

HOW NARRATIVES INFLUENCE THE ROLE OF ARTIFICIAL INTELLIGENCE IN SOCIETY

Abstract

This article consists of three parts. In the first part it is explained how narratives can influence the views people have about, amongst others, the role of AI in our society. In the second part, two narratives are discussed: one about human agency, the other about the nature of machines. In combination, these two narratives disallow the possibility that AIs act. Finally, the article discusses four ways to deal with the competing narratives.

Keywords: agency, cognitive dissonance, coherence, narratives

1 Introduction

Some decades ago, the present author had a debate with a professor in computer science.¹ The topic was whether computer programs could be agents. For instance, would it be possible that an application program, without human intervention, order new supplies for a store? My own position was that this would be possible and – more in general – that the law should also recognise juridical acts such as contracting if they were performed by computer programs. My opponent fiercely opposed: an application program could not perform acts, just as a pen could not write a letter. Humans use pens to write letters and computer programs to buy on the internet, but neither pens nor computer programs could act all by themselves. Therefore, the law should not provide facilities for autonomous agency of computer programs.

With the recent large-scale introduction of artificially intelligent applications ('AIs', for short), the question of whether these computer programs can act has become more urgent. However, the answer remains as controversial as it was several decades ago. My present view is that the controversy continues because the opposing answers – AIs can, or cannot, act autonomously – are based on different narratives about what these programs are and how they are different from human agents.

The claims that AIs can act autonomously, and the related claims that AIs can act for reasons, and that they can act intentionally, belong to one narrative. According to this narrative, some artificially intelligent computer programs are like human beings, at least to the extent that agency is concerned. Therefore, it is possible to ascribe them agency. According to a different narrative, AIs – no matter how 'intelligent' – are machines that operate in accordance with the laws of physics. Such machines are not 'really' intelligent, have no intentions or free will, cannot 'really' act, and should not be empowered to perform juridical acts. The opposition between these narratives explains that some people are inclined to treat AIs for some purposes analogous to human beings, while others oppose this parallel.

It will be clear that if the above claim about the roles of narratives is correct, narratives are a crucial factor in determining the role of AIs in our society. It is the purpose of this article to

¹ Dietz, Wat doen computers als ze iets zeggen?; Hage, Handelende computers.

elucidate this claim by offering a view of what narratives are and how they influence human behaviour, the development of our society, and the role of AIs in it (section 2). In connection to modern technological developments, including the creation and use of AIs, two narratives play a significant role. They are the topic of section 3. One narrative deals with agency and human beings as the main kind of agent. The other narrative is about machines and their mode of operation. The narratives are incompatible in the sense that something cannot both be an agent and a machine. For AIs, which are typically implemented in hardware that is machine-like, but which also possess intelligence and capacities that used to be seen as typically human, the question arises in which narrative they fit. Section 4 addresses the question of whether it possible to merge the narratives about machines and about agents.

2 Narratives and related notions

2.1 Narratives

The word ‘narrative’ means ‘story’. In this light, calling a particular account of facts or events a narrative does not provide much information. In this article, a more specific notion of a narrative will be used. According to this notion, humans have many beliefs, and these beliefs cohere. Narratives are a means of structuring the beliefs of persons in coherent wholes. These narratives do not only contain the beliefs, but also links between them.

For example, a person’s self-image (a narrative) contains facts from his life but also links between these facts.² Take, for example, the story that contains the following facts:

- Caroline went to school.
- She took along a box with some food.
- She left home at 8 A.M.

The story also contains the following links between the facts that have an explanatory function:

- The box contained her lunch.
- She had to walk 15 minutes and school started at 8:20 A.M

These links connect the facts of the story into a coherent narrative.

The following is a longer example. The links are *italicised*.

Pedro studied law but found insufficient intellectual satisfaction in dealing with the rules of his country. *Therefore*, he decided to study philosophy as well. After finishing a PhD-thesis on legal philosophy, he found that his interest in the exact sciences was insufficiently satisfied. *In an attempt* to combine this interest with his earlier studies, he switched to the field of AI and Law in which he performed research on the philosophical aspects of AI in law.

