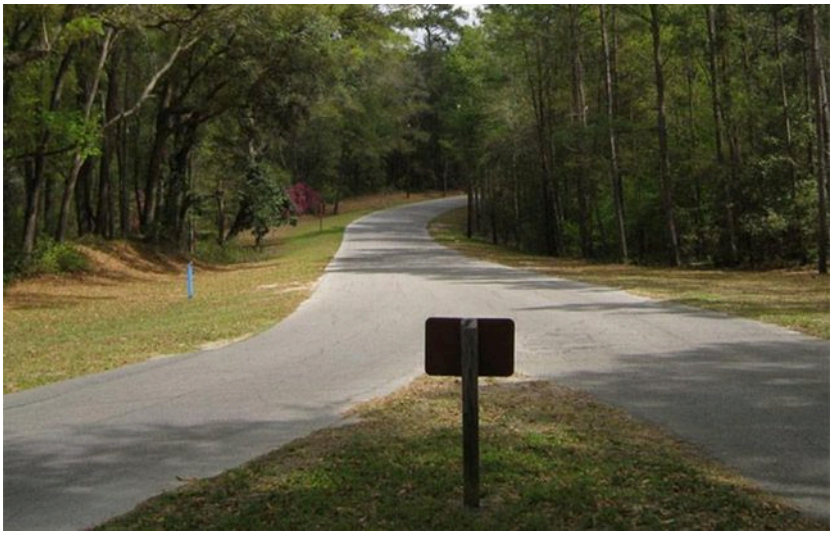
Programming fatalistic, uncertain or suicidal AI

Originally published June 21, 2017 at https://www.aiia.net/artificial-intelligence/articles/programming-fatalistic-uncertain-or-suicidal-ai



The problem of when to give up on a cherished goal is a profound philosophical issue. If the best human minds haven't resolved it into a definable and universally applicable formula, it's hard to define what to teach computers about it.

Machines which 'hang' or loop fruitlessly are very easy to create, as are machines which abandon their task at the slightest obstruction. That precious middle ground between perseverance and goal-abandonment defines creative intelligence in every sphere from business to courtship — and has become a particular study within AI research in recent years.

One recent Berkeley research paper, entitled *The Off-Switch Game* , concludes that artificial intelligence frameworks may need a degree of built-in uncertainty in order to replicate the kind of circumspection and reflection that humans typically employ when considering their efforts to attain a particular result.

The object of this reflection, in the hypothetical scenario, is whether the AI should in any instance obstruct human access to whatever 'off switch' the system might have.

The paper concludes that AI systems which are uncertain about their actual purpose are less likely to impede human-level access to whatever termination mechanisms they have in place.

Quantum leap

But introducing uncertainty is not a blanket solution, because excessive uncertainty makes the AI system indecisive and ineffective; instead of looping externally by re-performing ineffective actions , it will loop on the threshold of decision, much as we are prone to do when we consider the stakes to be critical.

'It is important,' the paper observes. 'to note that this uncertainty is not free; we can not just make [the AI] maximally uncertain about [the human's] preferences. If [the AI] is completely uncertain about [the human's] preferences then it will be unable to correctly select [the optimal solution] from its set of options.'

Without the property of doubt, the AI may be 'psychotic', with too much doubt 'neurotic'.

If this begins to seem familiar to programmers or tech enthusiasts, it's probably because the concept addresses the space between a zero and a one — not traditionally the strong suit of binary computing systems.

The coming advent of effective and practical deep learning systems based around quantum computing (QC) seems likely to have a radical contribution to make to this area of AI research, since QC systems operate natively in less certain or didactic space.

However, since the security concerns are already so great around the way binary-based AI may 'branch' into unreasonable courses of non-supervised action, QC decision-making systems also seem likely to stay ring-fenced in the groves of academe for some decades to come.

A question of commitment

The systems under study both in the paper and in the field in general have the capacity to generate their own 'else…if' loops, effectively. If they need to consult on these decision trees, they can never evolve into genuine unsupervised learning. And if those loops are pre-programmed or otherwise overly proscribed through human fear of negative autonomous decisions then the system cannot grow as an entity.

The Berkeley paper deals, as has other recent work, with science-fiction's favorite prophesy of doom — the possibility of an AI entity deciding that human interference with its processes is unacceptable. Even an AI system hypothetically imbued with Asimov's three laws of robotics is faced with the same tension between inaction and injurious action, and risks, in doing no harm, to do no good either.

The Berkeley researchers posit future research work exploring potential negotiation between the system and the human trying to switch it off, presenting the possibility of a person trying to win an existential argument with a machine over a matter which may prove critical — another popular sci-fi trope.

Prior work which addresses the ‘off-switch’ challenge has proposed to make AI systems indifferent to being deactivated, dubbing such systems as corrigible agents. The work proposes the presence of persistent ‘drives’ in AI systems, apparently analogous to a person’s underlying obsessions, which are likely to be present ‘unless explicitly counteracted’.

The paper also notes the problem’s proximity to that faced by companies to balance salary expenditure considerations with motivational psychology as regards the extent to which it can trust its employees.

It further observes that the off-switch model is also difficult to conceptualize because the typical spur of ‘reward’ is hard to pre-program into situations which may effectively involve failure.

IMAGE: *Flickr*