# Mindreading in Law

ŁUKASZ **KUREK**
*Jagiellonian University in Krakow*

ABSTRACT
This chapter will discuss legal reasoning about mental states. Law involves reasoning about mental states, i.e. mindreading, whenever it requires lawyers to interpret human behaviour. Behaviour interpretation is involved, *inter alia*, when the issue in question is whether a person is legally responsible. Typically, legal investigations of mindreading are theoretical or conceptual and result in descriptive claims about how lawyers mindread and normative claims about how lawyers ought to mindread in order to, let's say, attribute legal responsibility. In contrast to this approach, the present chapter will adopt an empirically-oriented perspective and focus on the psychological underpinnings of lawyers' capacity to mindread. The motivation for adopting the empirically-oriented perspective on this issue is twofold. First, in comparison to theoretical or conceptual investigations, empirical research provides us with more reliable data about the actual psychology of mindreading. What is more, empirical research in question points to important features of legal reasoning about mental states which are difficult to observe from the theoretical or conceptual perspective. Second, what cognitive science tells us about mindreading is relevant to the normative claims concerned with legal reasoning about mental states.

KEYWORDS
folk psychology, theory of mind, behaviour explanation, mental state attribution, legal responsibility

CITATION
Kurek Ł. 2023. *Mindreading in Law*, in Brigaglia M., Roversi C. (eds.), *Legal Reasoning and Cognitive Science: Topics and Perspectives*, «Diritto & Questioni pubbliche», Special Publication, August 2023.

DIRITTO & QUESTIONI PUBBLICHE / RECOGNISE
*Legal Reasoning and Cognitive Science: Topics and Perspectives*
Special Publication / August, 2023, pp. 143-157


ISSN 1825-0173

# Mindreading in Law

ŁUKASZ KUREK



## 1. *Introduction*

"Mindreading" refers to the human cognitive capacity to attribute mental states - such as thoughts and emotions - to other people and oneself. People use this capacity to make sense of the behaviour of the target of mental state attribution. Mindreading is used in various domains which include, but are not limited to, everyday life and law. The capacity in question is studied by cognitive scientists who push forward our understanding of it by describing the psychological mechanism which underpins it. Although there are still some disagreements about how this mechanism works, what we already know about should be of interest to those who want to better understand legal reasoning about human behaviour. The goal of this chapter is to survey the cognitive-scientific research on mindreading and discuss its implications for law.

## 2. *What is mindreading for?*

As it was mentioned, mindreading is used to interpret behaviour. To illustrate the relationship between mindreading and behaviour interpretation consider the following example. Imagine there is a person, Jones, who, upon leaving a party, took someone else's umbrella instead of his own. Both umbrellas were very similar. In fact, both umbrellas were two exemplars of the same model and the only noticeable difference between them was that the opening mechanism in Jones' umbrella was malfunctioning. However, before leaving the party, Jones did not open the umbrella that he took.

At least two competing interpretations of Jones' behaviour are available. According to the first interpretation, which is more favourable to Jones, he made an honest mistake. He was simply unaware that the umbrella he took belonged to someone else. This interpretation is corroborated by the fact that both umbrellas were very similar and the only noticeable difference between them could be observed upon their opening—and Jones did not open the umbrella he took upon leaving the party. According to the second interpretation, which puts Jones' in a far less favourable light, he stole the umbrella. This would have been the case if had known that he was taking someone else's umbrella and he just had it with his own, broken one.

Deciding between these two interpretations requires deciding which mental state to attribute to Jones. The interpreter needs to determine whether Jones believed that he was taking someone else's umbrella or whether he believed that the umbrella is his own. This is because stealing requires that the person who takes an object belonging to someone else is aware of this. A person cannot steal an object she believes to be her own.

Given the above, we may conclude that mental state attribution is essential if our goal is to determine whether someone made a mistake. Although this is true, the role of mindreading is

* I would like to thank the anonymous reviewer for offering insightful comments and raising hard questions which helped me strengthen this chapter.

much more prominent than this. We mindread in order to interpret behaviour for various practical purposes which fall into a broad category of behaviour management (MALLE 2004, 63-82; HUTTO 2008, 23-40; SPAULDING 2018, 42-61). Managing behaviour consists in, for example, attributing moral or legal responsibility, criticizing or praising as well as administering more substantial rewards or punishments.

## 3. *Cognitive science of mindreading: Some preliminary remarks*

Mindreading can be approached from three perspectives. First, mindreading is a cognitive *capacity*. Second, it was mentioned that cognitive scientists are interested in the features of the *information-processing* associated with this cognitive capacity. Finally, it was claimed that the capacity to attribute mental states is exercised in various domains which means that there is a diverse range of concrete *cases* in which we people use mindreading. Some readers may want to know a little bit more about the relationship between cognitive competences, information-processing and concrete cases in which these cognitive capacities are exercised as well as about the relationship between mindreading, behaviour interpretation and behaviour management.

Cognitive capacities tell us what a mind can do. Typically, they are individuated by their goal—by spelling out what minds endowed with these capacities can achieve. Thus, the most immediate goal of mindreading is to attribute mental states. Cognitive capacities are often nested, which means that having them is linked with having other, lower-level capacities. For example, the capacity to interpret behaviour—the most immediate goal of which is to make sense of behaviour—is linked with having mindreading. Behaviour interpretation is itself a lower-level cognitive capacity in comparison to the capacity to manage behaviour. To illustrate this, recall Jones, who had taken someone else's umbrella instead of his own. Our capacity to interpret Jones' behaviour—to decide whether he stole the umbrella or just made a mistake—is linked with our capacity to mindread, in the sense that the former involves being able to attribute to Jones a belief with a particular content. Going further, our capacity to decide whether Jones' should be punished involves our capacity to interpret his behaviour, in the sense that administering this punishment involves being able of making sense of what he did.

