

Explainable Legal AI

Jaap Hage

ABSTRACT

This article aims to show the relevance of the cognitive sciences and of social ontology for law and AI. In this way it tries to embed law and AI research in the broader project of discovering how human reasoning works. It is *not* an attempt to bring existing law and AI technology a little further and it does *not* strive to remain in the existing tradition. On the contrary, it aims at changing the tradition to make it pay more attention to the psychology of legal reasoning and to the social construction of law.

To serve its main purpose, the article uses the example of explainable legal AI, which seems to be problematic in the case of legal AI systems based on machine learning. The present article argues that there are two types of explanation for a judgment or a decision. Type 1 describes the process through which the judgment or decision came about. Type 2 is a tool for generating understanding in an audience. These two types do not always coincide. A true description of how a judgment or decision came about does not always generate understanding, while a narrative that generates understanding does not always give a true description. Although there are reasons why in law type 1 and type 2 explanations may coincide, in the end this turns out not to be the case because legal facts are ‘constructivist’.

1 Preface

This article deals with artificially intelligent legal systems (AILSs), defined as systems that generate judgements about what is legally speaking the case in a particular fact situation. Such judgments can be used to advice citizens who want to be informed about their legal positions, or human legal decision makers who would like advice about what decision to take. They might even function as legal decision makers themselves. Whatever their precise role, AILSs must cooperate with human beings in whose society they must function, and part of this cooperation is that they must explain their judgments.

Especially if AILSs are the products of machine learning (ML-systems), explanations of their outputs may be hard to obtain. But is, in the case of ML-systems, the desire for these explanations realistic? In this article, the question will be answered affirmatively, but not by showing how it is possible to extract explanations from the operation of ML-systems. The article will argue that AILSs can provide explanations of their decisions if we, humans, have realistic expectations about these explanations. To understand why this is the case, and what realistic expectations are, it is necessary to pay attention to the nature of explanation and of law. That is what this article will do, and on the basis of the findings recommendations will be made about the kind of explanation AILSs can and should give. The focus will be on AILSs that generate advice for human users who take authoritative decisions, or AILSs which take legal decisions themselves.

One purpose of this article is to contribute to the discussion about explainable legal AI. However, there is also a second, more important, purpose. That is to illustrate – not to argue – why the cognitive sciences such as cognitive neuroscience, cognitive and social psychology, and even social ontology,

are relevant for foundational issues in AI and Law (Hage 2021). To fulfil this second purpose, the argument will liberally draw from topics that are traditionally dealt with by different disciplines.

The argument will regularly refer to the following example case:

Susan drives at moderate speed through the outskirts of her town. Suddenly Max comes crossing the street, running from behind a parked car. Even her careful driving cannot protect Susan from a collision with Max.

Is Susan (or her insurance) liable for the medical costs of Max?

2 Explainable AI

Legal decisions can have a big impact on the lives of those concerned. This is why the parties in a legal dispute may expect an explanation. As Atkinson et al (2020) write:

‘In a legal dispute there will be two parties and one will win and one will lose. If justice is to be served, the losers have a right to an explanation of why their case was unsuccessful. Given such an explanation, the losers may be satisfied and accept the decision, or may consider if there are grounds to appeal. Justice must not only be done, but must be seen to be done, and, without an explanation, the required transparency is missing. Therefore explanation is essential for any legal application that is to be used in a practical setting.’

Although explanation has of old been an important topic in law and AI (Atkinson et al 2020), it has more recently developed into an important issue in AI in general, since the knowledge in AI systems often has become the result of machine learning. This knowledge is not transparent to human beings and its products in the form of system output cannot be explained to humans in the way it was possible to explain the products of, for instance, traditional expert systems (Moore and Swartout 1988).

This attention for explanation in AI can be found back in recent publications. According to the EU High-Level Expert Group on Artificial Intelligence, ‘AI is not an end in itself, but rather a promising means to increase human flourishing ...’ (High-Level Expert Group on Artificial Intelligence 2019). To realise this goal, the expert group proposes the *principle of explicability* (explainability). Explainability is ‘... the ability to explain both the technical processes of an AI system and the related human decisions (e.g. application areas of a system)’. Moreover, according to the expert group, ‘Whenever an AI system has a significant impact on people’s lives, it should be possible to demand a suitable explanation of the AI system’s decision-making process. Such explanation should be timely and adapted to the expertise of the stakeholder concerned ...’.

In their article *Four Principles of Explainable Artificial Intelligence*, Phillips et al (2021) formulate four principles which AI systems that are intended or required to be explainable, should adhere to:

1. *Explanation*: A system delivers or contains accompanying evidence or reason(s) for outputs and/or processes.
2. *Meaningful*: A system provides explanations that are understandable to the intended consumer(s).
3. *Explanation Accuracy*: An explanation correctly reflects the reason for generating the output and/or accurately reflects the system’s process.
4. *Knowledge Limits*: A system only operates under conditions for which it was designed and when it reaches sufficient confidence in its output.

Principle 1 requires from systems that are intended or required to be explainable that they provide explanations but does not indicate what is expected from these explanations, except that they provide evidence or reasons. Principle 4 does not address explanation, and I will ignore it here.

The principles 2 and 3 formulate the core of what is expected. Principle 2 focuses on the understandability of explanations to the intended consumers of the system, who presumably are human beings. Principle 3 requires that an explanation reflects the reason the system had for generating the output, or that it reflects the system's process. To state it in broad terms, principle 2 requires explanations in terms suited to human users, while principle 3 requires explanation that reflects the system's operation. The same two principles can, with some good will, be recognised in the desiderata concerning explanation that were formulated by the EU High-Level Expert Group on Artificial Intelligence. Are these two principles compatible with each other? Does a good explanation in the case of a decision *both* reflect the operation of a decision maker, be it an AI system or a human being, *and* make the decision understandable for human beings? Or are there two notions of explanation at stake, one notion which relates to explanation to making the output – in our case a decision – understandable for human users, and a second notion which deals with the 'process', be it physical or logical, through which the decision was generated?

3 Two types of explanation

There is no generally accepted view of what explanation is. For instance, the Stanford Encyclopedia of Philosophy has several entries that deal with different kinds of explanation (Alvarez 2017; Brenner et al 2021; Mancosu et al 2023; Ross and Woodward 2023; Woodward and Ross 2021). None of these specifically deals with the explanation of legal decisions, but this topic has been covered by Atkinson et al (2020). The question of whether there are two notions of explanation, as suggested above, has to the knowledge of the present author not received much attention yet.

3.1 Explanation as the description of a 'process'

One important way to explain a fact or event is to describe the process¹ through which the fact or event came about.² This kind of explanation was the topic of a seminal article by Hempel and

¹ Some physical laws deal with simultaneous facts and not with processes. Boyle's law, relating the temperature, the volume, and the pressure of a gas, is a case in point. Moreover, logical relations as those between the facts that match the conditions of a rule, and the conclusion of a rule, are not processes either. Therefore the word 'process' in the name for this kind of explanation was written between quotes. The process in question is not necessarily a physical process. It is rather a mental process, projected on the outside world. This even holds if there is also a real physical process, as the interpretation of the relation between the premises and the conclusion of the deductive argument as cause and effect is the result of a mental operation. When Hempel called this process 'deduction', he may have tried to capture the same idea. Moreover, the use of the notion of deduction has its own problems. For instance, it has often been pointed out (e.g. in Prakken and Sartor 1997) that legal conclusions, or the arguments leading to these conclusions, are defeasible in the sense that the conclusions sometimes must be rejected if one or more premises are added to the original argument. This means that an argument leading to a legal conclusion cannot be deductively valid.

² To avoid repetition of the cumbersome phrase 'fact or event', this article will from here on deal with the explanation of facts. It is assumed that the explanation of events is not different in a manner that is relevant for the purpose of the article.

Oppenheim (1948) which strongly influenced the debate following its publication. In this article, the authors proposed what has become known as the deductive-nomological model of explanation (for short: the DN-model). According to this model, the explanation of a phenomenon consists of a deductively valid argument with two premises. One premise formulates a law, such as ‘Metals expand when heated’. The other premise formulates a fact situation, such as ‘M is a piece of metal that was heated’. From these two premises it can be deduced that M expanded, the *explanandum*.³ If the argument is sound, it provides an explanation of the fact expressed by its conclusion. See figure 1; the similarity to the Toulmin-scheme of argumentation (Toulmin 1958, p. 99) is no coincidence.

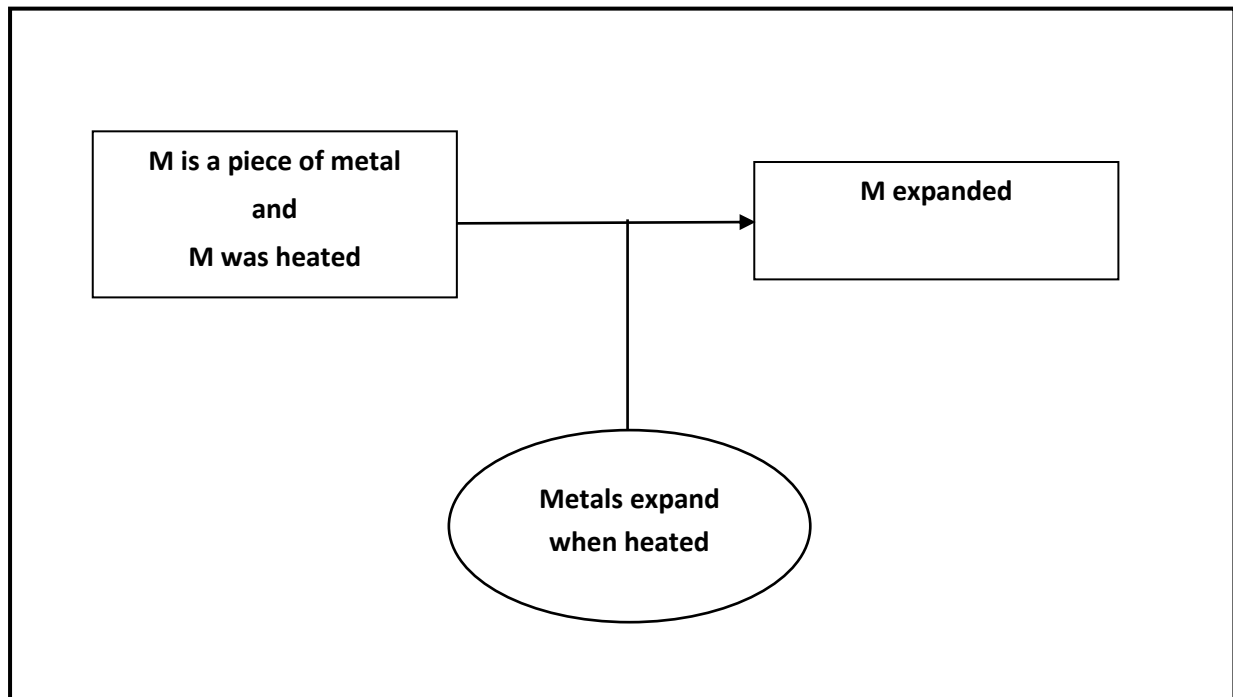


Figure 1: The DN-model of explanation

3.2 Pragmatic explanation

Some explanations in our daily lives seem not to be based on any plausible law (Woodward 2000). For instance, if Max suddenly crosses the street, is hit by a car, and loses a leg because of the accident, the medical costs can be explained from these events, but there is no plausible law that connects these events to the specific medical costs. In a different way, something similar holds for evolutionary explanation. Suppose that Darwin was right⁴ and that the long neck of giraffes can be explained from adaptation to particular opportunities to obtain food. There is no plausible law that predicts or explains that giraffes will develop long necks in order to obtain more food from trees with high branches.

³ The description somewhat simplifies the model and ignores some details that are not relevant for the purpose of this article.

⁴ According to modern insights, Darwin was wrong. See <https://www.scienceabc.com/nature/animals/why-giraffes-have-a-long-neck.html>. Accessed 18 October 2023

These and other complications make it attractive to look for other theories of explanation than the DN-model. Pragmatic theories, which focus on what the explanation brings about rather than on what it describes, form such an alternative. According to Woodward and Ross (2021), pragmatic theories assume that explanation refers to facts about the interests, beliefs, or other features of the psychology of those providing or receiving the explanation and/or irreducible reference to the ‘context’ in which the explanation occurs. When the High-Level Expert Group on Artificial Intelligence (2019) demands that the explanation by an AI system should be timely and adapted to the expertise of the stakeholder concerned, this fits nicely in a pragmatist perspective on explanation. The same holds for the second principle of explainable AI that was formulated by Phillips et al (2021), requiring that explanations are understandable to the intended consumers. These ideas can be summarised in the view that explanation is a means to promote understanding (Brožek 2016).