This fictitious narrative not only mentions several facts about the development of Pedro, but also the links between these facts which in Pedro’s self-image are based on his motivation to develop various aspects of his personality. The links in this narrative are inferred from the actual

² More on the role of narratives in defining a person’s self-identity, see Schechtman, The narrative self.

events in Pedro's life, but at the same time provide these events with the coherence and the meaning they have for Pedro. In this way the narrative is for Pedro part of the larger narrative of what person he 'really' is. More in general, narratives as they are discussed in this article are a mix of fact descriptions and links between these facts. Unlike the fact descriptions, the links are not primarily descriptive, but interpretative. They cannot easily be falsified by empirical research, although factual information can make some interpretations more plausible than others.

2.2 Belief sets, inference rules, and priorities

The *belief set* of a person is everything this person believes, together with the links this person uses to 'derive' beliefs from other beliefs, and with priorities this person uses to make these derivations. This sounds abstract, but an example may make it more concrete. Let us consider the belief set of Jane. Jane believes that there is smoke, and this belief is supported by her experience of seeing smoke. The inferential step from the experience of seeing something and believing that this something is the case is so obvious that it will usually go unnoticed. That there is a step at all can be noticed from the fact that it is possible to 'see' something without believing it to be the case. This may, for instance, happen if one knows to be under the influence of a drug that can cause hallucinations.

The link between what one experiences as seeing and the belief of what is seen is an instance of an 'inference rule'.³ If a person uses an inference rule, this means that she is disposed to believe the consequence of this rule if she believes the rule's antecedent.⁴ More concretely: if Jane uses the inference rule from 'I have the experience of seeing X, and therefore X is the case', this means that Jane is disposed to believe X if she has the experience of seeing that X is the case. If Jane also uses the inference rule that 'Smoke means fire' she will be disposed to believe that there is a fire if she believes that there is smoke.⁵

Some inference rules are based on explanatory relations. The presence of X explains why Jane has the experience of seeing X, and therefore seeing X can lead Jane to believing that X is present. For instance, the presence of smoke can be explained from the existence of a fire, and therefore believing that there is smoke may lead to the belief that there is a fire. Another example would be that Pedro's studying law may be explained from his desire to know the law. Believing that Pedro is studying law may therefore lead to the belief that Pedro desires to know the law.

³ The term 'inference rule' stems from proof theory in formal logic (Smith, The laws of truth, 387), but here it will be used for a psychological mechanism.

⁴ The use of an inference rule is not the reason for believing the rule's consequent based on the rule's antecedent. If a person is disposed to believe one kind of thing in case she believes some other kind of thing, being thus disposed is the use of the rule, and not something for which the (use of the) rule is a reason.

⁵ Logicians should notice the word 'disposed' here. Psychological inference rules work different from inference rules in logical calculi, although the difference becomes smaller if the logic is one for defeasible reasoning.

However, there are also inference rules that have nothing to do with explanation. If a benevolent government believes that measure M will increase the prosperity of the country, this may lead the government to believe that it should take this measure. The inference rule in this case leads from a belief about what promotes prosperity to a belief about what the government should – and will – do.

2.3 Coherence and cognitive dissonance

Belief sets are not stable; they change in the course of time. One cause of change is sensory perception; another cause is forgetting. A special cause of why a person changes her belief set is that this person experiences her belief set as incoherent. Such an experience leads to mental discomfort and motivates this person to modify her belief set in a way that takes the incoherence and its accompanying discomfort away. The discomfort has received a technical name in psychological theory: *cognitive dissonance*.⁶ Cognitive dissonance is the uncomfortable experience that one's belief set is incoherent.⁷ ⁸ Let me give an example. Jane sees and believes that there is smoke. She uses the inference rule that smoke means fire. As a result, she believes that there must be a fire. However, even after investigating the situation carefully, she does not find any fire. So, she believes that there must be a fire, as there is smoke, but also that there cannot be a fire since she did not find it in her careful investigation. Jane is confused – the mental discomfort – because her belief set is incoherent. This confusion will typically lead her to modify her belief set in a way that coherence is restored.

There are three ways to restore coherence. Jane can give up her belief that there is smoke, but then she should find an alternative explanation for her experience of seeing smoke. She can also adopt the belief that there is a fire, but then she must think of an explanation why her careful investigation did not lead to the discovery of a fire. Or, finally, Jane can give up her inference rule that smoke means fire.⁹ Since the use of an inference rule is typically the result of several events in which the rule turned out to be reliable, inference rules are usually less easily abandoned than the beliefs that, together with the inference rule, lead to the mental discomfort. From this example we can see that if a belief set is experienced as incoherent, it must be changed to take away the resulting discomfort.

