to solve it in a particular manner. Singled out for this category are those that require slow and deliberate conscious cognitive effort. The emphasis on *conscious effort* is particularly important. For example, catching a ball is not viewed as being intelligent but calculating the trajectory of a space rocket is. However, both require the application of similar physical principles, it is just that in the first case they are applied, or at least approximated, subconsciously. Furthermore, the class of problems that are hard for humans to solve in this way is different from those that are hard for a (Turing) machine to solve and vice versa. For example, consider mental arithmetic. Alexa can quickly factorize relatively large numbers into their prime factors. If a person could do this then they might be view as being very intelligent, but it is not particularly impressive for a computer. It needs an extensive list of known prime numbers and a lot of computing power, but then the algorithm is simple. Terms like "thinking", "intelligence" and "decision making" are very human centric and certainly one wonders how useful they are when applied to machines. However, I want to go further and argue that they in fact lose their meaning when applied outside of a human context. Machines can't think or make decisions because thinking and decision making are peculiarly human activities inherently linked with human conscious experience.

Thinking does not simply refer to the process of problem solving but to the conscious experience of what it is like for human beings to attempt to solve problems in a certain way. Thinking requires the conscious sensations of deliberation, puzzlement, inspiration etc. The term "decision making" makes similar implicit reference to conscious mental experience. To decide is not simply to choose, but to consciously choose; it requires us to weigh up alternatives, to imagine the impact of our actions on others, and to know that we have done so. We choose conscious of the fact the we could have chosen differently. There are of course no such conscious experiences involved when autonomous systems select actions, so terms like "decision making" just don't apply. Machines do not decide or

choose, they apply functions. Some inputs are received, e.g. from sensor measurements and internal monitoring, and a policy function is applied to these inputs which then selects an action. This is what "choosing" is for a machine and it involves none of the conscious aspects that we associate with human decision making. There is no weighing up of options just prior to deciding. There are no ethical considerations entertained at that moment. All such things, if they are to be considered at all, must be built into the policy function itself. For machines, ethics is a "design time" issue, not a "run time" one.

To be clear, I am not suggesting that there is anything special about conscious experiences that means computers are incapable of having them. I am not advocating some form of panpsychism in which brains are made of the kind of matter that can exhibit consciousness whereas computers are not. Even less am I adopting a dualist position in which conscious thought is separate from the mere computation of subconscious thought. It may well turn out to be that consciousness is an emergent effect arising from running computational processes on a biological computer with a particular architecture, but if so, it is a real emergent effect, nonetheless. We really experience consciousness and terms like "thinking" and "decision making" implicitly refer to this experience. Of course, it may one day, in the far future, be the case that machines come to experience a form of consciousness also as a result of emergence, but if this happens there is no reason to suppose that it would be anything like human consciousness and hence words linked to human cognition would still not apply to it. Just as for Thomas Nagel we cannot know what it is like to a bat, so we will never know what it is like to be a conscious machine.

Does any of this matter? We have borrowed words that apply to humans and applied them to machines because aspects of their meaning are still relevant. Autonomous systems still choose even if they don't consciously choose. Surely this is something that happens all the time; the uses of words are adapted to new scenarios and this is part of the power of language. For example, planes fly in a very different way to that in which our ancestors would have observed birds flying but extending the word "fly" from the natural to the artificial has not caused us serious problems. In many respects this is also true for cognitive words. When designing AI algorithms, it is often convenient to refer to the system as "making a rational decision". However, I want to argue that when we come to consider the issues around ethics and trust for autonomous systems then our borrowing of words leads to confusion. This is because human conscious experience is central to our understanding of ethics, trust, and even law. To judge an action right or wrong implies conscious agency. The agent is required to consciously reflect on what they are about to do, to consciously consider the impact of the different possible actions as viewed through the lens of ethics, be influenced by feelings of compassion and empathy, and then choose according. If their action causes harm they can be judge because they are deemed to be capable of conscious choice. If we trust people it is because we believe that they will consciously choose, to the best of their ability, to do the right thing. If we trust an institution it because we trust those that built it and wrote its constitution. If we trust a law, it is because we trust those that composed it and the institutions to which they belong. If we trust a machine it is because we trust its designers and engineers. In this way trust can be inherited but its source must always trace back to the human experience of choice.

Our use of terms like decision making or autonomy to apply to machines causes confusion because it misleads us into thinking that the source of trust (or mistrust) or the root of responsibility can be the machines themselves. We then worry about whether machines will always make ethical decisions. Instead, we need to worry about how to design and build them in such a way to ensure that their decision making is ethical. The two things are different in that for the latter the source of trust and responsibility continues to lie with human beings and their institutions. However, herein lies the real

difficulty. We are not really talking about deploying autonomous systems, we are envisaging a new type of automation. In this new wave we will be automating processes that have until now required significant conscious human involvement. This means that the chain of inherent responsibility linking back to human judgement is a short one. Responsibility for an accident is often quickly traced to the drivers involved. If such processes are automated then there may be many more links in this chain, making it difficult to trace ultimate responsibility. Perhaps a chain here is even the wrong image. We have an event, an accident, and lines of responsibility running backwards in time like a tree. If a driverless car causes a fatality who is responsible? Is it the car owner or the factory where it was built? Maybe there is a bug that can be traced back to a particular software engineer? Or is it an inherent design flaw for which a whole design team are responsibility? But then what about the independent quality control processes or regulatory checks, have they also failed and if so, who is responsible? In the end it will be up to society acting through laws and regulation to decide how such branching trees of responsibility should be traversed.

The problem of trust is made more difficult because of the type of tasks which we are now attempting to automate. They tend to be carried out in a complex and unbounded environment, making it almost impossible to exhaustively verify the behaviour of the system. Indeed, this is one of the reasons why up to now they required significant human involvement or at least supervision. Consider the task of driving through a large city centre for which the space of possible scenarios that might be encountered is huge. This makes it even more difficult to trace the lines of responsibility since inevitably machines will behave in unexpected ways when presented with unforeseen circumstances. Who can be held responsible for failing to see the unforeseeable? Nonetheless, the decision to automate is ours and we must try to understand and clarify the complex sources of trust and responsibility on which we base this decision. Remembering that machines don’t really think or make decisions should help us to better focus on where these sources can and should lie.