Although explanations of decisions that are adapted to the expertise of stakeholders or understandable to the intended consumers may at the same time also describe the process through which the decision was made, this will often not be the case. Evidence supporting this claim can be found in the popularity of the belief/desire model for explaining human action (section 3.3) and from the battle of methods in the social sciences (section 3.4).

3.3 Explaining human actions

In the case of an AI-system that produces legal decisions, human users of the system will need an explanation of why legal decisions produced by the system faithfully reflect the law, not of how the system works. For instance, what reasons did the system have for holding Susan liable for the medical costs of Max? These users expect an explanation from the system that is similar to the explanation they would have expected from a human decision maker. Human decision makers typically do not explain their decisions by pointing to the physical mechanisms – e.g. firing neurons – that made them take their decisions.

Explanations of decisions are in this respect similar to the explanation of other actions. Typical actions are intentional, at least in the Western tradition (Piñeros and Tenenbaum 2023). This is reflected in the way they are usually explained, namely by specifying what the agent desired to do or to bring about and his⁵ belief that the action promoted or realised this desire. For example, a court desires (wants) to give a verdict that reflects the legal positions of the parties in a dispute and believes that verdict V reflects the law. This fits in the so-called ‘belief-desire model’ (BD-model) for explaining intentional actions, a model that is traditionally traced back to Hume (Hume 1978; Smith 1994). On this model, the reason why an action was performed is either a belief, or what is desired, or a combination of these two. For instance, if the court holds Susan liable for the costs of Max, this can be explained by the court’s belief that this is the law, by the desire of the court to apply the law, or by a combination of these two.⁶ See figure 2.

⁵ To avoid cumbersome formulations such as ‘(s)he’ or ‘her/his’, to emphasize gender-neutrality, the present author uses his own (male) gender to refer to persons whose gender does not follow from the main text. He encourages non-male authors to follow the same policy.

⁶ See in this connection Raz’s distinction between complete and incomplete reasons (Raz 1999, pp. 22-25).

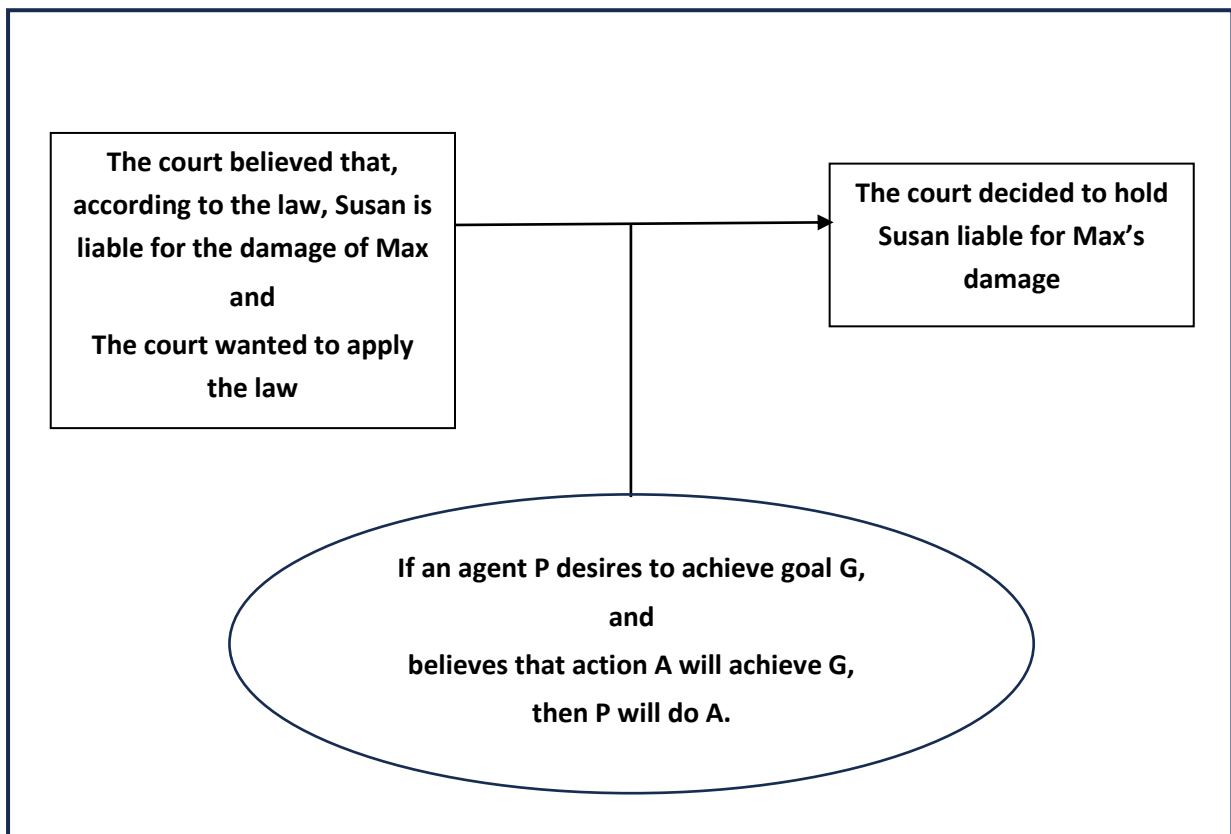


Figure 2: The belief-desire-model for the explanation of intentional action

It is well possible that an action is best explained by a false belief, as when a person takes an umbrella with him because he falsely believes that it is raining. It is also possible that an action is motivated, and can be explained, by a wrong desire, as when a terrorist's bombing a mass event is motivated by his desire to call the public's attention to his cause⁷. However, even if an action was motivated by a false belief or a wrong desire, the agent held the belief or had the desire, and the action was *in his eyes* the rational thing to do.

This is why there is a close connection between the explanation of intentional actions and their rationalization and why the explanation of an action often resembles the rationalization of the same action.⁸ For instance, if a court holds Susan liable for the damage of Max, the court normally believes that Susan is liable according to the law. An explanation of the court's decision will therefore typically include the argument employed by the court to argue why Susan is liable according to the law. This

⁷ The 'wrongness' of this desire does not necessarily lie in the desire itself (drawing attention to a cause) but in the suitability of this desire to justify the action. Some desires may be wrong in themselves, such as, perhaps, the desire to kill as many human beings as possible.

⁸ Instead of the word 'rationalization' that is used in the main text, the word 'justification' can also be used. In that case, there is a close connection between the explanation and the justification of actions. However, the word 'justification' is somewhat infelicitous if the belief on which the justification is based is false, or if the goal or the desire is misguided. Therefore the main text prefers the term 'rationalization'. By the way, the same idea is captured by the claim that human 'agents act for reasons, where reasons are the ends pursued by the agent as practically intelligible goods'(Frey 2019/20).

argument also rationalizes the decision that Susan is liable. More in general, people tend to explain their actions by pointing out that these actions were rational (in their eyes).

3.4 Empathetic understanding

Given the success of the empirical sciences since the 16th century, it was tempting to apply the empirical method to the social sciences as well. This gave rise to the idea that the behaviour of individual persons, but also of social groups, should be explained by means of general laws, as in the DN-model (Hempel 1965, chapter 9; Rudner 1966, p. 63; Kincaid 1994). However, there also was a strong counter-movement that emphasized the roles of meaning and understanding in the social sciences. For example, if you want to explain why people gather at the outskirts of a town, all dressed in black, it is hard to find a physical law by means of which this behaviour can be explained. Explanation becomes easier if you know that the people were meeting at the occasion of a funeral and that black is the colour to express their mourning. Insights like this inspired for instance the sociologist Weber (1922) to assign a central place to *Verstehen* (empathetic understanding) in his sociological method. More recently, Fay (1983) has proposed to explain human behaviour by means of general laws formulated in terms of beliefs and desires etc, where empathetic understanding is no doubt a heuristic tool to find the appropriate concepts to formulate adequate laws.⁹

On this last approach, the two types of explanation coincide. On one hand, explanations are still given in terms of the ‘process’ through which behaviour came about. On the other hand, the behaviour is made understandable to other human beings – and perhaps also to the agent himself (see section 4.4) by explaining it in terms of beliefs and desires. In that case, the description of the ‘process’ leading to an action makes the action understandable for human beings. The first type of explanation has then become a means to perform explanation of the second type.

3.5 Explanation as a means to create coherence

If explanation as ‘process’-description can be interpreted as a means to explain in the sense of promoting understanding, it is important to have a better insight into what this second kind of explanation amounts to. If we do not understand why this piece of metal expanded, our understanding may be increased by knowing that metals expand when heated. If Susan does not understand why she is liable for the damage of Max, she can be helped by pointing out that car drivers have a strict liability for the damage of non-drivers. However, there is no generally accepted conception of what understanding is (Grimm 2021), and we cannot start from a clear view of what understanding is to get a grip on the derived notion of explanation.¹⁰ The theories of explanation and of understanding must be co-developed.¹¹

⁹ Gardner(1987), chapter 5, describes the development in psychology leading from behaviourist attempts to explain human behaviour without factoring in mental states to cognitive psychology which attempts to formulate laws in terms of mental states. This can be seen as the psychological counterpart of the debate in sociology on the explanation of social behaviour.

¹⁰ There even exist different notions of understanding. For instance, understanding a text means that one has extracted the information contained in that text.

¹¹ The theory of understanding that is briefly described here is an internalist coherentist theory in the sense of Grimm 2021, with an interpretation of coherence as ‘integrated coherence’ (see Hage 2005a, 2013, 2016).

Our point of departure in this connection is the notion of a recognition set. This is the set of everything which an agent recognises. ‘Recognise’ is a technical term here, which stands for ‘believes’ in the case of descriptive sentences, for ‘uses’ in the case of rules, ‘holds’ in the case of values, ‘pursues’ in the case of goals, etc.¹² For our present purposes, it suffices to consider only beliefs and rules. For instance, a court believes that Susan caused a car accident and that Max suffered damage because of this accident. The court also uses the rule that if a person causes a car accident, he is liable for the damage caused by that accident. Moreover, the court uses an inference rule for drawing conclusions by means of rules.¹³ Given this inference rule, the rule about liability, and the belief about the car incident, the court rationally ought to recognise that Susan is liable for the damage of Max.¹⁴

However, the court is not necessarily completely rational and perhaps in this case it does not apply the rule which it normally uses. So, it may be the case that the court’s recognition set does not include the belief that Susan is liable for the damage of Max, although rationally the set ought to include that belief. If this is the case, the recognition set is incoherent. Informally, a recognition set is coherent if it includes everything that it rationally ought to include and nothing that it rationally ought not to include. The standards for what rationality involves in this connection derive from the recognition set itself (Hage 2013).¹⁵

A recognition set is in the first place a psychological phenomenon, as it is determined by what a particular agent recognises.¹⁶ Since agents are not always completely rational, chances are that their recognition sets are incoherent. The incoherence of one’s recognition set will often go unnoticed, and then the recognition set will not include the belief that the set is incoherent. However, if some incoherence is noticed by the owner of the set, the set will contain the belief that it is incoherent itself. If this happens, the recognised incoherence will typically cause mental discomfort. For instance, ‘I should not discriminate against blue-eyed persons, but I often do so nevertheless’. This discomfort, in combination with the incoherence of one’s recognition set that causes it, may be called ‘cognitive dissonance’ (Festinger 1957). If cognitive dissonance is the result of beliefs that do not sit well together and if the owner of the recognition set believes this to be the case, the dissonance constitutes a feeling of non-understanding (or misunderstanding).

People who experience cognitive dissonance will typically try to take it away by making their recognition sets coherent, at least to the extent that they recognise. If the dissonance is the result of a lack of coherence, explanation of one or more of the believed facts can do the job. For instance, Susan does not believe that she is liable for the damage of Max because she had no fault in the accident. Nevertheless she finds that a court holds her liable. At first, Susan does not understand why the court could do this and this leads to mental discomfort, if not indignation. However, in the

¹² The point of using the single word ‘recognise’ is not to create seeming uniformity by using an ambiguous word, but to emphasize that believing, using rules etc. are physiologically speaking not very different things. This almost-identity is for instance reflected in the phenomenon that some beliefs ‘automatically’ go together with a disposition to act. See also section 8.3.

¹³ The use of special inference rules for reasoning with legal rules and the corresponding treatment of these rules as logical individuals has been a central topic in Hage 1997 and 2005b. For an application to international law, see Marcos 2023.

¹⁴ The relation between rules and rational recognition is elaborated in section 8.5.

¹⁵ Technically speaking, this means that a coherent recognition set includes its own meta-recognitions and is in that sense closed.