⁶ Festinger, A Theory of Cognitive Dissonance.

⁷ In traditional psychological theory the term is most often used for the situation in which people did something that they find hard to endorse. See Tavris and Aronson, *Mistakes were made (but not by me)*. So, the use of the term here is not completely in line with what is usual in psychology, although fully in its spirit.

⁸ This formulation is perhaps too limited, as it only deals with beliefs. Cognitive dissonance also occurs if an agent did what he believes he should not have done. The incoherence is then not between beliefs, but between a belief and a disposition to act. This difference deserves more attention than it can be given here.

⁹ It is also possible to make an exception to the inference rule, but then Jane must find a reason for making an exception. I will leave this complication out of consideration here.

2.4 Entrenched narratives

In considering what change to make, which belief should be given up, or which inference rule abandoned, some elements of a belief set turn out to be more important than other elements.¹⁰ A person facing an incoherent belief set will typically restore coherence by giving up a less important element of the set, to rescue more important elements.¹¹ Which elements of a belief set are less important than other elements and therefore more prone to be given up in case of experienced incoherence, should be a matter of the belief set itself. It is presumed that a belief set does not only contain ‘ordinary’ beliefs about the facts in the world, but also meta-beliefs about, amongst others, what beliefs should be given up in the case of incoherence.

In constitutional law theory we have the phenomenon of rigid or entrenched constitutions. Constitutions are said to be entrenched ‘[...] if the procedure to modify them is more demanding than the one applied to ordinary legislation ...’.¹² The effect of entrenchment of constitutions or constitutional clauses is that these items are less prone to change and therefore more stable.¹³ Not all narratives are entrenched, but this article focuses on the entrenched ones. They include fact descriptions, inference rules, and priorities which are less prone to change if a belief set has become incoherent and is in need of updating.

Let me give an example. A scientist – perhaps inspired by the work of Popper¹⁴ – endorses a narrative about science and scientific method which holds that in scientific research there are no dogmas. According to this narrative, every scientific claim can be criticised and is open to falsification or revision if there is sufficient evidence that the claim is false. Suppose that somebody attacks this narrative and claims that even in science there are dogma’s¹⁵ and that this narrative is precisely an example of such a dogma. Now there are two possibilities. The first is that the scientist admits that his view of science seems dogmatic and – given the alleged nature of science – false and should therefore be given up. The second possibility is that the scientist sticks to his narrative and admits that his view of science is itself not a scientific claim. The test for being scientific should therefore not be applied to the narrative. The first possibility denies

¹⁰ This idea and its application to the notion of a narrative was inspired by the work of Quine (Two dogmas of empiricism; Quine and Ullian, *The Web of Belief*, 3f.).

¹¹ This is a psychological claim, not a logical one. It is a claim about what a person normally will do, not a claim about what this person rationally ought to do. The strict distinction between psychological and logical claims is not meant to suggest that there is no connection between the two. In my opinion, logical theory is a rationalised version of psychological mechanisms. However, the difference between psychology and logic is not a topic for this article. For more details, see Hage, *How Law’s Nature Influences Law’s Logic*.

¹² Pereira, *Entrenched clauses*.

¹³ Tushnet, *Constitution*.

¹⁴ Popper, *The Logic of Scientific Discovery*.

¹⁵ Kuhn, *The Structure of Scientific Revolutions*; Lakatos, *Falsification and the Methodology of Scientific Research Programs*.

that the narrative about science is entrenched. The consequence is that its claim about science may have to be given up in the light of the attack. The second possibility treats the narrative as entrenched but must admit that the narrative about science can therefore not be a scientific claim itself, as scientific claims are not entrenched.

2.5 Narratives and the meanings of words

This second possibility deserves more attention. The entrenched narrative about science and its method imposes limits on what can be called ‘science’ or ‘scientific’. Therefore, it may be said that the entrenched narrative (co-)defines the meaning of the word ‘science’.¹⁶ Being science literally *means* to fit in the narrative of what science is. Something that does not fit in the narrative can, for that reason, not count as science. Following this line of reasoning, our scientist who treats the narrative about science as entrenched had to admit that the narrative itself is not a claim of science. More in general, entrenched narratives can define what counts as belonging to a particular category. The nature of science may be seen as defined by such an entrenched narrative, just as the nature of religious faith, or – more relevant for our present purposes – the nature of agency and the nature of a machine.