One way of describing how the mind processes information when a particular cognitive capacity is exercised consists in providing an algorithm—a set of instructions—for achieving the goal of this capacity. Below, we will spend some time discussing the information processing associated with mindreading. This is where the cognitive-scientific research on mindreading will be of particular relevance.

Finally, it is worthwhile to consider concrete cases in which people use mindreading. First, concrete cases in which people attribute mental states supply most of the observable phenomena in every discussion about mindreading. Not so long ago these cases supplied the only observational phenomena in this context but this state of affairs changed when cognitive scientists began to investigate how mindreading is implemented in the human brain. Notice that in concrete cases in which people attribute mental states we do not actually observe how mindreading works. What we observe in concrete cases is human behaviour—linguistic or non-linguistic—which suggests that people who engage in this behaviour mindread.

For example, we may hear Smith saying to Brown something like 'Jones did not steal your umbrella. He took your umbrella because he thought it was his and this may be taken to suggest that in order for Smith to utter this sentence, it is not enough that Smith knows the English language—that he knows what to say if he wants to communicate to Brown that Jones took Brown's umbrella by mistake. Instead, we may suggest that in order for Smith to utter this sentence Smith needs to be capable of solving a cognitive, not a linguistic problem—the former

consisting in attributing to Jones a belief with a particular content. This is precisely what cognitive scientists are suggesting in their study of mindreading.

Although what people say when they mindread is an important source of data in these studies, when cognitive scientists investigate mindreading they do not target a linguistic competence but a cognitive one. The drawback of this approach is that cognitive competences and the information-processing which is associated with them are non-observable phenomena in the sense that their existence and their features need to be inferred from what can be directly observed.

Second, it is worthwhile to take a closer look at concrete cases in which people use mindreading because there are good reasons to think that there will be differences in the features of this cognitive capacity depending on the domain in which it is exercised. In particular, we will discuss differences between how mindreading is exercised for legal purposes and in everyday situations.

## 4. *Is mindreading modular?*

Cognitive science is often described as an interdisciplinary attempt to explain how (human) cognition works (THAGARD 2005, IX; BERMÚDEZ 2014, 88-90). This description is not inaccurate, but it is also not very informative. Specifically, it does not put into view an important assumption made by cognitive scientists according to which cognition is not a unified phenomenon: it is not only useful, but also correct to break down cognition into different cognitive capacities such as the capacities to perceive colours, shapes and movements, the capacity to mindread or the capacity to interpret behaviour. These distinctions are useful because, instead of tackling one general and quite complex issue, cognitive scientists study more specific and manageable ones.

Most of all, however, this partitioning of cognition seems to be a correct assumption about how the mind works. Numerous empirical and theoretical studies of the mind suggest that the mind is divided into parts which are sometimes called "mental modules" (CARRUTHERS 2006). We will focus on three features of mental modules: specialization, inaccessibility and automatism. Specialization means that each module deals with problems of a particular type. Inaccessibility refers to the fact that what happens within a module cannot be monitored from the outside. The upshot of inaccessibility is automaticity: once modules are provided with input, they process it until completion without any outside influence. These features are not an all or nothing matter. A particular module may be more or less specialized in the sense that the problems that it deals with may be more or less diverse. More inaccessible modules will be less available for external monitoring in comparison to more accessible modules. Finally, it will be more difficult to influence what happens inside highly automatic modules in comparison to modules that operate less automatically.

The more modular information processing associated with a given cognitive competence is, the less flexible this competence will be, in the sense that it will adapt more poorly to circumstances in which too much is demanded of it. Consider vision, for example. We know very well that reality consists in much more than medium-sized objects which lie around us. Still, our knowledge about reality cannot change the fact that we perceive the world in this particular way. The problem with perception is that it is not very flexible and it adapts poorly to investigations aiming to determine what kinds of things there are in the world different from medium-sized objects which lie around us. You need to go beyond what you perceive in order to construct an adequate image of reality. In the case of mindreading, the less flexible this competence is, the more poorly it will adapt to circumstances in which too much is demanded of it. So it is worthwhile to investigate whether there is any evidence that the information processing associated with mindreading is modular. If such evidence exists, it will provide us with preliminary reasons and a conceptual framework to

investigate a more specific issue: how mindreading works if it is used for the purposes of legal interpretation and management of behaviour.

However, mindreading does not look to be associated with information-processing of a modular nature. On the contrary, this cognitive capacity looks to be closely related to general reasoning. General reasoning is one of our most flexible capacities and the information processing that is associated with it does not look to be specialised, inaccessible or automatic. This is because the information processing in question deals with problems of a diverse nature, it can access information stored in various parts of the mind and we may control it. Information processing associated with general reasoning looks to be a good example of a non-modular process.

Discussing the relationship between mindreading and general reasoning in more detail, it may appear that mindreading is just a special case of general reasoning. It may appear that when we attribute mental states we use our general reasoning capacity—that is, the reasoning capacity we use to solve problems in various domains—for the purpose of mental state attribution. The reasoning required to decide whether Jones—who took someone else's umbrella—was aware that he was taking someone else's umbrella does not look very different from the reasoning required to decide, for example, which animal left the footprints that we observed when walking through a forest—and in the latter case there is no need for mental state attribution. Thus, if mindreading is only a subspecies of general reasoning, then the information processing associated with the former is not specialized.

What is more, mindreading looks to be associated with information processing which is accessible from the outside. When we attribute mental states we can consciously access what we know about the relationship between mind and behaviour. Mental state attribution does not look to be automatic either. If an actor in a play convincingly pretends despair, initially you may truly believe that this person actually suffers despair. But if you remember that what you see is a play you may override your initial reaction and attribute to the actor the correct mental state which is pretend despair.