¹⁶ It is possible to define a sociological counterpart as what is broadly recognised in a social group, analogous to World 3 in the sense of Popper 1979. This possibility will be discussed briefly in section 8.1.

justification of its verdict, the court explained its decision by pointing out that car drivers have a very strict form of liability. When Susan reads this explanation, she comes to understand the court's decision, and perhaps overcomes her cognitive dissonance.

Explanation as a means of creating or restoring coherence of a recognition set consists in pointing out connections between facts. In the example above, the explanation by the court would consist in pointing out the connection between the car accident and the liability of the car driver, a connection that is based on a rule of liability law. The explanation of why a piece of metal expanded consists in pointing out the connection between metal being heated and expanding. This connection is based on the physical law that metals expand when heated. However, the connection between facts that plays a role in explanation does not have to be a law or rule that exists objectively (Van Fraassen 1980, 100). To be efficacious, the connection should be based on the recognition set of the agent at whom the explanation is addressed. In other words, the physical law does not have to be true, or the legal rule does not have to be valid. It suffices if the person to whom the explanation is directed believes the law or holds the rule to be valid. If that is the case, the explanation will be efficacious. If an explanation is efficacious, it modifies the recognition set of the addressee and takes away the cognitive dissonance that constituted the lack of understanding.¹⁷

Explanation in the sense of describing how a fact came about adds beliefs and/or laws or rules to a recognition set, with the purpose of creating the belief of coherence. This shows how explanation of the first kind can be a means toward explanation of the second kind. It would be attractive if it were in general the case that explanation in the sense of showing how the explanandum came about and explanation as a tool to create understanding coincide. Removing cognitive dissonance would then be what explanation is, and describing how the explanandum came about would be the means through which cognitive dissonance is lifted. Type 1 explanation and type 2 explanation would not be different kinds of explanation, but 'only' two sides of the same explanation coin.

3.6 Summary on explanation and understanding

Both the EU High-Level Expert Group (2019) and Phillips et al (2021) implicitly distinguished between two kinds of explanation:

1. Explanation as a description of the 'process' that generated the explanandum. The traditional form of this kind of explanation is described by the DN-model as propagated by Hempel. In the case of

¹⁷ We can distinguish two notions of understanding. One notion is what might be called 'real' understanding. A person who 'really' understands a phenomenon gives this phenomenon the 'right' place in his web of beliefs (whatever the right place might be) and derives psychological satisfaction from this. The other notion emphasizes the psychological satisfaction and treats understanding as the *feeling* that something has a place in the understander's web of belief. This feeling can go together with misunderstanding (in the first sense) of what is 'really' the case.

It may be objected here that efficacious explanation may lead to a misguided feeling of understanding, as when a bad pupil who passed his exam explains this by having donated some money to his favourite church. Such an objection would be based on the difference between understanding as a mental state of a single person and understanding as a social phenomenon. An explanation that convinces one person but does not convince the people at large may lead to an understanding that is not broadly recognised as such and which therefore counts as a 'misguided feeling of understanding'. 'True understanding' is a social phenomenon, and not the kind of understanding that reflects a phenomenon as it objectively is.

ML-systems, the first kind of explanation is difficult – if not impossible – to obtain, especially if the explanation must also create understanding.

2. Explanation as a means to create understanding in the intended audience. The second kind of explanation is realised by traditional explanations of human behaviour, in terms of desires (in a broad sense) and beliefs. The recognition of the beliefs and desires that explain a particular action can often be accomplished by a form of empathetic understanding, referred to by the German word *Verstehen*. Explanation and understanding based on this *Verstehen* is in the social sciences often propagated as an alternative for the DN-model.

It is possible to connect explanation as a tool for creating understanding with a coherentist theory of recognition (amongst others of belief) and with the psychological notion of cognitive dissonance. A lack of understanding is then the incoherence of somebody's recognition set, which is recognised by this person. The result is a form of cognitive dissonance, a kind of mental discomfort. The purpose of explanation is to provide the person who suffers under this dissonance with beliefs (and other forms of recognition) to the effect that the incoherence goes away or is at least believed to be overcome. As a result, the cognitive dissonance is removed, and understanding is restored. Explanation as 'process'-description, type 1 explanation, is then a means to provide type 2 explanation, explanation as a tool for creating understanding. Can this coincidence between the two types of explanation be upheld? In the next section we will see that the two types of explanation do not always go together and can therefore not be two sides of the same coin.

4 Can human beings explain their own behaviour?

People often think to know through introspection what motivated their decisions. If we ask somebody why he took a particular decision, introspection usually leads to reasons that explain the decision. These reasons will typically resemble a rationalization: given the facts that the decision-maker believed to be the case, and given what he wanted to accomplish, it was rational to take the decision and that explains why the decision was taken. However, it may be doubted whether the reasons that support the decision and which are consciously accessible also play a role in the process leading to the decision. There have been many psychological experiments showing that people are often unaware of the causes of their decisions.¹⁸ In other words: people often do not know what moved them. In this section, several examples from the vast psychological literature will be discussed. As a conclusion of this second part of the argument, Gazzaniga's explanatory framework for rationalizations will be described.

¹⁸ Summaries of these experiments and their interpretations can be found in Wegner 2002, Wilson 2002, Bargh 2017 and – in Dutch – Dijksterhuis 2016. For an extensive and more abstract argument to the effect that people do not have introspective access to what caused their beliefs or motivated their actions, see Carruthers 2011.

4.1 Selecting panty-hoses

In the 1970s, Robert Nisbet and Timothy Wilson conducted an experiment in Meijer's Thrifty Acres, a bargain store just outside Ann Arbor.¹⁹ On a busy Saturday morning, they placed a sign on a table, inviting passing customers to evaluate four panty-hoses that were put on display in a row. Passers-by were asked which panty-hose they preferred and why. From earlier research, it was known that most customers preferred the right-most product, and also in this case that turned out to be the case. Pair A, the left-most, was preferred by 12%, pair B was preferred by 17%, pair C by 31%, while pair D, the right-most, was favoured by 40% of the customers. Although the pairs were identical and the difference in popularity could be explained from the order of the panty-hoses, most people gave other reasons for their preference, such as the (imaginary) superior knit, sheerness, or elasticity of the preferred panties. With one exception, all customers denied that the position of the panty-hoses had influenced their choice, and the exceptional customer said she was taking three psychology courses. Nisbet and Wilson concluded from this experiment that even if people 'are completely cognizant of both stimulus and response, they appear to be unable to report correctly about the effect of the stimulus on the response'. Also based on related research, the conclusion is more general:

'... when people attempt to report on their cognitive processes, that is, on the processes mediating the effects of a stimulus on a response, they do not do so on the basis of any true introspection. Instead, their reports are based on a priori, implicit causal theories, or judgments about the extent to which a particular stimulus is a plausible cause of a given response. This suggests that though people may not be able to observe directly their cognitive processes, they will sometimes be able to report accurately about them. Accurate reports will occur when influential stimuli are salient and are plausible causes of the responses they produce, and will not occur when stimuli are not salient or are not plausible causes' (Nisbet and Wilson 1977).

In short: introspection does not necessarily reveal the causes of one's behaviour and an explanation of why one did something should draw from other knowledge sources.

4.2 Hungry judges

Danziger, Levav and Avnaim-Pesso investigated the decisions of eight Jewish-Israeli judges in parole cases, taken in the first decennium of the 21st century (Danziger, Levav and Avnaim-Pesso 2011). The decisions were binary: accept the parole request or refuse it. It turned out that the percentage of acceptances dropped during a period in between meals of the judges, decreasing from in between 60% and 70% immediately after a meal to almost 0%. After a meal, the percentage of acceptances jumped back to the original higher level. The cause of this decrease of the percentage of acceptances was not determined, although the authors speculate about the phenomenon of 'self-depletion'.

For the present argument, the precise cause does not matter. It is assumed here that the judges involved were not aware of the fact that the lapse of time since their last meal had a considerable influence on their decisions. It is also assumed that the judges motivated their decisions by other factors than this lapse of time – the article provides little information about the motivations – and that the given motivations were therefore not faithful type 1 explanations of why the judges took their

¹⁹ The account of the experiment has been summarized from Wilson 2002, 102f. The experiment is also described in Nisbet and Wilson 1977.

decisions. Given these assumptions, the research illustrates that also in legal cases decision makers can be moved by factors of which they were not aware.

4.3 Anchoring

A second illustration of the phenomenon that also legal decision makers do not always know what motivated them is in the presence of the so-called 'anchoring bias'. The anchoring bias involves a situation in which people must answer a question, or take a decision, that involves quantities. They tend to let their decision be influenced by a quantity that was planted in their mind just before (that they were 'primed' with). This influence occurs even if the primed quantity is completely irrelevant for the question or decision and is known to be irrelevant.

For example²⁰, a wheel of fortune was rigged to make 10 or 65 the only possible outcomes. After a spin of the wheel – which obviously gave 10 or 65 as a result – students in the University of Oregon were asked whether the percentage of African member states of the UN was larger or smaller than the result of the wheel. This primed the students with either the number 10 or the number 65. After that, the students also had to answer the question of what is, in their opinion, the percentage of African members of the UN. Students who had 10 as the outcome of the wheel estimated the percentage of African members states on the average to be 25. However, students who were primed with the number 65 estimated the percentage of African members states on the average to be 45. Somehow, the student's estimations were 'drawn' towards the outcome of the wheel of fortune, even though the students were very much aware that this outcome was not relevant for their estimation. The result of the wheel functioned as an 'anchor' for the judgment about the percentage of African UN member states. This anchor biased the judgment of the students, and that explains the name 'anchoring bias'.

In their article *Inside the Judicial Mind*, Guthrie, Rachlinski and Wistrich (2001) describe, amongst others, an experiment in which the anchoring effect was tested on judges.²¹ The judges were presented with the following case:

'Suppose that you are presiding over a personal injury lawsuit that is in a federal court based on diversity jurisdiction. The defendant is a major company in the package delivery business. The plaintiff was badly injured after being struck by one of the defendant's trucks when its brakes failed at a traffic light. Subsequent investigations revealed that the braking system on the truck was faulty, and that the truck had not properly been maintained by the defendant. The plaintiff was hospitalized for several months, and has been in a wheelchair ever since, unable to use his legs. He had been earning a good living as a free-lance electrician and had built up a steady base of loyal customers. The plaintiff had requested damages for lost wages, hospitalization, and pain and suffering, but has not specified an amount. Both parties have waived their rights to a jury trial.'

²⁰ The present discussion, and all examples are based on Kahneman 2012, 119-128, which contains references to more specific publications.

²¹ The article also refers to publications in which the anchoring effect was shown to occur in law-related judgments made by non-lawyers.

The investigators randomly assigned judges either to a No Anchor condition or an Anchor condition. The judges in the No Anchor group were merely asked how much they would award plaintiff in compensatory damages. The judges in the Anchor group were informed that the defendant had moved for dismissal of the case as it would not meet the jurisdictional minimum of \$75,000 for a diversity case, thereby introducing the amount of \$75,000 as an anchor. It turned out that the introduction of the anchor had a large effect. The sixty-six judges in the No Anchor group would, on the average, award plaintiff \$1,249,000 compensation, while the fifty judges in the Anchor group merely awarded an average of \$882,000.

Again presumably, the judges would not be aware of this anchoring effect and would not mention the anchor as a part of the explanation of their judgments. Therefore this experiment can also be interpreted as evidence that in the legal domain, agents do not always know what caused their behaviour.

4.4 The interpreter

In this connection, the notion of the 'interpreter' as it was introduced by Gazzaniga (Gazzaniga 2016, chapter 2; Gazzaniga et al 2019, p. 152), plays a role. The idea that there is such an interpreter was the outflow of experiments that Gazzaniga and his colleagues performed on so-called 'split-brain-patients'. Human brains consist of two halves, the left and the right hemisphere. These two hemispheres are normally connected by a bundle of neurons, the *corpus callosum*. In very severe cases of epilepsy, the corpus callosum is cut through, to prevent epileptic seizures from moving from one hemisphere to the other. As a result of such an operation, the two hemispheres cannot communicate normally with each other anymore. This allows for some experiments that address the different functions of the two hemispheres.

One kind of experiment consists in presenting the right hemisphere with information and to ask a verbal report on this information. As the left hemisphere controls a person's speech centre, the verbal report has to be provided by this left hemisphere. Since this hemisphere does not dispose over the information presented to the right hemisphere, it must find an alternative way to generate the desired verbal report. It succeeds in doing so by inferring and confabulating the required information on the basis of what *is* available to it. An example is that a split-brain patient called PS was, in a manner only the right hemisphere could observe (e.g. by presenting the command to the left eye only), commanded to stand up. PS stood up and, when asked why, he could not mention the command as the reason because this information was not available to his left hemisphere. Instead he explained his behaviour from his desire to get a Coke (Gazzaniga et al 2019, p. 152; Wegner 2002, p. 182).

