VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

21
Mental causation, folk psychology, and rational action explanation

Tomi Kokkonen

When I give a reason for someone's action, do I identify a cause for the person's external behaviour? This is one of the issues within the multitude of philosophical problems that are bound together as "the problem of mental causation". This particular issue stems from the fact that in our folk-psychological practices, we seem to both rationalize the action and to give it a causal explanation at the same time – but, given the causal exclusion of physical reality and the non-reducibility of rationalizations to physics, this cannot be the case. Therefore, something has got to give. The now-standard solution to the problem, pioneered by Panu Raatikainen (2007 & 2010; see also Menzies 2007 & 2008; Shapiro & Sober 2007; Woodward 2008), uses the interventionist idea of causation and the contrastive theory of causal explanation (Woodward 2003) to argue that the reasons for action indeed identify the cause, while identifying the underlying physical processes do not answer to the relevant causal explanatory question. While I sympathize with this solution, I will argue that since our folk-psychological explanations are inherently ambiguous, there is no solution to *the* problem of mental causation, but a clarification of the issue leads into a more multi-layered explication of mental causation events. The standard solution gives an account to one but only one of the issues.

Kokkonen, T. (2025). Mental causation, folk psychology, and rational action explanation. In A. Laitinen, M. Keinänen, J. Reinikainen & A. Honkasalo (eds.). *Language, Truth, and Reality*. Tampere University Press. (pp. 261–279). https://doi.org/10.61201/h5fbdk64

The ambiguity of folk psychology

Folk-psychological descriptions such as "I opened the can because I wanted to drink what is inside the can" are ambiguous on whether they refer to *propositional attitudes ascribable to a person as a whole* or *representational states participating in cognitive processes within the person's cognition*. They seem to do both at the same time, which is one source for the problems we have with mental causation. I suggest, however, that these are two different functions that the folk-psychological descriptions have, but our everyday practices need not to distinguish between the two functions, so they do not do so. As a preliminary, let us take a look at what these practices are supposed to be about.

There are several theories of propositional attitudes and the nature of folk psychology and several possible ways to classify them. For current purposes, I will group philosophers of mind into *cognitivists*, who think that propositional attitudes are causally effective psychological states (for example, Fodor 1981; Kim 2005; Shea 2018), and *ascriptionists*, who think that propositional attitudes are the states that we attribute to agents. The ascriptionists include interpretationists (for example, Davidson 1963; Anscombe 1967; Dennett 1987) and dispositionalists (for example, Ryle 1949; Marcus 1990). The theories put forward by these philosophers are attempts to give the semantics for folk-psychology and account for how it relates to the causal structure of the world. However, if we approach folk psychology as a natural human practice (see Bogdan 1997; Gopnik & Meltzoff 1997; Wellman & Liu 2004; Hutto 2007 & 2008; Call & Tomasello 2008; Duval et al 2011; Henry *et al* 2013; Zawidzki 2013; Andrews 2012; 2015a & 2015b; Apperly 2020), it may be that no philosophical theory is a correct explication of what folk psychology.

Consider the following example. A person, call him Aaron, is drowning. Another person, call her Bea, sees him in trouble. Bea jumps to the water and saves him. She does this because she wants to, and she has no further external aims or reasons for doing so, such as glory or gratitude. This seems to be an altruistic act. However, when we take a closer look at Bea's motivation, we might get more confused. There is a debate in psychology whether the perceived distress in others (Aaron in this case) can be a directly motivational factor (*the empathy-altruism hypothesis*) or does the motivation always go through some self-regarding process, such as the distress that Bea undoubtedly feels in the situation, caused by the perception of Aaron in distress (see Batson 2011). The latter view holds that all helping is always motivationally selfish. This is not a conceptual issue: if the latter is true, motivation to help can be blocked by blocking the agent's distress, but if the empathy-altruism hypothesis is correct, Bea would help Aaron anyway. But does it really make sense Bea's action would be selfish even if the egoistic theory of human motivation is correct? Let us assume that Bea is Aaron's mother. Seeing him drowning creates distress, sure, but she might try to help him even to the point of self-sacrifice. How would this be selfish? We can make a distinction between the opposition between egoism and altruism on the "deeper" psychological level of description (that has to do with the psychological

motivation mechanisms) and the opposition between egoism and altruism on the agentive level of description (that has to do with the aims of the action; see Kokkonen 2021, chapter 6 for a throughout discussion). The need to make this distinction is clear in this example where the folk-psychological understanding of egoism and altruism becomes inherently contradictory, but the distinction itself is between two general levels of description.

