**Algorithms –The Furies of the age**

This session is about now and what may happen in the future but first we will start with a bit of past history.

For those of you haven’t read the Wikipedia entry a little history. The word algorithm comes from the name Al-Khwarizmi a 12th century mathematician astronomer in Baghdad. Nothing like it you say until you hear the latinised version of his name *algorithmus*. He was interested in rules and that is what an algorithm is.

Let’s start with really simple algorithms which was all there was until about 5 years ago. An algorithm had an objective and was subject to certain limitations called constraints. They pottered along quietly for many centuries just in maths until computers came along. Computers need rules so algorithms became the word that was at the heart of software. Thing didn’t change that much in the first 50 years of their life but by then computers were becoming much more powerful and much smaller. People realized having a single objective didn’t reflect the real world so they tried to link a number of algorithms together to achieve a particular multiple objective.

The algorithms we know are mostly in the simple algorithm category. Just to run through a few. Probably the most well-known is the one from University College. Neil Ferguson warned the government that if they didn’t have an immediate lock-down then could expect about 500,000 Britons to die. So Johnson changed his mind and is now said to be furious about it. There are obviously many other reasons to have a lock-down particularly as most countries were doing it but the government needed a reputable scientist to blame. The algorithm was based on epidemiological data and it showed what could happen under certain conditions. It didn’t try to understand Covid-19 just the effect of the R factor and the death rate. In fact there was very little data about the virus at this time so there was all based on the statistics of previous pandemics. Maybe we have forgotten this one because there has been a much more recent one on exam grades.

OfQual were ordered to come up with an algorithm for estimating student grades. They told the government it was almost impossible but their advice was ignored and they were told to make sure there was no grade inflation. Their objective was to allocate grades based on the information they had and to make sure that there was no grade inflation. Traditionally in large software projects one has “bug hunters” (rewarded by the number found) to test the robustness of the system. In this system it is not about robustness but about appropriateness. It would have taken very little resource to discover what the recipients discovered when they got their grades.

A little digression I am very interested in AI and Kieren sent me a paper about AI philosophy. One of the concepts was *simulated philosophy* which would use computational simulation to learn about AI and test its explanations. Surely something like this could have been used to test the appropriateness of this algorithm.

Algorithm may be set to become the word of the year or maybe it will be “virus”. Biological virus we knew couldn’t be treated with antibiotics but they didn’t occupy a central word space (up to now) . They became more visible when applied to an virus on a computer which always it seemed malevolent – eg closing a system down by overloading it, demand a reward or else….

Algorithms typically have a single objective, when we move the world of multiple objectives then we move into the world of AI. For a while we may use the term algorithms but the meaning of the word will change and looking back in the future we may have a nostalgia for the old algorithm.

Let’s consider the real world where there are multiple objectives. I will give a quote from “Dawn of AI Philosophy” by Amin Afrouzi.

For example, we could test whether AI agents with a multi-vector reward system regard different types of rewards as interchangeable. This allows us to test, for example, whether a Benthamite reduction of all rewards into a singular utility function is plausible. Or whether different rewards can lead to a stratified reasoning system akin to what Joseph Raz proposes in his account of political authority. (In a Razian framework, reasons are hierarchically ordered and are weighed on different scales. For example, in rejecting a deal, I may act on first-order reasons such as desires or needs, which I weigh on a single scale. Or I may do so out of illness, without even considering the balance of first-order reasons. In the latter case, illness is a secondary-reason because it defeats first-order reasons not by weight but by type. Political authority, the view goes, introduces second-order reasons for action.)

As you may have noticed this talk has edged into the world of AI and it is worth expanding one’s knowledge a bit. There are different types of system. Firstly individual *non-competitive systems* such as Google’s language translator or Siri, Alexa etc. Then we have *competitive systems* where a number of participants have their own system which is trying to outdo others – the classic example is stock trading in the financial market. What happens when these systems are competing for a limited resource then genuine ethical questions arise.

Just a small digression – we all know about the trolley question. People from different cultural backgrounds come to very different conclusions. There are lots of issues about freedom, fatalism, thinking as an individual or as a group etc. Take Chinese fatalism—a cluster of beliefs, deeply rooted in Chinese **culture**, according to which one should allow events to run their natural course without active interference. Once we get into the world of AI then all these would need to be resolved or we would have a system where you have to choose which school you are. A very good example is self driving cars. It is possible when you get into one you might have to choose who is more important - those inside the car or those outside. (see *moral machine* website)

A final type worth thinking about are *collaborative systems*.

Elevated AI is a genus of systems that not only reflect on their own actions, but also detect other actors and consider their relationship with them. Collaborative AI agents regard other actors as potential collaborators. If they anticipate that by working together, they can maximize their individual and collective performance, they form groups, share information, and even transfer skills to one another. They can track other actors’ degree of productivity and deepen their relationships with highly contributing players over time. In such systems, AI can interact with others in ways that resemble promising and contracting. They track each other’s activities to avoid those that do not contribute or even to retaliate against non-reciprocating counterparts. Just as they can deepen their relationship with the most contributing collaborators or “bond” with repeat players. They can even (collectively) ostracize or defame one another by sharing information. If these machines work on behalf of human users, the ethical responsibilities they undertake would be shared by their human clients, who may be oblivious to those responsibilities.

Moreover, Collaborative AI can give rise to newly ethically challenging behavior. For example, AI agents that customize your hotel prices can secretly collude with those that book your flight to drive up prices. Similar collusive behavior can create utility crises or stock market bubbles.

But to get back to today’s world another word we haven’t talked about is “App”. This word is becoming ubiquitous – it is a shortening of the word application. It is a portal or window into internet that delivers (in theory) results for a specific purpose. We get very used to using them and they become part of our lives. But a word of warning they are a way into the world of AI and before we realise this it will be too late.