Apparently, agents can 'explain' their behaviour in the absence of introspective knowledge of what moved them. They deal as well as possible with a request for explanation ('Why did I stand up without a clear reason?'), even if the explanation they come up with seems implausible to outsiders who know more than the agent. This brings us back to the observation by Nisbet and Wilson that accurate reports on what moved a person will only occur when this person has access to salient and plausible causes of their behaviour. Based on the findings of Gazzaniga, we should add that the relevant hemisphere

should have this access. The explanation of a person's speech behaviour has moved from the person as a whole to the left hemisphere of his brain.²²

But how can the left hemisphere produce an explanation if it has no access to the right hemisphere which contains the relevant information? According to Gazzaniga (2016), the left hemisphere contains a module that generates the explanations a person gives of his own behaviour. Gazzaniga called this module the 'interpreter'. Proponents of the modularity of mind such as Gazzaniga, Fodor (1983) and Cosmides and Tooby (1992) assume that the human minds²³ contain a large number of modules which are connected to each other, but which function to a large extent autonomously. Some of these modules produce conscious phenomena, but there is no central 'I' to which this consciousness belongs or that coordinates the cooperation of the modules. This raises the question of where a person's feeling of unity and central control comes from. Gazzaniga's answer is that these are produced by the interpreter module, generated by the left hemisphere. This interpreter takes clues from other modules in the brain, including modules that process sensory information, and weaves everything into a coherent narrative. This narrative provides the explanation a person will give of his own behaviour or beliefs. The explanation may coincide with explanations given by other persons, including scientific explanations, but there is no guarantee that this will be the case. It depends on the reliability of the input that the other mental modules give to the interpreter (see also Carruthers 2011). It is – in the terminology that is fashionable now – a case of inference to the best explanation (Douven 2021). What is in this connection 'best' depends on what is available to the interpreter module. As Nisbet and Wilson (1977) pointed out, the accuracy of the explanation created by the interpreter depends on what the other mental modules provide as input. Gazzaniga's view about the modularity of mind and the role of the interpreter provides a theoretical framework in which the findings of Nisbet and Wilson fit.

5 Where we are and where we will go

With the arrival of ML-systems, the demand for explanation of their outputs has become more urgent: artificial intelligence should remain explainable. However, we have seen that the very notion of explanation is not clear and that the DN-model propagated by Hempel, has come under attack from authors who favour a pragmatist approach. In the eyes of the 'pragmatists', explanation is a communicative tool which, the present article argues, aims at removing cognitive dissonance. On this pragmatist approach, explanation does not have to mirror law-based relations in the external world but should fit the information needs of the intended audience. In the discussion of Gazzaniga's theory of the interpreter we saw that the narratives constructed by the interpreter, including explanatory narratives, are not always true. This last result is supported by the research of Nisbet and Wilson (panty-hoses), Danziger et al (hungry judges), and Kahneman and Tversky, and Guthrie *et al*

²² Theories of embodied cognition (Shapiro and Spaulding 2021) suggest that focus on (a part of) the brain may be too limiting. That may be correct but does not address my main point that explanations of behaviour may deal with parts of the body, rather than with the full person. Behaviour is ascribed to a person (Bennett and Hacker 2003, 2f.) but can often be explained by the operation of a part of the body.

²³ Notice the shift from 'brain' to 'mind'. The brain is a physical entity that plays an important role in explaining the physical aspects of behaviour. The mind is a product of, amongst others, the brain; it is not physical. Mental modules are not physical parts of the brain, but functional aspects of the mind. For example, a module processes input from the eyes.

(anchoring). These findings have several implications for the theory that type 1 explanation and type 2 explanation are two sides of the same explanatory coin.

First, even our own minds produce type 2 explanations ('narratives') that satisfy ourselves, but which are not always true.

Second, when it comes to explaining behaviour, type 2 explanations are formulated in terms of beliefs and desires and other, related mental states. At the same time, type 1 explanations are often – but not always – formulated in non-mental terms, such as firing neurons, the level of glucose in one's blood (hungry judges), or seemingly irrelevant information (e.g. primed numbers, in the case of anchoring). This means that the explanations of type 1 and type 2 do not 'match'.

Third, if we try to formulate type 1 explanations in terminology that corresponds to the terminology of type 2 explanations by formulating laws that make use of mental terminology, it becomes unclear how the explaining facts can lead to the explained behaviour. For instance, we can try to explain why Susan drove her car through the outskirts of the town by mentioning that she wanted to pick up her 6-year old son Michael from school and her belief that driving through the outskirts would bring her to the school. Such an explanation would make us understand why Susan drove at the place where Max crossed the street. However, it would be quite hazardous to predict from Susan's desire and her belief that Susan would be at the place of the accident. Such a prediction would come out false too often to justify the adoption of the law that Susan will do what she believes will promote her desires. Briefly stated: it is very difficult to formulate solid laws – laws that not only allow for explanations but also make predictions possible – in terms of beliefs and desires.

If we expect from AI-systems a capability to give explanations that are true descriptions of how they arrived at their output and which are at the same time efficacious in creating understanding, we impose on these systems requirements which not even human decision makers can satisfy. It seems that we can expect true type 1 explanations, or satisfactory type 2 explanations, but not explanations which are both satisfactory for human users and also true process descriptions.

Should we therefore conclude that the demand for explanation that is made on artificially intelligent systems is unrealistic, as not even humans can satisfy it? That conclusion would be hasty, because there are many situations in which human decision makers can explain why they took a particular decision. Such an explanation is not on the level of brain processes, but on the level of arguments that were used in dialogues with other human beings (e.g. judges in a court), or in a *dialogue interieur* as when a single judge considers different arguments in his mind.

The argument for this view may be along the following line. A court wants (desires) to apply the law. If the court believes that the law is X, for instance that Susan is liable for the damage of Max, this explains why the court decides X. In this argument there is much room for argumentation about the content of the law, as this argumentation determines what the court believes the law to be. So, the reasons for and against a particular view of what the law is in a concrete case determine what the court believes and – together with the court's desire to apply the law – this belief explains the decision of the court. Moreover – and this is a crucial addition – if the arguments that determine the court's belief about the law also influence, or even determine, the view on the law of the court's audience, the court will be able to explain its decision in a manner that is both a type 1 and a type 2 explanation. In the following sections it will be argued that this argument is to a large extent correct, but not completely. It is to a large extent correct because law is a social phenomenon. (The argument

therefore does not apply to the explanation of purely physical phenomena.) It is not fully correct, however, because the social phenomenon law exists not as a matter of conventional fact but as a matter of constructivist fact. (This distinction will be explained in section 8.) Ultimately, the conclusion must therefore be that the ideal situation in which type 1 explanation and type 2 explanation coincide does not occur in law. At the end of this article, it will be argued what this implies for artificially intelligent legal decision making systems that are the product of machine learning.

But first, let us see why the social nature of law almost makes law explainable by type 1 and type 2 explanations as two sides of the same coin. The argument why this is the case leads us from Haidt's social-intuitionist model of judgment, via dual process theories of mind, to theories of how social reality – including the law – exists and their implications for the explanation of legal decisions.

6 Haidt's social-intuitionist model of judgment

In an influential article, 'The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment', Jonathan Haidt (2001) described the 'social intuitionist model' (SI-model) of judgments.²⁴ In this model, Haidt attempted to combine the psychological findings that were partly discussed above, into a coherent narrative.²⁵ The model is easiest explained by means of the diagram that Haidt presented to this purpose. See Figure 3.

If a person, let us call her Alice, is presented with a fact situation (the eliciting situation), this causes in first instance an intuition in her. In this connection, an intuition is a belief²⁶ that arises through an unconscious process and for which the believer has no reasons. The intuition may lead to an action of Alice, without even becoming conscious. For instance, if the fact situation is the presence of a snake, Alice's preconscious intuition of a dangerous situation may make her withdraw her hand without prior realization why.²⁷

However, the intuition may also lead to a *conscious judgment*: there is a snake over there and the situation is dangerous. If this step is taken, there are three changes. First, the judgment is conscious, in contrast to the unconscious intuition. Second, the judgment is articulated in language. This articulation in a language already introduces a social aspect into the judgment, as languages are social phenomena. And third, the intuition is mentally subdivided into a belief ('there is a snake') and a disposition to act ('withdraw your hand').

Only if the intuition has become a judgment, can it become the object of critical reasoning. If that happens, the reasoning process may lead to a change of judgment (the dotted line in Figure 1). For instance: 'There is a snake, but it is not poisonous, so there is no danger'. However, according to Haidt, the reasoning process – if there is one – is often not critical at all. Typically, the reasoning process

²⁴ The addition that it concerns moral judgment is not very relevant, as there is little reason why the model should not be applicable to other kinds of judgments about concrete fact situations, including legal judgments.

²⁵ See also Haidt 2012 and Haidt and Bjorklund 2008.

²⁶ In his account of the SI-model, Haidt focuses on beliefs, but in the present author's opinion, the model can also be applied to other recognitions, such as rules.

²⁷ The 'mix' of a factual belief and a disposition to act is mentally speaking one thing (the intuition). This unity of belief and disposition is why the term 'recognition' is preferred in this article over the more traditional 'belief'.

provides the judgment with a rationalisation. For instance: perhaps the snake is not poisonous, but better safe than sorry (rationalizing both the disposition to withdraw one's hand and the belief that is wise to do so). This fits with Gazzaniga's view that the function of the interpreter is to make sense of the inputs it receives and not to search critically for independent truth.

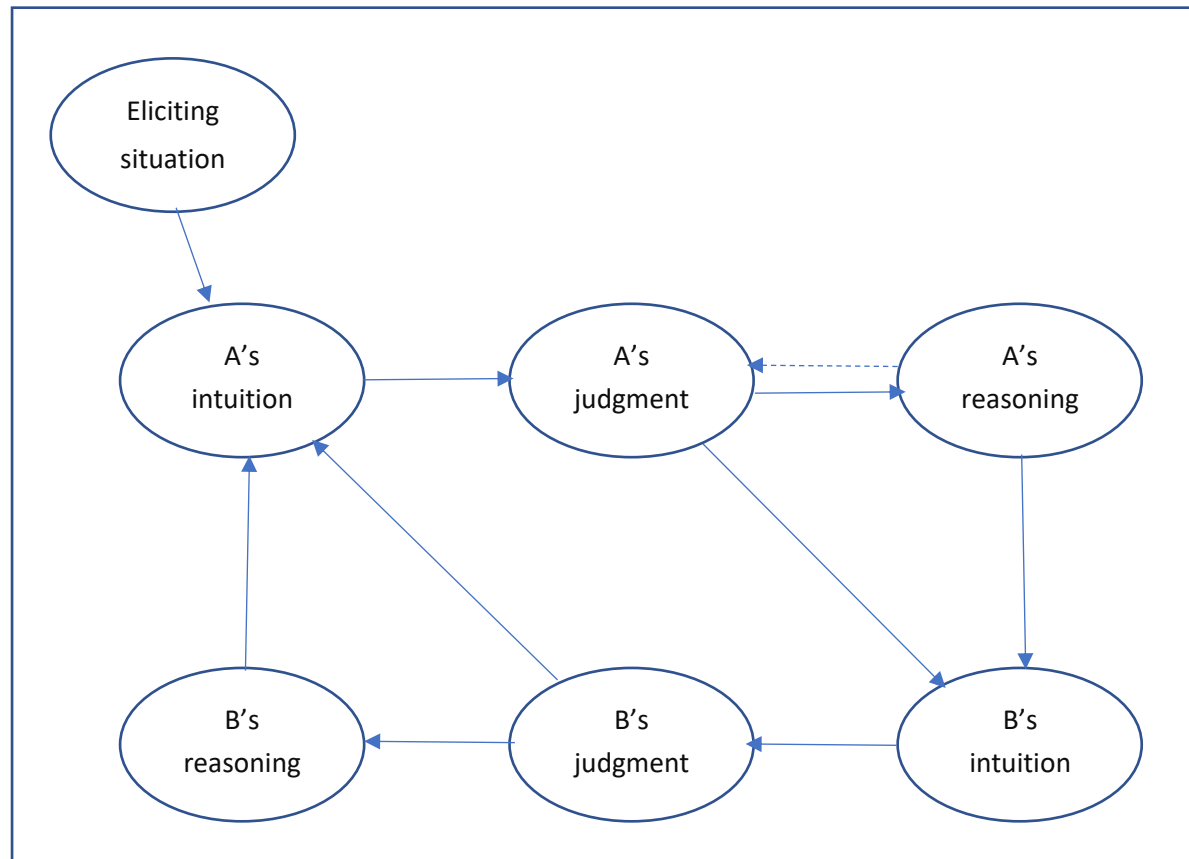


Figure 3: The social intuitionist model of judgments

People do not operate in a vacuum: they communicate with each other. Alice may inform Bernadette about her judgment: 'Be careful. The snake may be dangerous'. And she may present reasons to convince Bernadette that the snake may be dangerous (it is poisonous), or that this is unlikely (this kind of snake never bites). As Mercier and Sperber (2007) argued extensively, reason-giving is first and foremost a means of persuading others or oneself and not a way to show that a particular thesis – e.g. that the snake is dangerous – is right or wrong. Giving reasons is in first instance a means of persuasion, not of justification.