In sum, it appears that information processing associated with mindreading does not seem to be modular, which means that this cognitive capacity is a rather flexible one. Thus, we should not worry very much that mindreading will be particularly prone to fail if too much is demanded from it in comparison to cases in which we demand too much from one of our most flexible cognitive capacities which is general reasoning. However, if we go beyond these appearances concerning mindreading, we will observe that things look very differently. To illustrate that mindreading and general reasoning capacity are associated with different kinds of information processing—which means that they are separate cognitive capacities—and that the information processing associated with the former is much more modular than it looks at face value, we will briefly discuss the cognitive-scientific research on two issues related to mindreading: its neural basis and the autism spectrum disorders.

Research on the neural basis of mindreading indicates that the brain regions processing information associated with mindreading are different from the brain regions processing information associated with general reasoning capacity. The brain regions associated with the former include, in particular, dorsomedial prefrontal cortex (DMPFC) and temporoparietal junction (TPJ). The brain regions associated with the latter include, in particular, lateral prefrontal cortex (LPFC) and lateral posterior parietal cortex (LPPC) (Saxe & Powell 2006; Lieberman 2013, 116 f.).

Broadly speaking, to determine which brain regions process information associated with a given cognitive competence, researchers measure which brain regions are more active when people solve problems requiring this competence. These measurements are made with the aid of neuroimaging devices, notably functional imagining devices which measure brain activity in a particular period of time. A typical problem that involves mindreading consists in deciding whether Jones—who took someone else's umbrella—did it by mistake or on purpose. As for the

problems which require general reasoning, cognitive scientists often identify these problems as those which require that a person consciously holds and updates numerous pieces of information. The cognitive capacity to consciously hold and update numerous pieces of information is called "working memory". Although not all problems which involve working memory involve general reasoning—the latter cognitive capacity is more specific as it is associated with consciously holding and updating numerous pieces of information in an ordered manner: forming conclusions from premises—, all problems which involve general reasoning involve working memory.

One of the early neuroimaging studies illustrating the dissociation between information processing associated with mindreading and information processing associated with working memory involved participants reading three types of sentences (FLETCHER et al., 1995; other studies showing differences in how the brain processes linguistic information about mental states and linguistic information about non-mental phenomena include ROTTSCHY et al. 2011, TURKELTAUB et al. 2003, CASTELLI et al. 2002). The first type, "mindreading stories", described what happened to various fictional persons and their understanding required mindreading. One such group of sentences told a story about a burglar who dropped a glove while running past a police officer. The police officer yelled at the burglar to stop so that the burglar could take his glove back. However, the burglar assumed that the police officer wanted to arrest him and gave himself up. To understand the burglar's behaviour, participants needed to attribute to him the false belief that the police officer knew about his crime and wanted to arrest him. Understanding the other two types of sentences—"physical stories" and "unlinked sentences"—did not involve mindreading. Among physical stories was a story about a burglar breaking into a jeweller's store who steps on something soft which turns out to be an animal. The animal runs away and sets off the alarm. Examples of unlinked sentences included «Jill repeated the experiment, several times» or «She took a suite in a grand hotel». To control whether participants understood sentences belonging to each of the three types they were asked questions about what they read.

When participants read the sentences, they were scanned by a neuroimaging device which showed that understanding the three types of sentences was associated with three different patterns of brain activity. The most significant difference in brain activity was observed in the case of mindreading stories and unrelated sentences. Brain activity was more similar in the case of mindreading stories and physical stories, but in the case of mindreading stories increased activity was observed in a highly circumscribed region of the brain: DMPFC and TPJ. These regions were quiet in the case of physical stories and unrelated sentences. What is more, brain regions associated with working memory, LPFC and LPPC, were relatively quiet in the case of mindreading stories in comparison to physical stories and unrelated sentences.

Recall the three features of modular information processing: specialization, inaccessibility and automatism. The above-mentioned empirical research suggests to what extent the information processing associated with mindreading shares these features. First, it looks as though this information processing is specialized. That is, it looks as though there is a cognitive module that deals only with mental state attribution. This is corroborated by the fact that a highly circumscribed part of the brain shows greater activity when people solve problems of a particular type that involve mindreading and this brain region remains quiet when people solve problems of the same particular type but which do not involve this cognitive capacity. For example, this brain region is active when people read written sentences the understanding of which involves mindreading and it remains quiet when people read written sentences the understanding of which does not involve this cognitive capacity. This suggests that the cognitive module associated with mindreading comes online only in very specific circumstances: when the task at hand involves mental state attribution.

Second, the above-mentioned empirical study suggests that the cognitive module associated with mindreading is, at least to a certain extent, both inaccessible and not subject to the control of general reasoning. Recall that when participants in the above-mentioned experiment solved

problems involving mental state attribution their working memory was not activated. If this was the case, then they did not consciously monitor the process which resulted in, let's say, their attribution to the burglar the mistaken belief that the policeman knows about the robber's crime. This accounts for the claim that the mindreading module may be partially inaccessible to general reasoning—perhaps participants did not monitor this process because it was inaccessible to their general reasoning. Going further, if participants' working memory was offline when they exercised their mindreading capacity, then they did not influence what happened inside their mindreading module via their general reasoning. This accounts for the claim that the mindreading module may not be subject to the control of general reasoning and in this sense automatic. For it may be the case that participants did not influence what happened inside their mindreading modules via their general reasoning capacity because it was not possible for them to influence this module in this way.

It may appear that accessibility to general reasoning and controllability by means of general reasoning are associated with each other in the sense that the former is a necessary condition of the latter: in order for general reasoning to influence what happens inside the mindreading module, a person needs to be able to consciously monitor what happens there. This needs not to be the case, however. A module may not be accessible for conscious monitoring and still remain under the influence of general reasoning. In such cases, the influence of general reasoning will not be direct—in the sense that the premises in general reasoning will not concern what happens inside the mindreading module because this would require accessibility to this module—but it will be indirect. Indirect influence of general reasoning on the mindreading module consists merely in the fact that results of general reasoning are available to this module as an input. That requires that the mindreading module has the capacity to access the information processing associated with general reasoning and monitor what happens inside.

