

Would an intelligent car go joyriding?

joyride (noun)

"A pleasure trip in a motor car, aeroplane, etc., often without the permission of the owner of the vehicle"

Oxford English Dictionary

The car would not want to do wrong

People generally like to justify their existence by believing they are doing good. Many believe it is a defining concept of being human: they know right from wrong.

However, the idea of morals is tricky. As Nielsen (1968) explains, they cannot be touched or manipulated in the physical sense, since they are not objects. A defining characteristic of a moral is that you can never prove one more correct than another, since they are not verifiable.

Nevertheless, Baier (1958) lays down a number of rules that all morals must follow. To paraphrase, each must be:

- A matter of principle
- Able to be followed by everyone
- For the good of everyone – it needs to be acceptable whether it is being 'given' or 'received'

A problem with Baier's ideas is that it is inevitably ethnocentric. What is acceptable to one culture and set of traditions is not to another. Nielsen puts his argument across through a crass (from our viewpoint) example. Say you were to convince a soldier in Nazi Germany that he and his family were Jewish, enough so that he was prepared to send himself to the gas chambers. Then, by making him realise that it was all trick, he would most likely not want to follow through with his plan. How would he be able to justify this? How has his family changed from when he thought they were Jewish, and afterwards? Only his perception of the situation has altered.

Joyriding, as defined above, is a criminal offence. Although positivism is generally accepted as the basis for English law, an intelligent being would probably use this as guidance when deciding what is right and wrong. A full discussion of how we form and why we follow laws is beyond this essay, but I am using the Natural Law viewpoint that morality forms the backbone of the legal system. This enables the creator of an intelligent machine to give a certain amount of flexibility: it would not be necessary to program or otherwise list every law of every country it could be in.

However, this conflicts with the problems raised earlier. What is moral, or acceptable to one race, is not necessarily to another. For example, as humans we generally accept the death of another species of animal as undesirable, perhaps cruel, but it is not comparable to the death of another human. However, even in seemingly clear-cut cases such as 'murder is wrong', there are exceptions. The tribal instinct of humans, as Nielsen again explains, is perhaps the basis for the view that killing another person who is not in the same social group (such as two warring tribes) is acceptable.

Defeasibility is easily accepted in human cultures. For example, there are many phrases such as "you've got to be cruel to be kind". If it were possible to build morality into a car, it would help build its 'belief-filter', so it can decide exactly how to react to certain situations. For example, a slightly more callous model may try to convince our vehicle that it is worthwhile to scare its passengers. Should it accept this advice, since it could benefit through extortion?

Consider the situation in which an intelligent car, capable of making moral decisions, always tries to do the right thing. Should it transport a cheating husband to his mistress? Should it even matter to the car? If it decides that it is not, how does that affect other situations? How should the car react to the disposal of another intelligent vehicle? Should it always follow laws and principles designed for humans? Is going on a joyride acceptable when it would not harm anyone?

A machine has no concept of 'joy'

The issue of emotion in machines is complex. Many science fiction writers assign a lower form of intelligence to those with emotions. For example, Spock, in the TV show *Star Trek*, has no emotions, and is truly logical. The suggestion seems to be that he has a more advanced brain, and thus a higher degree of rationality.

However, Evans (2001) believes that this negative view of emotion is unrealistic, and suggests that emotional awareness in a computer could be very beneficial. He gives the example of a computer that could detect the fact you have had a bad morning, and act accordingly, either giving you space, or trying to cheer you up.

Having machines that do more than detect emotional states could also be needed to be truly intelligent. Simon (1967) predicted that machines that can quickly make decisions accurately would need some form of emotions. For example, having two goal states provides a conflict that Simon suggests would only truly be solved by having primitive emotions. They could serve as an interruption mechanism: fear would indicate to the machine that it should escape a dangerous situation, and quickly.

Evans suggests that the majority of emotional machines will be built for entertainment: to create better AI in games, or in robotic pets such as Sony's AIBO. As mentioned before, it would allow easier interaction if the computer could understand and respond intelligently.

However, giving machine emotions could also have unintended consequences. The Sci-Fi show *Red Dwarf* has an intelligent toaster, one that encourages you to eat toast. However, it ends up being destroyed by the crew due to its incessant nagging. It could be argued that its primary goal state is creating toast, and it does everything it can do to be 'happy' – to fulfil its goal – contrary to own survival.

There can be no truly intelligent car

In this context, intelligent means more than just 'processing data'. Instead, it comes from the main thrust of this essay: creating an Artificial Intelligence, a far grander meaning. Rather than spend several hundred words recursively defining words however, I shall take the direction that intelligence cannot be truly defined accurately. Instead, I shall follow Penrose in examining what it means to be intelligent, and how the result can affect the meaning of the word.