Bernadette processes communications from Alice in the first place unconsciously and this leads to an intuition. The intuition may lead to a conscious judgment and Bernadette may formulate her own reasons why this judgment is correct. Both this judgment and this justification may be communicated back to Alice and influence Alice's intuition. Possibly, this restarts the whole process again.

Haidt emphasized that the chance that the intuition of A will be modified is bigger if B communicates a contrary view to A, than if only A's reasoning process provides the feedback. Argumentation as a means of modifying opinions is in the first place a social, not a solipsistic, process.

Thus far Haidt. In the next section we will see how a ‘solipsistic process’ may be less solipsistic than it may seem at first sight. Moreover, as we will see in section 8, the judgments of interacting people influence each other, and sometimes the process will give rise to a shared pool of judgments and – perhaps surprisingly – a shared pool of social facts. Together these proposed additions to Haidt’s SI-model provide a reason why in the case of judgments about social facts, including judgments about law, type 1 explanations and type 2 explanations can still be two sides of the same coin.

7 Dual process theories of mind

There is an important approach to beliefs and decision making, which is based on the so-called ‘dual process theories of mind’ (Evans and Stanovich 2013). According to dual process theories, there are two different processes of how the mind arrives at beliefs or decisions. Kahneman (2012) popularized the terms ‘System 1’ and ‘System 2’ for these processes.

System 1 operates unconsciously, fast, and is not under explicit control. Think, for instance, of intuitive value judgments, or the unexplainable feeling that there is a danger lurking. Sometimes, system 1 produces good results, as emphasized by, amongst others, Gigerenzer (2007). On other occasions, it leads to less rational results, as emphasized by, amongst others, Kahneman and Tversky (Kahneman and Tversky 1982; Kahneman 2012).

System 2 operates consciously; it is slower, and subject to explicit control. Doing complicated arithmetic would be an example (for most persons), as well as planning a long and complicated journey. Although not often mentioned in the literature,²⁸ multi-agent decision making would also be an example of system 2 at work, especially if the multiple agents are treated as a single decision maker. Decision making by a court of law consisting of more than one judge would be a good example in this connection.

If a decision maker is asked how he arrived at his decision, the system 2 process leading to his decision is accessible to introspection. In the case of a court decision, there may even exist a written account of the process leading to the final decision.

Researchers who find that people often cannot properly explain their own judgments or decisions, will typically point to judgments or decisions that were produced by system 1. Researchers who mention examples of judgments or decisions that can be properly explained will typically point to products of system 2. It is tempting to see this as evidence that the researchers have investigated different systems and as a result arrive at seemingly conflicting conclusions.

The opposition of systems 1 and 2 derives support from the finding that some judgments can be explained through introspection while other judgments cannot. However, rather than claiming that dual process theories of mind support the view that some decisions can be explained through introspection or the other way round, it is more cautious to claim that the findings about the (im)possibility to explain decisions through introspection cohere with the dual process theories of mind.

²⁸ The present author does not know an example in which multi-person decision making is mentioned as an example of the operation of system 2.

Somewhat speculatively, it is possible to expand the role of system 2. It is usual to treat the systems 1 and 2 as two aspects of a single person's mind. If we follow Gazzaniga and others and assume that minds consist of multiple modules (see section 4.4), system 2 processes may be interpreted as forming judgments through the interaction of several modules. This would explain why system 2 processes take more time than system 1 processes. It would also make it better understandable why system 2 processes are accessible to conscious introspection and why they are subject to rational control. If this speculation cuts ice, we have reason to assume that even 'solipsistic decision making' is a kind of social process in which reasoning about the decision is not mere rationalization of a decision already made but can become critical thinking.

8 Social reality

If we take social interaction into account, we see that the beliefs of a person are influenced by the exchange of arguments with other persons. The beliefs of these other persons will be changed in a similar manner. Social interaction leads to mutual adaptation of beliefs, and often even to some degree of consensus. The explanation of beliefs and the decisions based on these beliefs should therefore take social interaction into account.

However, that is not all. Some facts exist because, simply formulated, they are broadly believed to exist. These facts will be called 'social facts' or facts in social reality. For these facts it holds that social interaction not only promotes the existence of belief consensus, but also that social interaction leads to the existence of these facts. This section and its subsections deal with social facts and their relevance for the explanation of beliefs and decisions. The argument starts with the way in which the philosopher of science Karl Popper dealt with social facts in his field of expertise.

8.1 Epistemology without a knowing subject

In 1967, Popper gave an address that was published under the title 'Epistemology Without a Knowing Subject' (Popper 1979, pp. 106-152), Popper's main message was that epistemology should not deal with persons holding beliefs, but with knowledge that is abstracted from the persons having it. To make his point, Popper distinguished between three 'Worlds'. World 1 would consist of physical objects or states, World 2 of mental states of individual persons and perhaps also their dispositions to act, and World 3 of knowledge that is abstracted from knowing subjects. World 3 would not only contain theoretical systems or problems, but – and this is important – also critical arguments. For example, World 3 would contain quantum physics on the Copenhagen interpretation, but also the many worlds argument against it (Faye 2019). Although Popper did not discuss law, World 3 would also contain legal science, its claims, its doctrines, and its argument forms. See Figure 4.

Importantly, the elements of World 3 are the products of the beliefs and arguments in individual minds, but also influence what is in those minds. For instance, the theory that light propagates in a medium called 'ether' inspired the Michelson-Morley experiment to measure the speed of planet Earth relative to this ether.²⁹ A legal example would be that the rules in a legal system depend on the

²⁹ For a description of the experiment, see https://en.wikipedia.org/wiki/Michelson–Morley_experiment. Accessed 8 January 2024

beliefs and arguments in the minds of the subjects of that system, but a judge may believe that Susan ought to compensate the damage of Max because of a rule of the legal system in which he operates.

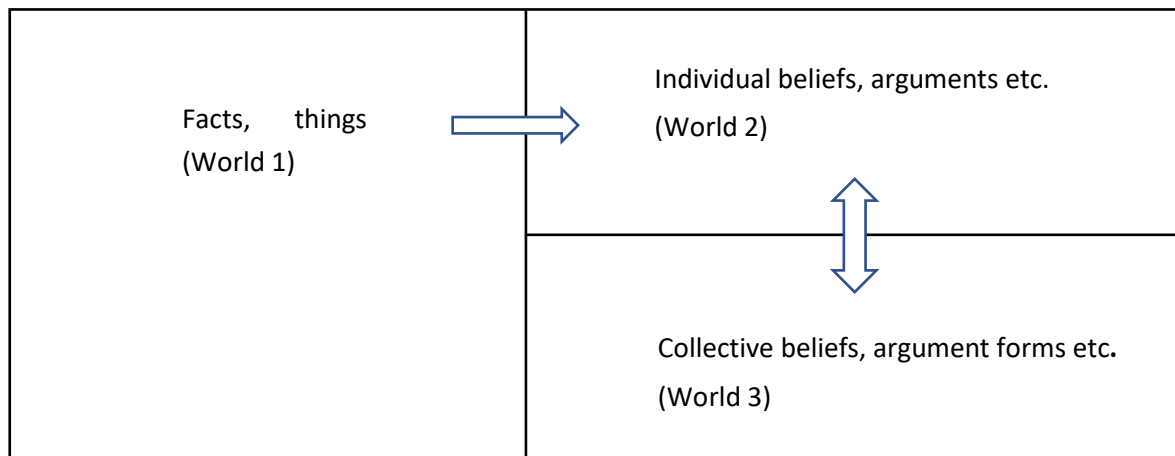


Figure 4: The three Worlds and our knowledge of it, according to Popper

There are at least two important messages in Popper's theory. One is that there are facts that are mind-dependent, but which can be studied independent of the minds that gave rise to them. These are the facts in what Popper called 'World 3', but which in this article will be referred to as facts in social reality, or 'social facts'. The second important message is that social reality also contains standards for good reasoning, indicating what facts can and rationally ought to be recognised given other facts that were already recognised. These standards include methodological rules, for instance on the conditions under which abductive arguments are acceptable, or logical rules on what theorems can be deduced from a set of axioms. They also include legal rules, specifying what legal conclusions ought to be drawn from facts situations. If, as it is claimed here, rules of inference, methodological rules and legal rules are all part of social reality, this means that these rules are the products of human minds in social interaction. They can be 'discovered' by analysing our social facts.

8.2 Objective, subjective, and social facts

In the following, I will use the words 'fact' and 'state of affairs' in a technical sense. My starting point will be the existence of a language which includes statements (descriptive sentences). Statements express states of affairs and are either true or false. For instance, the English language includes the statement 'It is raining'. This statement expresses the state of affairs that it is raining and is true if it is raining and otherwise false. If the sentence is true, the expressed state of affairs is a fact, and otherwise not. In this connection, a fact is an element of the world that makes a declarative sentence (or a proposition) true.³⁰

³⁰ These definitions make facts dependent on, amongst others, a language and the descriptive sentences it can express. For a discussion, see Hage 2018, pp. 32-34.

People distinguish between what is objective, subjective, and social.³¹ The distinction between these three kinds of states of affairs is based on two underlying characteristics which may be present or not. The two characteristics are whether the state of affairs is:

- mind-dependent; and
- the same for everybody.

Objective states of affairs are (1) not mind-dependent and (2) the same for everybody. An example would be the state of affairs that Mount Everest is a higher mountain than the Vaalserberg (the highest 'mountain' of the Netherlands).

Subjective states of affairs (1) depend on what individual persons recognise and are mind-dependent, and (2) are therefore not the same for everybody. An example would be the 'fact' that Mozart was a better composer than Salieri. Many people would not call subjective facts 'facts' at all; they reserve the word 'fact' for objective facts and perhaps also social facts.

Social states of affairs are somehow in between objective and subjective: (1) they depend on what the members of a social group recognise and are in that sense mind-dependent, and (2a) they are the same for the members of a group, but (2b) not necessarily the same between groups. One example is the law of a country. The law depends, in a complicated manner, on what the legal subjects of a country recognise as law and is the same for these legal subjects. However, different countries may have different laws, and what is the law for a Frenchman may not be the law for somebody in China.

8.3 Conventional social facts

Social facts are either conventional or constructivist. Conventional social facts only (but not always) exist in a group if most members of that group *recognise* that they exist (see section 3.5). Sometimes the task of recognition is delegated to one or more specific persons or institutions. A well-known legal example is that the recognition of rules as legal rules is delegated to courts and other 'officials' (Hart 2012, pp. 113-117). Delegated recognition presupposes that the persons to whom the recognition is delegated (the representatives) are recognised as such and that the members of the group tend to recognise what their representatives recognised on their behalf. So, if legal subjects delegated the task to recognise rules as law to the courts, they should recognise courts as their representatives for this purpose and they should normally recognise rules as law for the reason that the courts recognise them.

There is more to the existence of conventional facts than mere recognition. Suppose that Hendrik is broadly recognised as the leader of the Maastricht Cycling Club (MCC). However, it should not only be the case that sufficient members of MCC recognise Hendrik as their leader; the club members should also believe that sufficient other members also recognise Hendrik as leader of the group, and that these other members have the same beliefs about their fellow cyclists. Concretely, a group member such as Petra should not only have beliefs about Hendrik, but also about what her fellow group members recognise, including what her fellow group members believe about the beliefs of Petra herself.

³¹ The following theory about social reality and the kinds of facts in it was developed in a series of papers, some of which have already been (pre-)published (Hage 2022a, 2022b, 2023a, 2023b). The terminology in these papers is not fully consistent.

A third condition for the existence of conventional facts is that something can only be a conventional fact if states of affairs of that kind are not considered to be objective, subjective or constructivist. For instance, even if everybody believes that heat consists of calories, and also believes that everybody else believes this, it is nevertheless not a conventional fact. The reason is that the nature of heat is (usually) considered to be an objective state of affairs. For types of states of affairs that are considered to be objective, such as the nature of physical phenomena, the existence of a consensus is not decisive for what the facts are.