Some narratives are very personal. They are endorsed by a single person, but not necessarily by more. The narrative that defines a person’s self-image would be an example. Other narratives, with an important role for narratives that (partly) define what a word means, will be shared broadly. It does not make much sense for a scientist to endorse the falsificationist narrative that was described above, if she is the only one. However, if the narrative is broadly shared, it defines what science is¹⁷, and plays a role in the education of new scientists and in debates on what investigative practices deserve to be subsidized from public means.

2.6 Intermediate conclusion

These observations about entrenched narratives as relative stable subsets of belief sets, and about the social nature of some of these narratives, bring us to the first main conclusion of this article. There are narratives that are widely shared, and which sometimes even co-define the meanings of words. These entrenched narratives have a strong influence on how we ‘see’ the world, in particular those parts of the world with which these narratives deal.

Two important narratives deal with agency and with the nature of machines. Together, these two narratives strongly influence our ideas on the extent to which AIs can act and on how law should deal with AIs that ‘do’ things which would be called ‘actions’ if performed by human agents.

¹⁶ This is in line with inferentialism as a semantic theory. See Brandom, *Making it explicit*, chapter 2.

¹⁷ There is a complication here, which has to do with the distinction between conventional and constructivist facts (see Hage, *The Nature of Law and Constructivist Facts*), but for the present purposes this complication can be ignored.

3 Two narratives

In the debate about the agency of AIs there is an important role for the narrative about agency in general. This narrative is modelled after the prevailing view (narrative) of human agency, as the latter to a large extent determines what counts as agency in general. Therefore, the argument that AIs are (not) like human agents plays an important role for the question of whether AIs can be agents. Regrettably, the narrative about human agency has no canonical content. Its application to AI-agency is only possible by making assumptions about the narrative, and these assumptions can be challenged. With this proviso, let us start.

3.1 Intentionality of human action

A typical human action is intentional: the agent does something and is at the same time aware of doing it. For example, John drinks a glass of water and while drinking John is aware that he is drinking. This is a modified formulation of what Searle called ‘intention in the action’.¹⁸ Moreover, a typical human action is rational in the sense that the agent believed that it was the right thing to do. Not necessarily morally right, but the right thing to do for the agent. This means that the agent wanted to achieve some goal, and believed that the action was an adequate, perhaps even the best, means to achieve this goal. For instance, John wanted to quench his thirst. He believed that drinking a glass of water would quench his thirst and therefore John believes that drinking a glass of water would be the right thing for him to do. This belief goes hand in hand with the intention to drink the water. This intention is formed before the action is actually performed. It is what Searle called a ‘prior intention’.¹⁹ If this instrumental rationality is combined with the intentionality of the action, we arrive at the following model of human agency:

- The agent wants to achieve a goal.
- She believes that performing some action is an effective, or even the best, means to achieve the goal.
- Therefore, she forms the prior intention to perform the action, and
- this intention makes her actually perform the action,
- which is itself intentional as well.

3.2 Freedom of the will

This model of agency emphasizes the role of the intention with which an act was performed. Another aspect of agency is the freedom of the agent to perform an action, or to refrain from performing it. There are at least three aspects to this freedom. The one is that the agent is physically capable of performing the action. If John’s hands are tied, he cannot pick up the glass of water to drink it. In that case, John lacks the capability to drink the glass of water. So, even if he wants to quench his thirst and believes that drinking the water will satisfy his desire, he will not drink the water. If John knows that he cannot do it, perhaps he does not even form the intention to do so.

¹⁸ Searle, Intentionality. An essay in the philosophy of mind, 83f.

¹⁹ Searle, Intentionality. An essay in the philosophy of mind, 84f.

Another impediment to the performance of some action is that the law makes this action impossible. If John wants to marry his sister, the law prevents this by withholding from John (and all other legal subjects) the legal competence to marry close relatives, thereby making the marriage impossible. This is the second aspect of free agency.