Perhaps the most famous of all AI theories is the Turing Test. It is an attempt to answer the question of how to tell if machines are intelligent. However, many, including Bringsjord (1995), believe it to be inaccurate. In one of his texts, he explores possible extensions, designed to defend the test against Searle's classic Chinese Room problem. This includes ideas such as one that allows the judge to use the sensor and motor skills of the being as another indicator. However, as he argues, it is very possible that even a finely crafted machine with a wide range of impressive actions may simply be performing them randomly. In other words, is a machine that appears to be intelligent actually intelligent?

Another test forwarded by Harnad (1990) involves examining the physical processing structure: in a human's case, brain tissue. This however is not helpful to our quest: the heading under which Turing actually discusses his design is "Can Machines Think?". Instead,

this extension seems to solely focus on whether the entity under examination is human or not.

Instead Bringsjord brings a more logical framework that attempts to study and find if the subject is processing the information in a logical fashion. By looking at the high-level algorithms that process the data, it is possible to discount Searle's argument. However, as Bringsjord himself states: it is an intractable, possibly incomputable.

If we cannot detect if a computer is intelligent, then how else can we decide if a car would be able to go joyriding?

There are many different arguments, into which Penrose (2000) delves into in his papers. For example, he contends that:

"Appropriate physical action of the brain evokes awareness, but this physical action cannot be properly simulated computationally."

In other words, we can never hope to create an intelligent car, since there are quantum interactions in the human brain taking place that are impossible to imitate. Critics counter by saying that it is eminently possible to simulate the decisions in a lowly earthworm, which uses the same kind of low-level functions.

Penrose also suggests that even when a computer is generally considered good at a problem, it does not fully understand it. When playing chess he asserts, computers simply calculate a huge number of moves ahead, and do not always assess the situation wisely. His example is when a certain barrier is formed, consisting of a mainly of a line of pawns across the board, it is impenetrable. To a human it is obvious: it forms a boundary, instantly visible. Deep Blue however, decided to punch through the barrier, and face certain annihilation. Vitalists would argue the computer would never be able to 'see' the problem, since it lacks the 'spark' that would give true consciousness.

I maintain that even if true intelligence is not possible, it might not matter. If a chess-playing robot sometimes seems to make arbitrary decisions, it does not necessarily follow it is unintelligent. Deep Blue posed a challenge to the foremost chess players in the world: it chose highly intelligent moves the majority of the time. Even if a car arbitrarily decided to go joy riding, to me it might actually seem more human.

"Two things are infinite: the universe and human stupidity; and I'm not sure about the universe."

Albert Einstein

It would not be a joyride as defined: it is its own owner

'Robot rights' seems to be a fairly recent buzzword. Philosophers tend to view advocates of this trend with a certain amount of derision: in *Android Epistemology* Ford *et al.* tried to 'keep a straight face' when asking Minsky of his view.

Minsky's reply turns the argument on its head. Instead, he asks why humans have rights. He, as many others do (Evans 2001), believe that the decision is inextricably linked to how the entity thinks. If it is conscious, then perhaps this awareness should be preserved.

However, it is more than this. As Boden (1995) suggests, "when describing a robot as creative (or intelligent), [it] carries significant moral overtones". She wonders if we will ever actually believe a robot is truly intelligent, since we could be limited by our own perceptions. If we treated an intelligent car like a calculator, who would suffer the blame if things went wrong?

This is the crux of the argument. We are reminded that with rights come responsibilities – but does it work in the other direction? We entrust our lives to technology every day. If a machine became aware of this, should it have recompense? If we accept Asimov's famous robotic laws, then we are placing human life above that of the robots' existence.

Boden believes that we cannot know at this point exactly how we will treat androids, or other forms of intelligence. Do we treat them like animals, and give a subset of our own rights? Or will the attraction be so strong for some that like in a number of fiction works (such as *Do Androids Dream of Electric Sheep?*, *AI* and *Bicentennial Man*) that we might even fall in love with these beings?

A car could not be able to make the decision

Free will is continuously contentious part of philosophy. Indeed, so many thinkers have different opinions that some have started to avoid the issue entirely. (Nelkin 1996) Nevertheless, it a concept that even non-philosophers wrestle with, since it is perhaps the defining characteristic of being human. Why do we make decisions? How can we feel that we have the right answer?

Minsky (1995) discusses this topic in his entertaining answer to the issue of robot rights. He claims that the distributed nature of our brains make it impossible to decide exactly how we make decisions: that they just happen. He calls this idea of not knowing what your reasons are 'free will'.

With this definition, it is hard to see why an intelligent car would want to have such a characteristic. By calculating exactly every possible permutation, much like a chess robot, it could make decisions based upon its probability of obtaining a specific goal state.

To make a decision, you need to be aware of the situation - as defended by Nelkin: "All consciousness involves awareness". The classic example is that unconscious people generally are not aware. I contend there is a certain grey area: when asleep, humans are generally do not make conscious decisions, yet they are still aware of certain inputs. Nelkin does recognise this, and provides the apt example of having no memory after a long car journey home.