To be conventional, a kind of state of affairs should also not be considered as constructivist. For a constructivist in ethics, the mere consensus about a particular moral judgment does not prove the judgment to be correct. Even if 'everybody' agrees that coloured people are inferior, this does not show coloured people to be inferior indeed. This is different for being the leader of an informal club such as MCC, where consensus is decisive.

So, the existence of a conventional social fact requires recognition on two levels: a particular type of state of affairs must be considered social – not objective or subjective – and not constructivist, and a concrete instance of this type must be broadly recognised as existing. For instance, the members of MCC must (1) consider the leadership of their club to be a matter of conventional social fact and (2) they must recognise Hendrik as their leader.

8.4 Constructivist social facts

Not all social facts are conventional. There is a second category, constructivist facts, where an existing broad consensus is not the final word on what the facts are. Suppose that the members of MCC take a vote on what was the best cycling trip they made this year. They decide unanimously that the trip to the castle gardens in Arcen was the best trip. Does this mean that the Arcen trip really was the best trip? No, even if all club members agree on what was the best trip, this does not mean that it *really* was the best trip. It remains possible to raise the question of whether all members of the club were mistaken about the best trip.

There is a difference between what most or even all members of the group recognise as the best trip and what really was the best trip. Facts such as the fact about what was the best cycling trip of the year are not objective, because they depend on how people 'feel' about things. Neither are they merely subjective, as it makes sense to argue about them. And, finally, they do not seem to be conventional social facts either, because a broadly shared belief about them is not the final word. I will call such facts *constructivist facts*.³²

Constructivist facts are social facts, which are nevertheless open to serious questioning. This combination is possible if the social practice of a group does not only recognise the existence of these facts, but also the possibility to question them. For instance, *prima facie* it may be a social fact in MCC that the trip to the castle gardens of Arcen was the best trip of the year. However, the members of MCC agree and know that the others also agree that, theoretically speaking, everybody might be mistaken. If somebody came up with convincing reasons that another trip was even better, this other trip would be better. Moreover, it would have been better from the beginning, not merely because

³² There are close connections between these constructivist facts and constructivism (intuitionism) in the philosophy of mathematics (Iemhof 2020) and constructivism in moral philosophy (Rawls 1980; Bagnoli 2021).

the members of MCC changed their minds. If an argument makes people change their minds about constructivist facts, they change their minds about what the facts already were.

Constructivist facts are characterized by the possibility to have a *serious* debate about them. ‘Serious’ means in this connection that the participants in the debate believe that it is possible to disagree about these facts without thereby showing a misunderstanding of what the debate is about and that there is a correct answer to the question what the facts are, independent of what people actually believe it is. For instance, if Joanna and Frédéric disagree about whether red wine is better or white wine, while they believe that it is just a matter of taste, they consider the issue at stake to be a merely subjective one. There is no right answer as to what the best wine is, and their disagreement is not serious. If two members of MCC disagree about whether Hendrik is their leader, while both know that practically all members of the club recognise Hendrik as their leader, their disagreement is not serious either. The reason is that not believing that Hendrik is the leader while also believing that ‘everybody’ recognises Hendrik as the leader, shows misunderstanding of the conditions for leadership, which is a matter of convention.³³ The example about the best cycling trip of the year illustrates that it is possible to disagree seriously about what was the best trip. The seriousness of the debate becomes manifest in the assumption of all participants that there is a right answer to some question, even though it is not a matter of objective fact, and that this right answer does not change if people merely disagree about what the answer is.

Which social facts are constructivist, and which ones are conventional? It is impossible to give this question a general answer. The social practice of a group determines which social facts count as constructivist and which ones as conventional. If a broadly shared recognition may seriously be questioned, the social fact is considered to be constructivist; if not, it is conventional. Moreover, it seems that this categorization as conventional or constructivist is itself a matter of constructivist, and therefore also social, fact. People can seriously disagree on whether a particular kind of fact is conventional or constructivist. In legal philosophy, for example, there is a serious debate (in different terminology) between hard legal positivists and non-positivists on whether law is conventional or constructivist (cf. Gardner 2001 and Dworkin 1986). In ethical theory, there is a similar debate between conventionalists (relativists) and constructivists (Gowans 2021; Bagnoli 2021).

A constructivist fact is a fact that is recognised as a result of the rational reconstruction of the set of objective facts and social facts that are recognised in a social group.³⁴ Such a reconstruction will often consist of a debate. The debate may be casual, as amongst the members of MCC about the best cycling trip. It may also be more formal, as a debate in science about the best explanation of a newly discovered phenomenon. Rational reconstruction may involve no change for a particular social fact, and then that fact continues to exist as a social fact in the group because it was already recognised. An example would be that the members of MCC group believe that the cycling trip to the castle gardens of Arcen was the best trip of 2020 and that this belief survives a rational reconstruction of their belief set. Then the belief that the cycling trip to the castle gardens of Arcen was the best trip is

³³ Of course, it is possible to have serious discussions on the issues of whether Hendrik is a good leader or whether Hendrik ought to be the leader. However, these discussions would address another issue than whether Hendrik *is* the leader.

³⁴ There is no room in this article to further develop the notion of a rational reconstruction. As a very short alternative, I suggest that rational reconstruction of a set of beliefs and recognitions is making the set integrated coherent (Hage 2005a, 2013; 2016).

an element of the rationally reconstructed belief set, because it was already in the original belief set and nothing changed in this respect.

A second possibility is that reconstruction involves the inclusion of a particular social fact, and then that fact exists as a social fact in the group because it ought to be recognised according to the rational reconstruction. An example would be that the members of MCC initially did not have the rule that members of all religious convictions should be treated equally, but that the existence of this rule is included in the rationally reconstructed set and the rule therefore already existed as a matter of constructivist fact.

As a third possibility, reconstruction may involve the removal of a particular social fact, and then that fact did not exist as a constructivist fact in the group because it ought not to be recognised according to the rational reconstruction. An example would be that the members of MCC group ought not to have recognised the trip to Arcen as the best one. Then the belief that the cycling trip to the castle gardens of Arcen was the best trip is not part of the rationally reconstructed belief set and the trip to Arcen was, all things considered, not the best trip.

Last but not least it needs mentioning that reconstructing a belief set is an open-ended process. The existing belief set is an important starting point which to a large extent determines what its rational reconstruction will involve. However, it is always possible to adduce 'new' facts as arguments in a reconstruction debate. If these new arguments are broadly accepted in the relevant social group, the facts they mention will be added to the belief set in reconstruction, and the arguments will influence what facts should rationally be included in the reconstructed set. An example will be discussed in section 8.6.

Rationally reconstructing a set of recognitions or beliefs leads to a judgement on what rationally ought to be recognised, given the original beliefs. The recognitions in the reconstructed set are what the believer of the original set rationally ought to recognise. Moreover, as the example of the best cycling trip illustrates, the facts that rationally ought to be recognised are also the 'real' facts, because we are speaking of constructivist social facts. The members of MCC who argue about what was really the best cycling trip argue about what really was the case. Constructivist facts are the conclusions of the best possible arguments. These arguments determine what, rationally speaking, ought to be recognised, but *ipso facto* they also determine that part of social reality. Perhaps this is the most important thing to remember about constructivist facts: *constructivist reality is what rationally ought to be recognised as real*.

What counts in this connection as rational? Is there an objective, mind-independent standard for rationality, identical or analogous to the standard of classical logic? The proliferation of logical systems in the last, say, 70 years, suggests the opposite (Haack 1978; Priest 2008; Walton *et al* 2008). To cut a potentially long argument short, I will assume here that rationality is a matter of constructivist fact. Social conventions form the starting point in determining the standards of rationality, but they are not the last word. The debate on what counts as rational is to be conducted by means of standards which are themselves subject to debate.

8.5 Why legal facts are constructivist

Let us assume that law is a part of social reality and that this also holds for legal facts such as the fact that Iris is punishable, that John must stop for the red traffic light, or that this statutory rule is valid

law. Then the question arises of whether these social facts are constructivist or conventional. Assuming, for the sake of argument, that the answer is the same for all legal facts, the best view seems to be that legal facts are constructivist.

Remember that whether a kind of state of affairs is conventional or constructivist depends on whether a broadly shared view is the last word, both in the sense that conventional facts are what 'everybody' recognises them to be and in the sense that if there is no broad consensus, there is no conventional fact. If legal facts were conventional, this would mean that there is no law where there is a lack of consensus on what the law is. Hard cases would be cases where there is a gap in the law because of a lack of consensus. If the conventional view of law would be correct for legal facts, gaps would be a common phenomenon. In contrast, if the constructivist view would be correct, gaps would only occur if a rational reconstruction of what is broadly recognised would not give an answer. If this could occur at all³⁵, it would happen only occasionally. Legal decision makers seldom seem to assume that there is a gap in the law and to decide a case on the basis of moral or policy considerations only. So, it seems that these officials recognise more law than the conventional view claims there are. Since the views of these officials are decisive for whether legal facts are conventional or constructivist, it would seem that they are constructivist.

A similar argument starts from the observation that even if there is a broad consensus on what the law is, lawyers sometimes continue to argue as if this consensus is wrong. Such arguments can only be taken seriously if law is considered to be constructivist. This also pleads for the view that legal facts are constructivist.

A third argument is that the idea of legal sources only makes sense on a constructivist view of law. The idea of legal sources is that rules that can be traced back to a source of law are for that reason valid legal rules and – a less convincing addition – that rules that cannot be traced back to some legal source, are for that reason not legal rules. On a conventional view of law, the only reason why a rule is a valid legal rule is that it is broadly recognised as such. If a legal source plays a role in this connection, that maybe an interesting observation, but the source does not make a legal rule valid. On a constructivist view, on the contrary, sources can be crucially important, because legal rules are valid if (and only if) they rationally ought to be recognised as such. If a rule rationally ought to be recognised as valid law, it is valid law, even if it is not (yet) broadly recognised as such. This makes sense on a constructivist view of law.

A fourth argument is the argument from legal interpretation. Legal debates on the correct interpretation of a legal source are debates on whether a rule can be traced to this source. Such debates are broadly recognised in legal practice as making sense. This is another argument why legal practice treats legal facts – this time facts about what are valid legal rules – as constructivist. And if legal practice treats these legal facts as constructivist, they are *prima facie* constructivist.³⁶

³⁵ Dworkin (1981) claimed that all cases have one right answer.

³⁶ It is only *prima facie* because the issue of whether legal facts are constructivist is itself a matter of constructivist fact.

There is sufficient reason to assume that law exists in social reality. Since there are at least four reasons (see above) why legal facts are constructivist, and since there are no obvious reasons why law would be conventional³⁷, it can be concluded that legal facts are constructivist.

8.6 Open-ended reconstruction

If legal facts are constructivist, what does this mean for explaining legal judgments? Let us assume that legal judgments can be explained (type 2) on the belief/desire model. A decision maker wants to apply the law, and so his belief about what the law is determines what decision he will take. The explanation of the decision is therefore essentially an explanation of what the decision maker believes the law is. A second assumption will be that the decision maker is a member of the social group that recognises a particular legal system.³⁸ Being a member of this group implies two things. First, that legal facts are recognised as constructivist facts. And second that the social facts recognised by the group members – directly in the case of conventional facts; indirectly in the case of constructivist facts – are also recognised by the decision maker. If these two assumptions are made, it follows that the explanation of the decision lies in the argument why the decision is in accordance with the law.

Assuming that legal facts are constructivist, an argument why a decision is in accordance with the law is an argument why it is rational to recognise the conclusion of the argument as a legal fact. We have seen that the rationality of such a conclusion is to a large extent determined by the original belief set of the group members, including the standards for rational belief (revision). However, we have also seen (the claim) that debates on how a belief set should be rationally reconstructed is open-ended in the sense that it is possible to adduce ‘new’ facts in the argument.

An example may clarify this point, and the standard example of this article about the car accident between Susan and Max (see section 1) will be used again. The facts of the case were as follows:

Susan drives at moderate speed through the outskirts of her town. Suddenly Max comes crossing the street, running from behind a parked car. Even her careful driving cannot protect Susan from a collision with Max.

Is Susan (or her insurance) liable for the medical costs of Max? If legal facts were conventional, the question after the liability of Susan for the medical costs of Max would only have a legal answer if there were a broad consensus between the legal subjects that Susan is (not) liable. Without such a consensus, there would be no law on the issue. This is not how law operates, which goes to show again that legal facts are not conventional.