A third possibility is that the law does not withhold the competence to marry close relatives but forbids to do so. If such a prohibition is not combined with a lack of competence²⁰, John can, legally speaking, still marry his sister. However, if the prohibition is combined with the threat of a legal sanction, this may take away John's desire to marry his sister. John will not form the intention to do so, and he will not go through the wedding procedure. This possibility deals with influencing an agent's will. The third aspect of free agency is that the agent can freely determine whether he wants to perform an action. Traditionally, this topic goes under the heading of a *free will*.

Typical human agency consists of actions performed by agents whose will could freely determine whether to perform the action or not. The notion of freedom of the will is notoriously complicated.²¹ One reason that is often adduced to argue why there is never a free will, is that the will is determined by physical facts and causal laws (physical determinism). The narrative about human agency assumes that this reason does not apply to human agency and that the human will is free in the sense of not determined by physical facts and causal laws.

A second reason why the will is not free is that the will is sometimes influenced by factors that ought not have influenced it. Suppose that Chris wanted to marry John and that Henry, the brother of Chris, extorted John into marrying Chris. If it was not the government, but Henry, who unlawfully influences the will of John, we say that John's intention to marry Chris was not freely formed, or – in other words – that John's will to marry Chris was not free. In general, it is assumed that a will was not free if it was formed under the influence of a factor that ought not to have influenced it. Whether the will was free in this sense is therefore a normative, and not a physical matter. This normative freedom of the will, or rather the lack of it, is typically *not* considered to be an impediment to agency. If John was extorted to marry Chris, John's marrying Chris is still considered to be an action of John (and of Chris, of course).

²⁰ A legal prohibition of a juridical act often goes together with a lack of competence. Then the juridical act is impossible, and the prohibition is superfluous.

²¹ Kane, *The Oxford Handbook of Free Will*.

3.3 Why, allegedly, machines do not act

The law²², the philosophy of mind²³ and the philosophy of action²⁴, provide a lot of material to extract a narrative about mankind from. For machines it is difficult to find a source for a narrative. Therefore, the following account of our narrative about machines may be even more controversial.

Machines are tools, used by humans, to transform input into output. For instance, bread machines take as their input the ingredients for bread, together with some choices on a control panel, and some energy. As their output, they produce (loafs of) bread and heat. Cash machines take as their input an electronic bank account, an amount of (paper) money, some energy, and again choices on a control panel. Their output is an amount of money, heat, and a change in the bank account.

A typical machine consists of many parts, which cooperate with each other to produce the output. This ‘cooperation’ is typically not seen as coordinated ‘actions’, but as mere physical processes that – through clever engineering – are made to connect in a manner that allows the output of the one process function as the input of another process.

Although it is possible, and common, to describe the operation of a machine in terms of coordinated actions²⁵, such a description is also seen as metaphorical. In ‘reality’ there are only physical processes, governed by causal laws. There are no interfering ‘mental’ states, such as intentions or plans. All there is, is causal laws, operating on physical entities, and producing physical outputs. Machines do not have a will, let alone a free will. Neither do machines act for reasons. The factors that influence the operation of machines are causes in physical processes. The only way in which machines can non-metaphorically be involved in actions performed for reasons, is when humans use them. Then the human agent acts, and the machine is a mere instrument, like a pen used for writing. However, the use of a machine can transform the kind of action performed by a human being. E.g. filling the bread machine with ingredients and pushing a few buttons becomes baking bread. But even if the nature of the action is transformed through the use of the machine, it is still the human being who counts as acting. The machine only fulfils the role of tool, used by the human agent to perform her action.

²² From private law, especially the law about juridical acts such as contracts and marriages is relevant, because this law is based on a view of human agency. But liability law is relevant here as well, since a main ground for liability is responsibility, and responsibility is in turn intricately connected to agency (Honoré, Responsibility and Fault). The relevance of criminal law and its underlying image of man as an – in general – free and responsible agent, is obvious.

²³ See, for instance, McLaughlin, Beckermann and Walter, *The Oxford Handbook of Philosophy of Mind*.

²⁴ See, for instance, O’Connor and Sandis, *A Companion to the Philosophy of Action*.

²⁵ For instance, the Wikipedia page on machines (<https://en.wikipedia.org/wiki/Machine>) claims that a machine is ‘[...] a physical system that uses power to apply forces and control movement to perform an *action*’ (italics inserted).