The drawback of the experimental study discussed above is that the mental state attribution problem that the participants needed to solve in order to understand the stories that they read was relatively simple. It may well be the case that it is because of its simplicity that the participants did not engage into general reasoning to determine, for example, what was the robber thinking when he gave up. Other, more complicated mindreading problems may not be possible to deal with unless one explicitly reasons about which mental state needs to be attributed.

These more complicated mindreading problems are often encountered when mindreading is used for the purpose of legal interpretation and management of behaviour. Think of a lawyer who wants to determine whether someone, let's say Wilson, committed fraud. This type of crime requires that Wilson intentionally deceives his victim in order to unlawfully obtain some gain. Thus, interpreting Wilson's behaviour as an instance of fraud requires attributing to him several mental states such as: a belief that what he told his victim was false, a belief that by deceiving his victim he will obtain a particular gain and a desire to obtain this particular gain. In effect, substantiating the claim that Wilson committed fraud requires an explanation of why it is appropriate to attribute to Wilson these particular mental states. Explaining this point requires that one engages in careful consideration about why this attribution is supported by the available evidence. This consideration will heavily draw on the working memory and general reasoning capacity of the interpreter of Wilson's behaviour.

If in more complex cases mindreading is linked with general reasoning, then this could be taken to show that information processing associated with mindreading is not modular. For in these more complex cases, it looks as though there are no two separate information-processing mechanisms underpinning mental state attribution—one associated with mindreading and the other associated with general reasoning—, but only one such mechanism which looks to be accessible to general reasoning and subject to its control. What is more, this mechanism does not look to be domain-specific, because it encompasses information processing associated with general reasoning which is not, by definition, limited in its purposes. One could also argue that there is no qualitative

difference between simple and difficult mindreading problems in the sense that in simple cases information processing associated with mindreading is accessible to general reasoning and subject to its control, but because these cases are simple general reasoning is not required. In the case of simple mindreading problems people could, however, access the information processing mechanism associated with their mental state attribution if they were explicitly asked to do so.

The idea that information processing associated with mindreading is not modular in the sense that it can be accessed and controlled via general reasoning is tempting if only for its simplicity. The upshot of this idea is that there is no truly separate mindreading capacity. In this view, complex mindreading problems, such as the problems encountered in the domain of legal interpretation and management of behaviour, require using general reasoning capacity in a specific domain which is mental state attribution. However, there are empirical findings that undermine this idea. These findings concern autism spectrum disorder (ASD) which is associated with impaired mental state attribution.

In one such experiment, there were four groups of participants: high-functioning children with autism, children with general intellectual impairment, normally developing 8-years old children, and adults (Abell et al. 2000). The three groups of children were matched for their verbal age, which means that children in the first and second group were 3-4 years older than children in the third group. Participants were shown animations of geometrical figures some of which interacted with each other as if one figure responded to the mental state of another figure. For example, in one such animation one triangle was trying to persuade another triangle to leave an enclosure. This means that the former triangle was aware that the latter triangle was reluctant to leave and the former made an attempt to change the latter's mind. After watching the animations participants were asked: "What happened in the cartoon?" and their descriptions were evaluated for accuracy. In the case of animations that required mindreading, participants' descriptions were evaluated for the accuracy of mental state attribution.

Results have shown a marked difference between high-functioning children with autism and the other three groups. More specifically, the accuracy of mental state attribution in the case of children with autism was significantly lower in comparison to all the other groups. For example, in the case of the animation showing one triangle trying to persuade another triangle to leave the enclosure, one child with autism said that «they are trying to push each other and want to kiss one another». What is more, results of this experiment corroborated previous findings showing that a characteristic feature of high-functioning individuals with autism is not that they use mentalizing descriptions less frequently in comparison to non-autistic individuals, but that the former are less accurate in their mental state attributions than the latter.

These results are of significance to our considerations because they go a long way towards showing that mindreading and general reasoning are separate capacities and that the information processing of the former is modular in the sense that it is domain-specific as well as inaccessible to general reasoning and not subject to its control. Observe that the mindreading problems that the participants in this experiment encountered were complex in the sense that these problems required general reasoning. Participants needed to come up with an explanation that best accounted for the available evidence on their own. However, even though the general reasoning capacity of the children with autism who took part in the experiment was significantly higher in comparison to the general reasoning capacity of the children with general intellectual impairment—as it was attested by the IQ tests the children were administered—the former group of children still shown a marked inaccuracy in their mental state attributions. This suggests that in the case of children with autism a more domain-specific information processing is impaired and that this domain-specific information processing is associated with mindreading. What is more, these children could not compensate for their impaired mindreading capacity and inaccurate mental state attributions with their superior general reasoning capacity which suggests that information processing associated with their mindreading capacity was inaccessible to their general reasoning and not subject to its control.

## 5. *Reflexive and reflective mindreading systems*

Taking stock of the above considerations, we may hypothesize that the information processing associated with mindreading is composed of two systems: a reflexive mindreading system and a reflective one. The reflexive system is highly modular in the sense that it is domain-specific—it is designed to deal with mental state attribution only—and it is highly inaccessible to general reasoning and largely not subject to its control. The reflexive system processes information whenever you encounter problems that require mental state attribution. If these problems are simple, the mental state attributions made by this system will be sufficient to solve them. It is the reflexive system that is responsible for the fact that you immediately make sense of information that you hear or read that requires mindreading to understand. The reflective system comes online when more complex mindreading problems are encountered and it uses the information processing associated with general reasoning. The reflective system is less modular than the reflexive one because the former uses resources associated with a domain-general capacity. This means that in the case of the reflective system the process which leads to mental state attribution is accessible to general reasoning and subject to its control. Importantly, however, the reflective system is not capable of dealing with mindreading problems on its own. It always requires input from the reflexive system as the latter is responsible for providing the necessary semi-finished products to the former. These semi-finished products are intuitive attributions—that is, mental states which are attributed by default when mindreading problems are encountered. We will call these outputs of the reflexive system "mindreading intuitions" or, in short, M-intuitions. How important M-intuitions are in the case of mental state attribution is illustrated by the fact that if the reflexive system is somehow impaired and its outputs are flawed it may be difficult for the reflective system, provided with these flawed M-intuitions, to correct them and produce accurate mental state attributions—as it is the case with individuals with autism spectrum disorder.