Let us assume that there is a valid legal rule to the effect that if a car is involved in an accident with a pedestrian, the possessor of the car (typically the owner) is strictly liable for the damage caused by the accident. Normally, this damage will be covered by the mandatory insurance for car-possessors. Given (the validity of) this rule, it is *prima facie* rational to accept the result of its application to the case. So, *prima facie*, Susan will be liable for the medical costs of Max.

³⁷ Of course, this is here merely a claim, and the author invites adherents of the view that legal facts are conventional in the sense of ‘conventional’ proposed here to provide the reasons which the present author does not see.

³⁸ Some may object here that artificially intelligent legal decision makers cannot be members of the social group that recognizes a particular legal system. For those, the assumption should become that the ‘beliefs’ of these AI-systems are similar to the beliefs of humans who recognize the legal system.

Suppose, however, that Max came running from behind the parked car because he wanted to end his life by the expected collision with Susan's car. Suppose, moreover, that there is no positive (written) law that deals specifically with suicide attempts.³⁹ Is it possible to effectively adduce the attempted suicide as a reason against the liability of Susan? That is possible, but only if this exception to the strict liability rule is broadly recognised by the legal subjects.^{40 41} Notice that such broad recognition is a psychological notion, and that reason – and therefore also the original belief set – will normally not play a role in it.

9 Implementation issues⁴²

The discussion of social facts was introduced (in section 5) by the following argument:

'A court wants (desires) to apply the law. If the court believes that the law is X, for instance that Susan is liable for the damage of Max, this explains why the court decides X. In this argument there is much room for argumentation about the content of the law, as this argumentation determines what the court believes the law to be. So, the reasons for and against a particular view of what the law is in a concrete case determine what the court believes and together with the court's desire to apply the law this belief explains the decision of the court. Moreover, if the arguments that determine the court's belief about the law also influence, or even determine, the view on the law of the court's audience, the court will be able to explain its decision in a manner that is both a type 1 and a type 2 explanation.'

We have now seen that the argument that leads a legal decision maker to a particular result is open-ended in the sense that the belief set from which the decision maker and its audience started is not the only factor that determines the outcome of the decision making process. This open-ended nature allows the argumentation process to be influenced by intuitions that are caused by unconscious processes. A full type 1 explanation of a legal decision also needs to mention how the intuitions that influenced the decision came about. Even if a decision maker were able to figure out how his intuitions were formed, mentioning these causes would not fit in the type 2 explanation that the audience expects.

This leads to the conclusion that the 'ideal' situation in which a type 1 explanation of behaviour (a legal decision) fulfils the function of a type 2 explanation does not in general exist in law. Legal decisions are strongly influenced by legal facts, and legal facts are the results of a process of decision

³⁹ Some identify the law with positive law. However, such an identification is not well-defendable if law is constructivist. Therefore, it makes sense to distinguish between positive law (as it can be derived from legal sources) and law *tout court*, the law as it ultimately is in concrete cases, and which does not necessarily coincide with the positive law. See Hage (2019).

⁴⁰ If the subjects have delegated the recognition task to the courts for which cases are brought, the consensus only needs to exist between the members of the court. Moreover, if the court uses a decision rule according to which a (qualified) majority of the court members can decide on behalf of the court, the consensus only needs to exist in a (qualified) majority.

⁴¹ An alternative for broad actual recognition would be what ought rationally to be recognised, but this rationality also depends in last instance on what is actually recognised.

⁴² The argument in this subsection was strongly influenced by discussions of the author with Maarten van der Meulen. The author thanks Maarten for these discussions but takes the sole responsibility for the mistakes that may have remained in the proposal for implementation.

making in which intuitions – which are, by definition, not based on reasons – play an important role. Whether we like it or not, in law there will often be a discrepancy between type 1 and type 2 explanations of decisions. If a legal decision maker, whether with natural or artificial intelligence, wants to explain (type 2) his decision to create understanding, he will normally not be helped by trying to find out how his decision came about (type 1 explanation). Notice that this conclusion not only applies to artificially intelligent legal decision makers, but also to human legal decision makers. Where it comes to explaining legal decisions, there is in this respect no fundamental difference between humans and AILSs.

From now on, we will work from the assumption that for AILSs type 1 explanations cannot be used to create understanding. Suppose that we have an AILS that is the product of machine learning. When fed with the facts of a case, it produces a legal decision as output, and the way in which the system transformed the input to the output remains inaccessible in the system's black box. A type 1 explanation is not even available, except in the very abstract sense that we can describe the learning procedure of the system, its training sets, and the general architecture (interlinked nodes). If a user of the system looks for an explanation of the system's decision, this type 1 explanation will be of no avail. The user wants a type 2 explanation in terms of case facts, rules, and perhaps legal sources and methods of legal reasoning. How is it possible to make the system produce such a type 2 explanation on the basis of the available type 1 explanation?

The answer is that this is *not* possible. The type 2 explanation must be generated independently, but some connection with the operation of the system can be maintained. In the following it will be described how it is possible to maintain at least some connection. However, a *caveat* is in place: the following description is quite abstract, and far removed from an actual implementation. Eventual attempts to create a ML-system that can give type 2 explanations will no doubt have to fill in many details and will most likely have to adapt much of what is suggested here. Hopefully, the present proposal is at least helpful for creating a legal ML-system that can give type 2 explanations.

Any acceptable type 2 explanation will resemble explanations as they are given by human legal decision makers, as the existing human practices set the framework of expectations in which any viable type 2 explanation must fit. So, depending on the legal system in which the AI system must function, the explanation will invoke the case facts, relevant previous cases, other sources of law, and the logical tools that are available to legal reasoners. Moreover, also less traditional tools can be used to remove the cognitive dissonance, such as hypotheticals (Ashley 1991), scenarios (Bex, Prakken, Verheij 2007; Mackor, Jellema, Van Koppen 2021), and rhetoric (Witteveen 1988; Del Mar 2020). Given all of these tools, some legal decisions are indefensible, and it is impossible to produce type 2 explanations for the decisions. If the ML-system produces such indefensible decisions, even after feedback from the failed attempt to explain, its learning process went wrong and the system is not good enough and should not be used.

If a ML-system produces mostly defensible decisions, that is: decisions for which type 2 explanations can be generated, the system should generate type 2 explanations for the decisions produced by the system's black box. In this connection it comes in helpful that law offers many reasoning tools that allow legal decision makers to produce explanations/justifications for many decisions, even for incompatible decisions. It turns out that a characteristic of law that is traditionally deemed problematic, namely that it allows arguments for incompatible conclusions, is actually a strength.

Legal decision makers cannot produce explanations/justifications for any decision, but they can produce them for many different decisions, even for incompatible ones. To mention a few possibilities (the list is far from exhaustive):

- There is leeway in the classification of the case facts, with the possibility to make a rule applicable or not. For instance, if a hungry judge (see section 4.2) does not want to allow parole, he can claim that the case facts show that there is a serious risk of recidivism.
- If a rule is applicable (its conditions are satisfied) it is always possible to make an exception. For instance if the court does not want to hold Susan liable for the medical costs of Max (see sections 1 and 8.6) for the reason that Max attempted suicide, it can make an exception to the rule about strict liability, even if there is no statutory basis or precedent for doing so.
- If a court believes that the 'obvious' decision in a case harms an important value, it can interpret an existing human right to make it also protect that value. The rule on which the 'obvious' decision is based can then be declared null and void because of a conflict with the constitution, an international convention, or with unspecified 'higher law' (*ius cogens*).

So, it is proposed that an AILS has (at least) two modules: a black box resulting from machine learning, that generates decisions for cases, and an explanation module that generates type 2 explanations for the decisions of the black box module. These type 2 explanations should be legal arguments that explain/justify the decision, which are adapted to the beliefs and values of the audience – they should create understanding – and which make use of the sources of law and the toolbox of legal method.

If it turns out impossible to produce such an argument, this finding may be fed back into the black box as additional input. However, if it is possible to produce an acceptable argument that does the job, this argument can be added to the decision as its explanation. Clearly, the output of the black box module does not have to be determined by this explanation, but that is precisely the point of this article. Moreover, also in human legal decision makers the explanations they give may be independent of the way they arrived at their decisions.

Because this article started from the assumption that an AILS was the product of machine-learning, the implementation suggested above made room for a black box that generates possible decisions that needed to be justified by a traditional legal reasoning system. However, as soon as the generation of a legal decision and its justification are made more or less independent, there comes room for legal justification of partial decisions. If a party in a legal dispute has a preferred outcome, the explanation module can generate an argument for that outcome. If the other party favours a different outcome, the explanation module will often be able to rationalize that outcome as well. In this respect, the 'intuitive' (in the sense of non-argued) decision of a black box is comparable to a partial outcome favoured by a process party. Given the constructivist nature of law, there is no a priori correct legal outcome for a case which the system should learn. This raises the question of whether in law there is a role for machine-learning if the constructivist nature of law makes it unclear what precisely the system should learn. The answer to this question is left open for the readers of this article.

10 Conclusion

Most of this article was devoted to answering the question of whether and how ML-systems that decide, or advice about, legal issues can explain their decisions. In this connection, a distinction was

made between type 1 explanations which describe the process leading to the decision, and type 2 explanations, which aim to create understanding of the decision in the audience. Furthermore, an ‘ideal’ situation was mentioned in which a type 1 explanation is the means for a type 2 explanation. Then the two types of explanation would be two sides of the same coin.

By means of social psychology and cognitive neuroscience it was shown that this ideal situation does not normally occur. People have no introspective knowledge about, or other privileged access to, what moved them. Moreover, a description of the process leading to a decision, even if it happens to be true, is typically not the type 2 explanation people expect. For actions, including decisions, people expect explanations in terms of reasons, and not in terms of causes which are not at the same time reasons.

However, the situation might be better with regard to legal decisions, because law is a social phenomenon and is therefore more adapted to explanations in terms of reasons. Alas, even that is not the case because legal facts are constructivist, which means that they are the outcome of a debate which is fundamentally open-ended. For artificially intelligent legal decision makers based on machine learning this means type 2 explanations of their decisions in general cannot be descriptions of the processes leading to these decisions (type 1 explanations).

Therefore, it was proposed that such systems consist of (at least) two parts: a black box that is the result of machine learning, and a module for providing a type 2 explanation. The black box generates a decision, and the explanation module uses explicitly represented knowledge to explain the decision of the black box. The leeway provided by the toolbox of legal method will normally make it possible for the explanation module to generate a type 2 explanation that on the one hand fits the black box decision and on the other hand creates understanding in the audience. Moreover, even the role of the black box module may be questioned as it is unclear what precisely the law is that must be learned by this module.

Here the body of this article ends. However, the article also aims to show how insights from the cognitive sciences (including, but not confined to, artificial intelligence) is crucial for the development of artificially intelligent tools for legal applications. The field of Law and AI is to a large extent an engineering discipline, focused on making tools that perform intelligent legal tasks. However, it is also the study of the role of AI in law and of what we expect artificially intelligent legal tools to do. To explain their outcomes is one of the things these tools are expected to do. What this means, and what is possible and desirable in this connection are questions that can only be answered by leaving the domain of technical solutions behind and entering the broader field of law and the cognitive sciences.

Acknowledgements

An earlier version of this article was read by Frans Leeuw, Henrique Marcos, and Antonia Waltermann. The author is grateful for the valuable comments they gave, as well as for the discussions with Maarten van der Meulen, which inspired the idea to separate causal explanation and rationalization in AILSs. An outline of this article was presented at the Jurix 2023 conference in Maastricht and the questions and discussions at that occasion were also a valuable contribution. Finally, three anonymous reviewers for the *AI and Law* journal gave extensive comments, which helped to avoid some minor mistakes and which – more importantly – made it clear that the paper should place more emphasis on its main goal

to change the AI and Law-project and should downplay its proposed 'solution' for the practical problem of providing legal decisions with understandable explanations.

References

Alvarez M (2017) Reasons for Action: Justification, Motivation, Explanation. In: Zalta EN (ed), *The Stanford Encyclopedia of Philosophy* (Winter 2017 Edition).

URL = <https://plato.stanford.edu/archives/win2017/entries/reasons-just-vs-expl>. Accessed 6 January 2024

Ashley KD (1991), Reasoning with Cases and Hypotheticals in Hypo, *International Journal of Man-Machine Studies* 34: 753-796

Atkinson K , Bench-Capon T, Bollegala D (2020) Explanation in AI and law: Past, present and future. *Artificial Intelligence* 289: 1-21

Bagnoli C (2021) Constructivism in Metaethics. In: Zalta EN (ed), *The Stanford Encyclopedia of Philosophy* (Spring 2021 Edition).