4 AIs

The narratives about (human) agency and about the operation of machines that were sketched in the preceding subsections make it impossible that a machine acts. A typical human action of a particular kind is guided by reasons that influenced the agent's free will and based on the mental intention to perform precisely that kind of action. The typical operation of a machine is not guided by reasons, but is influenced by causes, which are not reasons in the relevant sense.²⁶ Moreover, a machine does not form the intention to perform some kind of action, and in the causal process leading to the output, no (free) will is involved.

4.1 Cognitive dissonance

This mutual exclusion of machine operation and action can become problematic if there are events which are like actions, but which also have traits of machine operation. Suppose, for example, that a computer program 'notifies' that the stock of a product is almost depleted, 'decides' to order new items of the product and 'concludes a contract' with the computer program of the good's producer. The steps which are mentioned in the narrative of agency also play a role in the computer program which performs this transaction.²⁷

Situations like the one in the example can produce cognitive dissonance. On one hand, the computer programs seem to act; on the other hand, the computers on which these programs 'run' are machines which 'therefore' can only act in a metaphorical sense.

4.2 The first solution: accept that machines cannot act

To take the cognitive dissonance away, three solutions seem available:

1. Accept that computer programs do not act, as they are 'only' causal factors that govern the operation of a machine (the computer).
2. Modify the narrative about agency, by downplaying the role of reasons, free will and intentions. This would have the effect that, at least in theory, some machines can act.
3. Modify the narrative about machines, to allow machines to act for reasons, have a free will and to form intentions. Also, this would have the effect that some machines can act.

The first solution is to downplay the mental capacities of AIs. After all, computers are only machines and computer programs are merely means used by human agents to make these machines 'do' something. In 'reality' the human beings who make and use the programs are the ones who are acting. The seemingly intelligent behaviour of AIs is precisely what the name says: only *seemingly* intelligent, and *seemingly* behaviour. This approach was taken by Searle in

²⁶ The word 'reason' has many uses, and one of them is for causes. For example, it is well possible to say that the reason why the vase broke is that it was dropped on the floor. These 'cause-reasons' are not the reasons that are relevant in the narrative about human agency.

²⁷ It is assumed here that the application program is based on a traditional algorithm, and is not the product of machine learning. However, this assumption is not essential for the argument of this article.

his famous essay 'Minds, Brains and Science'.²⁸ It can be strengthened by elaborating on, and emphasizing, characteristics of human agency which seem unattainable for machines.

4.3 The second solution: downplay the uniqueness of human agency

The second solution is to modify the narrative about human agency and indirectly also the narrative about agency in general. If humans do not act for reasons, if their intentions are mere 'epiphenomena' of bodily states and processes, and if the free will of humans is merely the absence of undesirable influences, human beings could just as well be seen as machines. Then there is no reason anymore to treat them different than 'mere' machines should be treated. Inversely, there would be no reason anymore to treat machines different from human agents. So, if we assume that human beings can act, so can machines.

Do human beings act for reasons? People tend to explain their actions by pointing to the reasons they had for performing them,²⁹ but are these reasons the 'real' reasons? There is quite a bit of psychological research showing that people have no mental access to their reasons for action.³⁰ So, if people claim to know through introspection what moved them, this claim should not be taken very seriously.

Also, the role of intentions in generating actions is open to questions. Human actions typically involve bodily movements. These movements can be traced back to impulses sent from the brain to muscles. Never a trace of evidence was found for the claim that the brain was influenced by purely psychological states, such as intentions. Moreover, if it is assumed that intentions are not only psychological states but also bodily (brain) states, actions are completely explained by the physical aspects of the brain states. There is no role in the explanatory narrative for the psychological aspect of intentions. This psychological aspect might just as well not have been there; they have no causal role and are mere 'epiphenomena'.³¹

Finally, if everything in human actions can be explained in terms of bodily states and causal laws, without role for how these states are experienced (technically: their 'phenomenology'), it is unclear what the role of the free will might be. This obscurity is often framed as the problem of determinism: how can there be a free will if physical facts and causal laws completely determine the movements of the body and thereby most human actions? In the causal explanation of human actions, there is no place where the free will can 'interfere'.

4.4 The third solution: ascribe intentionality to some machines

The third solution to the problem of cognitive dissonance is to modify the narrative about machines. Should machines necessarily be treated as the place where 'blind' causal processes

²⁸ Searle, *Minds, Brains and Science*, Chapter 2.