In the light of above considerations, our immediate concern should be the reflexive mindreading system. In particular, we should be concerned with its algorithm—the set of instructions that this system follows in order to generate M-intuitions. The reason for this concern is that even formally immaculate reasonings will lead to false conclusions if their premises are false. Thus, even the most advanced reflective mindreading system—correctly utilizing complex algorithms for deductive or probability reasoning—will generate inaccurate mental state attributions if the M-intuitions that it is provided with are flawed. Another reason why we should be concerned with the reflexive system is that, due to its modularity, we may expect that it will be difficult to monitor and influence this system so that it no longer produces flawed M-intuitions.

There is a broad consensus among cognitive scientists that the reflexive system has two components—a theory of mind component and a mental simulation component—which work together in order to generate M-intuitions (Stich & Nichols 2003; Goldman 2006; Stueber 2006). A theory of mind (ToM) is a body of knowledge about mind and behaviour that is stored in memory and used by default to solve mindreading problems. This body of knowledge is not innate and it grows as the individual develops and learns new information relevant to attribute mental states.

One of the most interesting findings pertaining to the development of ToM is that in the course of early childhood it grows in a strikingly similar fashion across different individuals. One such finding is that 4-years old children begin to understand that other people have beliefs and that these beliefs may be wrong. Younger children, on the other hand, tend to think that others always conceive the world as it actually is. This is illustrated by the fact that these younger children tend to think that if they themselves have seen an event happening, then other people will also be aware that this event happened—even individuals who cannot be aware of this because the event happened in their absence. Later on, however, increasing individual differences may be observed in the content of the ToMs of different subjects and in the ways in

which they use it. Some of these differences are culturally driven. For example, individuals grown up in Far East cultures—such as Chinese, Japanese and Korean—are less likely to interpret others' behaviour by appealing to their mental states in comparison to individuals raised in Western cultures, such as Americans and Europeans. Individuals raised in Far East culture will be more likely to make sense of others' behaviour by appealing to the features of the environment which influenced this behaviour.

Individual differences in the content of ToM and in the way that it is used do not need to result from different cultural backgrounds. We may hypothesize that some of these differences will be observed across individuals belonging to much smaller groups—perhaps even to such a small group as people working in a particular profession. This is because the differences in question may result from the fact that individuals who work in certain professions will be required to add to their ToMs new information which will be of little relevance outside this profession. In effect, it will be unlikely that this information is included in ToMs of individuals who work in other professions. What is more, the new information which some individuals are required to add to their ToMs may be even at odds with ToMs of individuals working in other professions. These considerations are relevant to our purposes because legal professions require adding to one's ToM new information which may not only be of little relevance outside the legal domain but sometimes it may even be at odds with ToMs of non-lawyers.

## 6. *Mindreading and the legal interpretation and management of behaviour*

A well-known piece of information included in the ToMs of many individuals who work in a legal profession that goes beyond and is at odds with the ToMs of non-lawyers is that in certain cases people may be attributed oblique intentions. Oblique intention is a mental state which, if attributed to an individual, means that an individual intended to bring about a particular event because he recognized that this event may be one of the consequences of his action. The upshot of introducing oblique intention into law is that, from the legal point of view, an individual may be attributed an intention to bring about an event even though he did not want to bring about his event.

To illustrate in more detail what is essential to oblique intention, consider *Hyam v. DPP* (1975), a famous case in English law (Ormerod & Laird 2020, 98). Mrs. Hyam, who was jealous of her former partner's new fiancé, Mrs. Booth, poured gasoline into Mrs. Booth's letter box, ignited it and left home without warning anyone about the fire. Unfortunately, the fire spread, burned down Mrs. Booth's house and killed her two daughters. Mrs. Hyam was tried for murder. In the trial, Mrs. Hyam claimed that her intention was only to frighten Mrs. Booth and cause her to leave the town. Thus, according to Mrs. Hyam's defense, she had no intention of hurting anyone. Mrs. Hyam was convicted, however, and appealed. Eventually, the case was decided by the House of Lords where the appeal was refused. Lord Hailsham claimed that in order to attribute an intention it is sufficient that the defendant

> «knew there was a serious risk that death or serious bodily harm will ensure from his acts and he commits those acts deliberately and without lawful excuse with the intention to expose a potential victim to that risk as the result of those acts. It does not matter in such circumstances whether the defendant desires those consequences or not».

Although oblique intention is a common feature of contemporary criminal law systems, non-lawyers will be uncomfortable with the above claim by Lord Hailsham. This is because according to non-lawyers' ToMs—which they use in ordinary, everyday attributions of intention—intention requires that an individual desires a particular event to occur and believes that this event will be the consequence of his action (Malle & Nelson 2003, 573). On this

account, undesired side-effects cannot be intended. According to a non-lawyer's ToM, it makes even less sense to attribute intention for an unexpected side-effect in the case of which the individual is consciously unaware that his action will bring it about. Yet, despite the oblique intention's peculiarity from the non-legal perspective, lawyers will often intuitively attribute this mental state to an individual if the relevant facts of the case at hand are similar to the facts in *Hyam v. DPP*.

The second component of the reflexive mindreading system is mental simulation. Mental simulation consists in using one own's decision-making mechanism in the production of M-intuitions. In the case of mental simulation, this decision-making mechanism is used in an offline manner which means that the M-intuitions that it produces do not influence how the simulator behaves. Instead, these M-intuitions are decoupled and attributed to someone else or provided to the reflexive mindreading system for further processing.