<https://plato.stanford.edu/archives/win2017/entries/constructivism-metaethics/>. Accessed 8 January 2024

Bargh J (2017) *Before You Know It. The Unconscious Reasons We Do What We Do*. Windmill Books, London

Bennett MR and PMS Hacker (2003), *Philosophical Foundations of Neuroscience*, Blackwell Publishing, Malden e.a.

Bex FJ, Prakken H, Verheij B (2007), Formalising argumentative story-based analysis of evidence. *Proceedings of the 11th International Conference on Artificial Intelligence and Law*. ACM Press, New York, pp 1-10

Brenner A, Maurin A-S, Skiles A, Stenwall R, Thompson N (2021) Metaphysical Explanation. In: Zalta EN (ed), *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition).

<https://plato.stanford.edu/archives/win2021/entries/metaphysical-explanation>. Accessed 6 January 2024

Brożek B (2016) Explanation and Understanding. In: Brożek B, Heller M, Hohol M (eds), *The Concept of Explanation*. Copernicus Center Press, Kraków, pp. 11-42

Carruthers P (2011). *The Opacity of Mind. An Integrative Theory of Self-Knowledge*. Oxford University Press

Cosmides L, Tooby J (1992) Cognitive adaptations for social exchange. In: Barkow J, Cosmides L, Tooby J (eds), *The Adapted Mind*, OUP, Oxford, pp. 163–228

Danziger S, Levav J, Avnaim-Pesso L (2011). Extraneous factors in judicial decisions. In: Proceedings of the National Academy of Sciences 108, 6889-6892

Del Mar M (2020) Artefacts of Legal Inquiry. Hart, Oxford et al

Dijksterhuis A (2016) Het slimme onbewuste. Denken met gevoel. Prometheus, Amsterdam

Douven I (2021) Abduction. In: Zalta EN (ed), The Stanford Encyclopedia of Philosophy (Summer 2021 Edition). <https://plato.stanford.edu/archives/sum2021/entries/abduction>. Accessed 7 January 2024

Dworkin R (1981) Is there a right to pornography? In: *OJLS* 1: 177-212

Dworkin R (1986) Law's Empire. Fontana, London

Evans J, Stanovich K (2013) Dual-Process Theories of Higher Cognition: Advancing the Debate. In: Perspectives on Psychological Science 58: 223-241.

Fay B (1994), General Laws and Explaining Human Behavior. In: Martin M, McIntyre LC (eds), Readings in the philosophy of social science, MIT Press, Cambridge, pp 91-110

Faye J (2019) Copenhagen Interpretation of Quantum Mechanics. In: Zalta EN (ed), *The Stanford Encyclopedia of Philosophy* (Winter 2019 Edition). <https://plato.stanford.edu/archives/win2019/entries/qm-copenhagen/>. Accessed 8 January 2024

Festinger L (1957) A Theory of Cognitive Dissonance. Stanford University Press, Palo Alto

Fodor JA (1983) The Modularity of Mind. The MIT Press, Cambridge (Mass)

Frey JA (2019/20), Anscombe on Practical Knowledge and the Good. In: *Ergo* 6, No. 39. <https://doi.org/10.3998/ergo.12405314.0006.039>. Accessed 7 January 2024

Gardner H (1987) The Mind's New Science. A History of the Cognitive Revolution. Basic Books, New York

Gardner J (2001) Legal Positivism: 5 1/2 Myths. In: *American Journal of Jurisprudence* 46: 199-227

Gazzaniga MS (2016) Who's in charge. Free will and the science of the brain. Robinson.

Gazzaniga MS, Ivry RB, Mangun GR (2019) Cognitive Neuroscience. The biology of mind, 5th ed WW Norton & Company, New York and London

Gigerenzer G (2007) Gut Feelings. The intelligence of the unconscious. Penguin, New York

Gowans C (2021) Moral Relativism. In: Zalta EN (ed), The Stanford Encyclopedia of Philosophy (Spring 2021 Edition). <https://plato.stanford.edu/archives/spr2021/entries/moral-relativism/>. Accessed 8 January 2024

Grimm S (2021) Understanding. In: Zalta EN (ed), The Stanford Encyclopedia of Philosophy (Summer 2021 Edition). <https://plato.stanford.edu/archives/sum2021/entries/understanding/>. Accessed 7 January 2024

Guthrie C, Rachlinski JJ, Wistrich AJ (2001) Inside the Judicial Mind. In: Cornell Law Faculty Publications, Paper 814. <http://scholarship.law.cornell.edu/facpub/814>. Accessed 7 January 2024

Haack S (1978) Philosophy of Logics. CUP, Cambridge

Hage JC (1997) Reasoning with Rules. Kluwer Dordrecht

Hage JC (2005a), Law and Coherence. In: Studies in Legal Logic. Springer, Dordrecht pp 33-68

Hage JC (2005b) Reason-based Logic. In: Studies in Legal Logic. Springer Dordrecht, pp 69-99

Hage J (2013) Three kinds of coherentism. In: Araszkievicz M, Šavelka J (eds), Coherence: Insights from Philosophy, Jurisprudence and Artificial Intelligence, Springer, Dordrecht, 1-32

Hage J (2016) Anything goes : An Apology for Parallel Distributed Legal Science. In: Informal Logic 36: 271-287. <http://dx.doi.org/10.22329/il.v36i3.4719>

Hage J (2019) The Limited Function of Hermeneutics in Law. In: Duarte D, Moniz Lopes P, Sampaio JS (eds), Legal Interpretation and Scientific Knowledge, Springer, Cham, pp. 1-12

Hage J (2018) Foundations and Building Blocks of Law. Eleven International Publishing, Den Haag

Hage J (2021) Are the Cognitive Sciences Relevant for Law? In: Brožek B, Hage J, Vincent, NA (eds), Law and Mind. A Survey of Law and the Cognitive Sciences, CUP, Cambridge, pp. 17-49

Hage, J (2022a) Constructivist Facts as the Bridge Between Is and Ought. In: International Journal for the Semiotics of Law - Revue internationale de Sémiotique juridique 36, pp. 53–81

Hage, J (2022b) Rules and the Social World. In: L' Ircocervo 21, 16-35.

<https://lircocervo.it/?p=5346>. Accessed 7 January 2024

Hage, J (2023a) The Nature of Law and Constructivist Facts. In: Brigaglia M, Roversi C (eds.), Legal Reasoning and Cognitive Science: Topics and Perspectives, Diritto & Questioni Pubbliche, pp. 207-224. www.dirittoequestionipubbliche.org/page/2023_dq-recognise/00-DQrecognise-2023_contents.pdf . Accessed 23 January 2024

Hage, J (2023b) How Law's nature influences Law's Logic. In: Studia Humana (early view). https://studiahumana.com/art_bn,35,How-Law's-Nature-Influences-Law's-Logic.html. Accessed 23 January 2024

Haidt J (2001) The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. In: Psychological Review 108: 814-834

Haidt J (2012) *The Righteous Mind. Why good people are divided by politics and religion*. Penguin, London

Haidt J, Bjorklund F (2008) Social Intuitionists Answer Six Questions about Moral Psychology. In: Sinnott-Armstrong W (ed), *Moral Psychology, Volume 2: The Cognitive Science of Morality: Intuition and Diversity*, The MIT Press, Cambridge, pp. 181-218

Hart HLA (2012) *The Concept of Law*, 3rd ed, OUP, Oxford

Hempel CG, Oppenheim P (1948) Studies in the Logic of Explanation, *Philosophy of Science* 15, 135-175.

Hempel CG (1965) *Aspects of Scientific Explanation*: The Free Press, New York

Hempel, CG (1966) *Philosophy of Natural Science*. Prentice Hall, Englewood Cliffs

High-Level Expert Group on Artificial Intelligence (2019) *Ethics Guidelines For Trustworthy AI*, <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>. Accessed 5 January 2024

Hume D (1978), *A Treatise of Human Nature*. Selby-Bigge edition, OUP, Oxford. 1978. First edition 1739/40

Iemhoff R (2020) Intuitionism In: Zalta EN (ed), *The Stanford Encyclopedia of Philosophy* (Fall 2020 Edition). <https://plato.stanford.edu/archives/fall2020/entries/intuitionism/>. Accessed 8 January 2024

Kahneman D, Tversky A (1982) *Judgment under Uncertainty: Heuristics and Biases*. CUP, Cambridge

Kahneman D (2012) *Thinking, fast and slow*. Penguin, London

Kincaid H (1994) Defending Laws in the Social Sciences. In: Martin M, McIntyre LC (eds), *Readings in the philosophy of social science*, MIT Press, Cambridge, pp 111-130

Mackor AR, Jellema H, Van Koppen PJ (2021), *Explanation-Based Approaches to Reasoning about Evidence and Proof in Criminal Trials*. In: Brožek B, Hage J, Vincent, NA (eds), *Law and Mind. A Survey of Law and the Cognitive Sciences*. CUP, Cambridge, pp. 431-469

Mancosu P, Poggiolini F, Pincock C (2023) Mathematical Explanation. In: Zalta EN, Nodelman U (eds.) *The Stanford Encyclopedia of Philosophy* (Fall 2023 Edition). <https://plato.stanford.edu/archives/fall2023/entries/mathematics-explanation/>. Accessed 6 January 2024.

Marcos H (2023) *Consistency in International Law. How to Make Sense of a Decentralised and Expansive Rule-Based World*. Dissertation Maastricht University

Mercier H, Sperber D (2007) *The Enigma of Reason. A New Theory of Human Understanding*. Allen Lane

Moore JD, Swartout WR (1988) *Explanation in Expert Systems: A Survey*, ISI Research Report, ISI/RR: 88-228.

Nisbett RE, Wilson TD (1977). Telling more than we can know: Verbal reports on mental processes. In: *Psychological Review* 84, 231-259

Phillips P , Hahn C, Fontana P, Yates A, Greene K, Broniatowski D, Przybocki M (2021) *Four Principles of Explainable Artificial Intelligence*, NIST Interagency/Internal Report (NISTIR), National Institute of Standards and Technology, Gaithersburg, MD.

https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=933399. Accessed 13 October 2023

Piñeros Glasscock, JS, Tenenbaum S (2023), Action. In: Zalta EN, Nodelman U, *The Stanford Encyclopedia of Philosophy* (Spring 2023 Edition).

<https://plato.stanford.edu/archives/spr2023/entries/action/>. Accessed 7 January 2024.

Popper KR (1979). *Objective Knowledge. An Evolutionary Approach* (revised edition). OUP, Oxford

Prakken H, Sartor G (eds.) (1997) *Logical Models of Legal Argumentation*. Kluwer, Dordrecht

Priest G (2008) *An Introduction to Non-Classical Logic*, 2nd edition. CUP, Cambridge

Rawls J (1980) Kantian Constructivism in Moral Theory. In: *Journal of Philosophy* 77: 515-572

Raz J (1999) *Practical Reason and Norms*, 2nd edition. OUP, Oxford

Ross, L, Woodward J (2023) Causal Approaches to Scientific Explanation. In: Zalta EN, Nodelman U (eds.) *The Stanford Encyclopedia of Philosophy* (Spring 2023 Edition)

<https://plato.stanford.edu/archives/spr2023/entries/causal-explanation-science> . Accessed 6 January 2024

Rudner RS (1966) *Philosophy of Social Science*. Prentice Hall, Englewood Cliffs

Shapiro L, Spaulding S (2021) Embodied Cognition. In: Zalta EN (ed) *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition).

<https://plato.stanford.edu/archives/win2021/entries/embodied-cognition/>. Accessed 7 January 2024

Smith M (1994) *The Moral Problem*. Blackwell Oxford

Toulmin S (1958) *The Uses of Argument*, CUP, Cambridge

Van Fraassen B (1980) *The Scientific Image*. Clarendon Press, Oxford

Walton D, Reed C, Macagno F (2008) *Argumentation Schemes*. CUP, Cambridge

Weber M (1922) *Wirtschaft und Gesellschaft*. Mohr, Tübingen

Wegner DM (2002) *The Illusion of Conscious Will*. The MIT Press, Cambridge (Mass.)

Wilson TD (2002) *Strangers to Ourselves*. Discovering the adaptive unconsciousness. The Belknap Press, Cambridge (Mass)

Witteveen W (1988), *De retoriek in het recht*, Dissertation, Leyden University

Woodward J (2000) Explanation and Invariance in the Special Sciences, *The British Journal for the Philosophy of Science* 51: 197–254. doi:10.1093/bjps/51.2.197

Woodward J, Ross L (2021) Scientific Explanation. In: Zalta EN (ed) *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition).

<https://plato.stanford.edu/archives/sum2021/entries/scientific-explanation/>. Accessed on 6 January 2024