²⁹ Gazzaniga, *Who's in charge*. Free will and the science of the brain; Haidt, *Social Intuitionists Answer Six Questions about Moral Psychology*.

³⁰ For example: Wegner, *The Illusion of Conscious Will*; Wilson, *Strangers to Ourselves*. Discovering the adaptive unconsciousness; Carruthers, *The Opacity of Mind*.

³¹ For elaboration, see Hage and Waltermann, *Responsibility, Liability and Retribution*.

produce some output if fed with the necessary input? Or would it also make sense if at least some causal processes are interpreted in mental terms?

The philosopher Daniel Dennett made an attempt to develop the latter road by distinguishing three ways of looking at causal processes.³² First, it is possible to look at processes from the *physical stance*. Events are from this stance explained in terms of physical and chemical states or events, and causal laws.

There is also a second stance, the *design stance*. This stance is often taken in engineering and in biology. A system in the world (notice: it is already a *system*) is best understood as a means to realize a particular goal, or to fulfil a particular function. The system was ‘designed’ to fulfil this function, and its operation or malfunction is best described in terms of how well it fulfils this function. In this connection, the notion of ‘design stance’ should not be taken too literally. There need not be a real human or divine designer to make this stance useful. The issue is not whether there was a real designer, but whether explanations in terms of the function something fulfils are more fruitful than physical or chemical explanations. To this purpose, design by, for instance, evolution is good enough.

Finally, the third stance is the *intentional stance*. If something is studied from this stance, its operation is explained in terms of its desires and its beliefs about how these desires can, or cannot, be realised. We encountered this stance in the narrative about humans, but the intentional stance can also be used to explain the behaviour of a cruise missile, or a chess playing computer program. The cruise missile ‘wants’ to hit its target, takes in information about, for example, its position and the weather conditions, uses causal laws to compute what it should do, and adapts its ‘behaviour’ to make sure that it will reach its goal. More innocently, the computer program wants to checkmate its opponent and uses its knowledge about the board positions and the rules of the game, and heuristics for selecting strategies, to choose the best moves which may lead to its goal.

Crucial in this account of the three stances is that the intention of something is not a state of the thing itself which comes on top of its physical characteristics, but a way to *interpret* the physical characteristics, including movements. This means that there is no problem in taking the intentional stance towards machines that can also be studied from the physical or the design stance. The identification of the intentional stance provides a way to combine the interpretation of something as a machine and explanation of its ‘behaviour’ in the terminology that fit in the narrative about human beings. It would also allow to consider human beings as a sophisticated kind of machines.³³

4.5 The fourth option: recognize the mystery

All three ‘solutions’ succeed in taking away the cognitive dissonance that results from the existence of ‘intelligent’ computer programs. The first by denying that AIs act: AIs are machines and machines cannot act. The second solution takes the cognitive dissonance away by giving up the narrative about (human) agency. The special nature of human agency is denied and the

³² Dennett, *The Intentional Stance*, Chapter 2.

³³ This ‘solution’ was, in a completely different setting, already propagated by the French physician Julien Offray de La Mettrie, who in 1747 published a book entitled *L’homme machine* (Man a Machine).

narrative of what agency is, is modified to allow AIs to be agents in this more modest new narrative of agency. The third solution takes the cognitive dissonance away by redefining the notion of an intention and showing how some machines are capable of performing intentional actions in this redefined way. Perhaps AIs belong to this category of machines.

Of course, there also is a fourth option – not a solution – and that is to live with the cognitive dissonance caused by AIs. Agency is a mystery, that occurs in the life of human beings, but not in the history of machines.

5 Conclusion

This article consists of two parts. In the first part (Section 2) it was explained what narratives are and how entrenched narratives that are broadly shared can influence our view of the world, including the issue of whether AIs can act.

In the second part (Sections 3 and 4), it was first (Section 3) shown how two narratives, one about (human) agency and the other about machines, make it impossible that machines act. Second (Section 4), it was argued that AIs, some of them being machines that seem to act, cause cognitive dissonance if combined with these two narratives. It was also shown how this cognitive dissonance can be managed, by:

1. applying the conclusion about non-acting machines to AIs,
2. modifying the narrative about agency,
3. modifying the narrative about machines,
and – finally – by
4. living with the cognitive dissonance.