In the early days of cognitive-scientific research on mindreading many authors thought that the reflexive mindreading system is driven exclusively by a ToM. However, various kinds of considerations—grounded in empirical findings and conceptual developments—suggested that this simple view is difficult to hold. To mention one of such considerations, observe that if the reflexive mindreading system were limited to using only a ToM, it would be very difficult for this system to deal with even simple mindreading problems. To be more specific, if this were the case, then many mindreading problems would be computationally quite demanding in the sense that solving them in real-time would require a lot of resources such as time, memory or processing power.

To illustrate this point, consider again *Hyam v. DPP*. It was mentioned that lawyers will often intuitively attribute oblique intention to an individual if the relevant facts of the case at hand are similar to the facts in *Hyam v. DPP*. But in *Hyam v. DPP* oblique intention attribution to Mrs. Hyam was by no means intuitive. The judges in this case carried out explicit and detailed considerations whether it is appropriate to attribute oblique intention to Mrs. Hyam. And the jury was carefully instructed to consider this issue as well. This indicates that it was the reflexive mindreading system that was responsible for the attribution of oblique intention to Mrs. Hyam. However, a closer look at the court's considerations suggests that there was a different mental state which was intuitively attributed to Mrs. Hyam: the knowledge that her actions make it highly probable that anyone who lived in Mrs. Booth's house may get seriously injured. Recall that Mrs. Hyam denied that she intended to hurt anyone. However, she also said that when she burned the letterbox she realized that what she did was dangerous to anyone living in the house and added that she thought that the house was empty. Mrs. Hyam's statement that when she committed the action she realized how dangerous its consequences may be was the ground for the intuitive attribution to her of the above mentioned knowledge that her actions make it highly probable that anyone staying in Mrs. Booth's house may get seriously injured.

What suggests that this attribution was intuitive is that it is made without any explicit consideration—the court does not offer any explanation why it is appropriate to attribute to Mrs. Hyam this knowledge apart from mentioning that she admitted being aware that what she did was dangerous to anyone living in the house. This intuitive attribution was provided as an input to the reflexive mindreading systems of the judges and the jury for further processing where it played a key role in the carefully considered attribution to Mrs. Hyam the oblique intention to kill Mrs. Booth's children.

Consider how the reflexive mindreading system would need to operate to generate the M-intuition that Mrs. Hyam knows that it is highly probable that her actions may cause serious injuries to Mrs. Booth's children if this system was driven exclusively by a ToM. If this was the case, then this system would need to make a series of inferences that ultimately resulted in this attribution. The premises in these inferences would be, on the one hand, the relevant psychological generalizations stored within the ToM—that is, generalizations that link mind

and behaviour relevant for attributing the above-mentioned type of knowledge to Mrs. Hyam—and, on the other hand, the facts of the case which relate to these generalizations. According to one such generalization, for example, if someone says that when she acted she realized that her action may harm someone else, then this person acted knowing that it is highly probable that her action may harm someone else.

A generalization of this kind was probably used in *Hyam v. DPP* in order to attribute to Mrs. Hyam that she knew about the consequences of her action. This generalization is not explicitly mentioned in the court's considerations about what Mrs. Hyam knew. Instead, it is an implicit premise in the inference from what Mrs. Hyam said to what she knew. However, taking into account the facts of the case, it is by no means obvious that this generalization is applicable. Notice that Mrs. Hyam mentioned that she thought that the house is empty. This undermines the claim that she knew that it is highly probable that her actions may cause serious injury. If she believed that the house was empty, then she could not simultaneously believe that she may harm someone inside the house. To deal with this apparent inconsistency between Mrs. Hyam's beliefs, an appeal to further generalizations is required which would exclude Mrs. Hyam's belief that the house is empty from the set of premises in the inference aiming to determine what she knew. What is more, Mrs. Hyam explicitly denied that she intended to kill or hurt anyone and she also denied that she was aware that she could kill anyone. What she admitted is only that she realized that her actions may be dangerous to those inside the house. This is further evidence that it is by no means clear that Mrs. Hyam knew that it is highly probable that her actions may cause serious injury or even kill those who were staying in fact inside the house. Even more generalizations are required to show why it is appropriate to attribute to Mrs. Hyam this knowledge.

In short, if the reflexive mindreading system was driven exclusively by ToM, then this system would need to apply numerous generalizations stored within ToM to finally attribute to Mrs. Hyam the mental state under discussion. This, however, would place a heavy burden on this system in the sense that it would need to process a lot of information before inferring Mrs. Hyam's mental state. To sum up, assuming that the reflexive mindreading system is driven exclusively by ToM it is surprising that the judges and the jury in *Hyam v. DPP* were capable of intuitively—that is, without an explicit consideration—attributing to Mrs. Hyam the knowledge that that it is highly probable that her action may seriously injure those inside the house.

According to a more plausible explanation of how this intuitive mental state attribution was made, the reflexive mindreading system did this in a much less complicated way. Namely, this system accessed and used the same information-processing resources which are used when the individuals who do the mindreading experience the kind of mental state which they attribute to Mrs. Hyam. In this way, the cognitive burden on the reflexive mindreading system is significantly diminished. This system no longer needs to compute every psychological generalization, stored into ToM, relevant to the case at hand: it just generates, in those who engage in mindreading, the pretended or imaginary mental states which are supposed to correspond to those of the target of the simulation. In other words, mental simulation allows the mindreader to realize what his own mental states would be if he were to find himself in the situation of his target. The idea behind mental simulation is that instead of using an elaborate ToM, the simulator takes a shortcut and uses her own mind as a model of the mind of the simulation's target. Thus, to attribute to Mrs. Hyam the knowledge that it is highly probable that her actions may injure or kill someone, the jury and the judges took a shortcut and solved a simpler mindreading problem: whether they themselves would be aware that it is highly probable that their actions may injure or kill anyone, if they were in Mrs. Hyam situation. The result of this mental simulation was then attributed to Mrs. Hyam.