Nelkin also submits that we like to be in control, and be aware of a situation. He points to the many studies that show people generally feel calmer and even have a higher tolerance for pain when they feel they have some power. He follows this by talking about the body/mind problem, but also contends we that we cannot even completely command our thoughts. Often, people complain about not being able to resist chocolate, or cannot stop humming a song.

This viewpoint is a very deterministic one. Nelkin wonders if "we are no more than mere pushed-and-pulled objects in the external world" – we actually have no control over anything, so there is no possibility of making decisions. "When our autonomy is questioned... everything – all thought, all world, our very self is at risk."

This bleak outlook is counted by Edmonds (2003), who discusses designing free will into an entity. His paper begins by paraphrasing the words of von Neuman: that "anyone who considers computational methods of implementing free-will is in a state of sin". von Neuman was referring at the time to random numbers, but Edmonds suggests that the same principles apply to our topic.

There are many reasons why generating random numbers through algorithms is a "sin", but it currently used in cryptography and for commerce transactions throughout the world. He puts forward a framework where an environment is produced that could develop an entity capable of free will. His argument is essentially that no enough effort has actually been

expended on creating a practical solution to the many problems. The result may not be identical to a human concept of free will, but it may be good enough.

The car could not understand the concepts

Having a creative machine is often quickly rejected, especially by Vitalists. How can mere mechanics create something original? Wondering whether machines can ever be truly creative seems common thought for philosophers. Boden (1995) is certainly one of them. However, she also asks the ancillary question: would we know?

Creativity seems to be an easy thing for humans to do. Indeed, as Chomsky (1965) said, we create new and original sentences everyday. Many of these may be new to the speaker, but have been uttered by people before: Psychological or P-creative ideas could not have been thought of before by the thinker. Historical or H-creative ideas are also new to everyone else. This split coined by Boden is useful, since it provides necessary context to examine fully these novel ideas.

One of the ways in which we can be creative is via exploiting the rules in a form of expression. Chomsky shows us that this is very possible in the English language: there can be an infinite number of adverbs strung together, which means that a particularly colourful description might never have been heard before. Boden also gives the example of how the rules of music have changed due to constant manipulation of the conceptual space.

This feature can easily be utilised by a computer: indeed many programs already exist. These can use genetic algorithms and other such techniques that search towards a goal state, such as a pleasing tune. Exactly what this goal means to the computer is another issue. Boden even argues that a machine that can improvise Jazz or other music would not suffer from the inherent inflexibilities in humans, such as our limited short-term memories, and therefore could create H-creative melodies of which humans would not normally be capable.

This shows one area where a machine would have a conceptual state different to our own, and would possibly express themselves in a different way. I submit that English itself could be redefined: we split ideas into sentences and paragraphs to suit the way humans think. An intelligent car might well express his desire for driving in a different way that we could not understand.

Even more than that: embodiment of a creative machine could radically alter the way it perceives and learns from its environment. Many of our metaphors as Boden explains are due to our position in 3D space, with the certain characteristics of our bodies. A car might observe the world in such a different way that it consistently draws different conclusions, and would understand the concept of joyriding in a thoroughly different way – if it could at all.

References

Penrose, R., Shimony, A., Cartwright, N. & Hawking, S (2000), *The Large, the Small, and the Human Mind*, Cambridge University Press, Canto edition.

Bringsjord, S. (1995), "Could, How Could We Tell if, and Why Should – Androids Have Inner Live?", in Ford et al. 1998, pp 93 – 123.

Minsky, M. (1995), "Alienable Rights", in Ford et al. 1998, pp 307 – 313.

Boden, M. A. (1995), "Could a Robot Be Creative – And Would We Know?", in Ford et al. 1998, pp 51-73.

Ford, K. M., Glymour, C & Hayes, P (eds) (1995), *Android Epistemology*, MIT Press

Baier, K. (1958), *The Moral Point of View*, Ithaca

Nielsen, K (1968), "On Moral Truth", in Rescher, N (eds), *Studies in Moral Philosophy*, Blackwell, Pittsburgh, pp 9 – 25.

Evans, D (2001), *Emotion: The Science of Sentiment*, Oxford University Press.

Simon, H., (1967), "Motivational and Emotional Control of Cognition", *Psychological Review*, 74 29 – 39.

Harnad, S. (1990), "The Symbol Grounding Problem", *Physica*, 42 335-346.

Chomsky, N. (1965), *Aspects of the Theory of Syntax*, MIT Press.

Nelkin, N. (1996), *Consciousness and the Origins of Thought*, Cambridge University Press.

Edmonds, B. (2003), "Implementing Free Will" In Davis, D. N. *Visions of Mind - Architectures for Cognition and Affect*, IDEA Group Publishing.