REFERENCES

- Brandom, Robert B. *Making It Explicit*. Harvard University Press, 1998.
- Carruthers, Peter. *The Opacity of Mind*. Oxford University Press, 2011.
- Dennett, Daniel C. *The Intentional Stance*. The MIT Press, 1987.
- Dietz, Jan. *Wat doen computers als ze iets zeggen?*. Inaugural Address Technical University Delft, 1996.
- Festinger, L. *A Theory of Cognitive Dissonance*. Stanford University Press, 1957.
- Gazzaniga, M.S. *Who's in charge. Free will and the science of the brain*. Robinson, 2016.
- Hage, J. "Handelende computers". *Computerrecht* 1997, no 5, 219-223.
- Hage J. "The Nature of Law and Constructivist Facts". *Diritto & Questioni Pubbliche / Recognise* (special issue, edited by M. Brigaglia and C. Roversi), August 2023, 207-224. DOI: https://www.dirittoequestionipubbliche.org/page/2023_dq-recognise/e-03-Hage.pdf
- Hage, J.C. 2024. "How Law's Nature Influences Law's Logic". *Studia Humana* 13 no 3 (2024), 4-17. DOI: 10.2478/sh-2024-0015.
- Hage, Jaap and Antonia Waltermann. "Responsibility, Liability and Retribution". In *Law and Mind. A Survey of Law and the Cognitive Sciences*, edited by B. Brożek, J. Hage and N.A. Vincent, 255-288. Cambridge University Press, 2021.

Haidt, Jonathan and Fredrik Bjorklund. "Social Intuitionists Answer Six Questions about Moral Psychology". In *Moral Psychology, Volume 2: The Cognitive Science of Morality: Intuition and Diversity*, edited by Walter Sinnott-Armstrong, 181-218. MIT Press, 2008.

Honoré, Tony. *Responsibility and Fault*. Hart, 1999.

Kane, Robert (ed.). *The Oxford Handbook of Free Will*. Oxford University Press, 2002.

Kuhn, Thomas S. *The Structure of Scientific Revolutions. 50th anniversary Edition*. The University of Chicago Press, 2012.

Lakatos, I. "Falsification and the Methodology of Scientific Research Programs". In *Criticism and the Growth of Knowledge*, edited by I. Lakatos and A. Musgrave, 91-196. Cambridge University Press, 1970.

McLaughlin, Brian P., Ansgar Beckermann, and Sven Walter (eds.). *The Oxford Handbook of Philosophy of Mind*. Clarendon Press, 2009.

O'Connor, Timothy and Constantine Sandis (eds.). *A Companion to the Philosophy of Action*. Wiley-Blackwell, 2013.

Pereira, Jane Reis Goncalves. "Entrenched Clauses". In: *Oxford Constitutional Law*, DOI: <https://oxcon.oup.com/abstract/10.1093/law-mpeccol/law-mpeccol-e31> (2021).

Popper, K.R. *The Logic of Scientific Discovery*. Routledge, 2002. (First edition 1959).

Quine, W.V. "Two Dogmas of Empiricism". *The Philosophical Review* 60 (1951), 20-43. Reprinted, with minor changes, in W.V.O. Quine. *Two dogmas of empiricism*. Harvard University Press, 1953, 20-46.

Quine, W.V. and J.S. Ullian. *The Web of Belief*. Random House, 1970.

Schechtman, Marya. "The Narrative Self". In *The Oxford Handbook of The Self*, edited by Shaun Gallagher. Oxford University Press, 2011, 394-416.

Searle, John R. *Intentionality. An essay in the philosophy of mind*, Cambridge University Press, 1983.

Searle, John. *Minds, Brains and Science*. Harvard University Press, 1984.

Smith, Nicholas J.J. *Logic. The laws of truth*. Princeton University Press, 2012.

Tavris, Carol and Elliot Aronson. *Mistakes were made (but not by me)*. Mariner Books, 2007.

Tushnet, Mark. "Constitution". In *The Oxford Handbook of Comparative Constitutional Law*, edited by Michel Rosenfeld and András Sajó. Oxford University Press, 2012, 217-232.

Wegner, Daniel M. *The Illusion of Conscious Will*. The MIT Press, 2002.

Wilson, Timothy D. *Strangers to Ourselves. Discovering the adaptive unconsciousness*. The Belknap Press, 2002.