Some researchers found it tempting to suggest that mental simulation is all there is if you want to explain how mindreading works. This view, however, is also problematic. The main difficulty here is that if the results of mental simulation are to be plausible, the simulator needs

to adjust for the differences between himself and the target of the simulation. For example, in order to simulate another individual, the simulator needs to adjust for the possible difference between how he views a particular situation and how this situation is viewed by his target. What is more, the simulator needs to adjust for the difference between what he would want to achieve in this situation and what his target wanted to achieve. In short, the simulator needs to adjust for any relevant differences in beliefs, desires, emotions and other mental states between himself and his target. These adjusted mental states serve as the input to the simulation process which generates output in the form of a pretended set of mental states. However, the adjustment in question cannot proceed without the simulator knowing which differences in mental states between himself and his target are relevant to the case at hand as well as knowing what do these differences amount to. For example, adjusting for a difference in beliefs between the simulator and his target requires that the simulator knows what his target believes. Thus, if a mental simulation is to generate plausible outcomes, it needs to be augmented with a ToM in order for the simulator to adjust for the differences between himself and his target.

Taking into account the proposal that both mental simulation and ToM are involved in intuitive mental state attributions, we may propose the following algorithm for these attributions:

(1) adjust for the differences between yourself and the target of your mental state attribution,
(2) imagine what mental states you would have if you found yourself in the situation of your target,
(3) attribute the imagined mental states to your target (STUEBER 2006, 120).

Thus, the above-mentioned claim that, in order to attribute to Mrs. Hyam the knowledge that it is highly probable that her actions may injure or kill someone, the jury and the judges imagined whether they themselves would be aware that this would be the case if they were in her situation, does not completely account for this attribution. What is missing in this description is that their reflexive mindreading systems needed to adjust for the differences between themselves and Mrs. Hyam. One such difference could be that—to give a perhaps not so hypothetical example—the discrepancy between their belief that if they were in her situation, it would be obvious to them that their actions could lead to the deaths of those living in the house and Mrs. Hyam's lack of this belief or even her belief that her actions will surely not cause anyone to die. Noticing differences of this kind requires that the reflective mindreading system applies information stored in its ToM and attributes to the target the discrepant mental states.

## 7. *Why and how mindreading fails*

The previous discussion concerning how intuitive mental attributions are made can be used to predict when these attributions will be incorrect. This happens when the interpreter fails to adjust for some important difference between his own mental states and the mental states of his target. In the case of such a failure, the unadjusted mental states will be provided as input to the simulation process and this process will generate simulated mental states which do not correspond to the mental states of the target person. Still, despite their lack of correspondence, these mental states will be intuitively attributed by the interpreter to his target.

Taking into account that adjusting for the difference between the interpreter and his target involves applying the information stored in a ToM, this adjustment will fail if there is information about mind and behaviour that is relevant in a particular case of mental state attribution and which is not included in this interpreter's ToM. One important type of information about mind and behaviour that may not be included in the ToM of a person who

uses mindreading for the purposes of legal interpretation and management of behaviour is information provided by scientific research.

It is perhaps not too far-fetched to assume that the two dominant sources of information about mind and behaviour stored in this person's ToM are ordinary or folk psychology and its legal counterpart. The former consists of numerous psychological generalizations recognized by most people, such as the principle according to which a person has beliefs that may be mistaken, desires which motivate her to act or the principle that if a person says that when she carried out a particular action she realized that the consequences of this action may be dangerous to others, then she acted with a belief that the consequences of her action may be dangerous to others. These folk psychological generalizations are used in everyday situations as well as for the purposes of legal interpretation and management of behaviour. On the other hand, the legal counterpart of folk psychology consists in generalizations designed specifically for legal purposes. Many of these legal-psychological generalizations are in essence elaborations of posits of folk psychology such as the legal-psychological generalization that strong emotional disturbance inhibits control over action. Other legal-psychological generalizations appear to be refinements or developments of folk psychology such as the legal distinction between premediated and unpremeditated actions. In some cases, however, generalizations for mental state attribution of a legal nature may be at odds with what folk suggests as it is the case with oblique intention.

A characteristic feature of both folk psychology and its legal counterpart is that for the most part their principles are fixed and developed through linguistic considerations about how words associated with mind and behaviour ought to be used. That is, because in ordinary cases it makes sense to say that if someone admitted that he realized what will result from his actions, then he knew what will result from his actions—and it makes little sense to deny him this knowledge unless there are some circumstances which undermine his admission—, we think that we are allowed to attribute to him this knowledge. And because for the purposes of legal interpretation and management of behaviour it makes sense to say that if someone knew that it is highly probable that his action will result in someone else's injury, then he intended this injury, we think that we are allowed to attribute to him this intention.

What scientific research shows, however, is that there is a lot of information about how mind and behaviour are related which will go unnoticed even for those who are most careful in their linguistic considerations of these matters. As it was mentioned, a lot of this research focuses directly on how cognition works instead of studying people's competence with ordinary words associated with psychology. In the case of mindreading, a lot of this research is focused on the features of the reflexive mindreading system which is not really available for investigation for someone equipped only with linguistic competence. To support this claim with an example, consider the psychological research on memory—in particular, the research on misinformation effect.

Misinformation effect occurs when a person's memory of an event is distorted due to information that she acquired after this event occurred. Much of the research on this phenomenon was concerned with how false or misleading information acquired after experiencing an event influenced eyewitness testimony. In one of such experiments, participants were shown a film of a car accident (Loftus 1979). Afterwards, the participants were asked to describe what they have seen in the film. One group of participants was asked the question «About how fast were the cars going when they smashed into each other?». Another group was asked the question «About how fast were the cars going when they hit each other?». Subjects in the first group estimated that the car was moving faster than subjects in the second group. What is more, one week after seeing the film both groups were asked the following question about the film: "Did you see any broken glass?"—even though there was no broken glass in the film. Still, there were participants in both groups which responded affirmatively to this question and participants in the first group were

much more likely to provide this answer. As we can see, participants' memories were influenced by how the questions were formulated and what they were about.

Empirical findings pertaining to the misinformation effect were surprising because memory was not considered to be malleable to information acquired after the event—or at least it was not considered to be malleable in this way and to this extent. However, we can explain the surprising nature of these findings by appealing to the way in which our mindreading works—in particular, to the operation of the reflective mindreading system. Observe that deciding whether someone's testimony is reliable is an exercise in mental state attribution. This decision consists in attributing to the witness a set of beliefs about what he experienced in the past—a set of beliefs that correspond to his testimony—and determining whether these beliefs are true or false. Judging this last issue involves appealing to various generalizations stored in one's ToM about how memory works. One such generalization, which is probably shared by most, is that recent events tend to be better remembered than events that occurred in the distant past. Thus, if one is to judge whether a particular testimony is reliable, an issue under consideration will be how recent was the event that the testimony is about. If the event occurred in the distant past, we will be more willing to question the witness' reliability in comparison to cases in which the event was recent. Perhaps this willingness may even be observed at the level of the reflexive mindreading system. This means that intuitively—that is, without explicitly considering this issue—we will be more willing to judge a witness' reliability in accordance with the generalization that recent events tend to be better remembered than events that occurred in the distant past. However, even if this is will not always be the case—for example, because this generalization is not stored in the ToM which is accessible to the reflexive mindreading system of a particular person—it may be possible to apply this generalization via the reflective mindreading system and assess whether the intuitive attribution is correct.

Misinformation effect proved to be surprising which illustrates that the psychological generalization according to which a person's memory of an event may be distorted due to information that she acquired after this event occurred was not stored in the ToM's of most mindreaders. This concerns ToMs which were applied both intuitively and reflexively. The observation that before the empirically oriented research on memory revealed its shortcomings the legal systems relied too much on eyewitness memory shows that this generalization was not recognized by the law as well. At the present moment, however, this state of affairs appears to change and knowledge about how post-event interventions influence memory formation is more commonly applied in legal settings. We may expect that future empirical research on mind and behaviour will result in discoveries of more generalizations which will need to be included in the ToMs of those who mindread for the purposes of legal interpretation and management of behaviour.

*Further readings on mindreading in law:*

BROWN T. R. 2022. *Demystifying Mindreading for the Law*, in «Wisconsin Law Review», 1, 1 ff.

GREELY H.T. 2013. *Mind Reading, Neuroscience, and the Law*, in MORSE S.J., ROSKIES A.L. (eds.), *A Primer on Criminal Law and Neuroscience. A Contribution to the Law and Neuroscience Project, Supported by the MacArthur Foundation*, Oxford University Press.

GREGORY D. 2019. *Judging the Mental States of Others: 'Mindreading' in Legal Decision-Making*, in «Jurisprudence», 11, 1, 48 ff.

*References*


Abel F., Happe F., Frith U. 2000. *Do Triangles Play Tricks? Attribution of Mental States to Animated Shapes in Normal and Abnormal Development*, in «Cognitive Development», 15, 1 ff.

Bermúdez J.L. 2018, *Cognitive Science. An Introduction to the Science of the Mind*, Cambridge University Press.

Castelli F., Frith C., Happé F., Frith U. 2002. *Autism, Asperger Syndrome and Brain Mechanisms for the Attribution of Mental States to Animated Shapes*, in «Brain», 125, 8, 1839 ff.

Carruthers P. 2006. *The Architecture of the Mind. Massive Modularity and the Flexibility of Thought*, Oxford University Press.

Fletcher P. C., Happe F., Frith U., Baker S. C., Dolan R. J., Frackowiak R. S., Frith C. D. 1995. *Other Minds in the Brain: A Functional Imaging Study of "Theory of Mind" in Story Comprehension*, in «Cognition», 57, 109 ff.

Goldman A. 2006. *Simulating Minds. The Philosophy, Psychology, and Neuroscience of Mindreading*, Oxford University Press.

Hutto D. 2008. *Folk Psychological Narratives. The Sociocultural Basis of Understanding Reasons*, MIT Press.

Lieberman M. 2013. *Social. Why Our Brains Are Wired to Connect*, Oxford University Press.

Loftus E. 1979. *Eyewitness Testimony*, Cambridge University Press.

Malle B. 2004. *How the Mind Explains Behaviour. Folk Explanations, Meaning and Social Interaction*, MIT Press.

Malle B., Nelson S. 2003. *Judging* Mens Rea: *The Tension Between Folk Concepts and Legal Concepts of Intentionality*, in «Behavioral Sciences and the Law», 21, 563 ff.

Ormerod D., Laird K. 2020. *Smith, Hogan, & Ormerod's Text, Cases, & Materials on Criminal Law*, Oxford University Press.

Rottschy, C., Langner, R., Dogan, I., Reetz, K., Laird, A. R., Schulz, J. B., Eickhoff, S. B. 2011. *Modelling Neural Correlates of Working Memory: A Coordinate-Based Meta-Analysis*, in «NeuroImage», 60, 830 ff.

Saxe R., Powell L., 2006. *It's the Thought That Counts: Specific Brain Regions for One Component of Theory of Mind*, in «Psychological Science», 17, 692 ff.

Spaulding S., 2018. *How We Understand Others. Philosophy and Social Cognition*, Routledge.

Stich S., Nichols S. 2003. *Mindreading. An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*, Oxford University Press.

Stueber K. 2006. *Rediscovering Empathy. Agency, Folk Psychology and the Human Sciences*, MIT Press.

Turkeltaub P. E., Gareau L., Flowers D. L., Zeffiro T. A., Eden G. F. 2003. *Development of Neural Mechanisms for Reading*, in «Nature Neuroscience», 6, 7, 767 ff.

Thagard P. 2005. *Mind. Introduction to Cognitive Science*, MIT Press.