Folk-psychological concepts may refer, then, to both holistic states of an agent (call them *agentive level* states) and entities within the cognition (*psychological level* proper), even at the same time, since the practice itself does not recognize the difference between the two. (The debate on the correct level of description for the reference of folk-psychological concepts is also a debate on how to understand intentionality, so I prefer to use the term "agentive" instead of "intentional.") Furthermore, there is no *need* to make the distinction between the two levels in folk-psychological practice. Here I will follow the Pluralistic Folk Psychology idea of Kristin Andrews (2012, 2015a & 2015b), the view that folk psychology is inherently pluralistic as a theory and as a practice, whether in the psychological processes involved in the practices, what its function in social life is, or what the references of its core concepts are. I have discussed both philosophical and empirical reasons for thinking this is the case elsewhere (Kokkonen 2021, chapter 5), but I will discuss some philosophical issues as a prelude to my argument for the nature of mental causation now. My aim is to discuss how the two levels of description interact in folk psychology, not to present an alternative interpretation for the correct level of description of folk psychology. Later, I will discuss how this issue fundamentally changes the issue of mental causation.

The different intuitions about the nature of folk psychology in philosophy – and these are not only theoretical disagreements about folk psychology but also normative disagreements about what the philosophical theories using its conceptual framework should be about – may be symptomatic of its pluralistic nature. On one hand, folk psychology seems to have theory-like characteristics: we explain, predict and manipulate others' future behaviour by manipulating their mental states. Mental states, whatever they are, seem to work as if they are causal factors, according to the Woodwardian understanding of causality. At the same time, propositional attitudes are intentional and rational, and this seems to be foundational for the semantics of folk psychology. These are different aspects of human agency and psychological phenomena related to it that are unified in folk psychology for pragmatic reasons. Daniel Dennett's distinction between intentional, design, and physical stances (Dennett 1987) is one way to make sense of this and to make ascriptionism compatible with a causal interpretation of psychology proper. In this view, the ascribed states are abstracted properties of the system, rather than parts of the system, and thinking of them as parts with causal role would be a category mistake. According to Dennett's (1991) metaphor, beliefs and desires are more like the centres of gravity than the concrete states of a mechanism.

All this seems somewhat vague, however, and there have been more recent attempts to analyse how intentionality arises from brain processes in a more

detailed way from the representationalist perspective. One idea is that intentional states can be understood as robust outcome functions that have been stabilized by evolution and learning (for example, Godfrey-Smith 2006; Sterelny 2015; Shea 2018). These processes are controlled by sub-personal sub-systems, and their functional operations have representational content. Person-level attributions of beliefs and desires, however, are robust states of the individual (or the whole "system") that describe their cognitive relations with the world, descriptive and directive, and these relations are constituted by the parts of the representational system. This is plausible and I will not challenge the idea as such. However, it would still be a category mistake to reduce the states of the system to the parts of the system. Beliefs and desires are dependent on the system of representations, not parts of it. If the robust outcome function approach works, it explains the constitution of systemic states, but this does not build a conceptual link between the levels of description. What I suggest, instead, is that both agentive and psychological level are sensible levels of analysis, even if we also understand intentionality and representations at the psychological level. The latter is a separate question asking what explains, on the cognitive/psychological level, agentive-level intentionality – if anything does.

There are also interesting alternatives to the representational theory of mind within the naturalistic context that are still compatible with describing humans as agents. The most extreme is radical enactivism (Hutto & Myin 2013 & 2017), which takes the biological processes outside the central processing system more seriously as part of cognition. It proposes that much of cognition lacks any representational content at all and has more to do with how the sensory-motor system functions as a whole. The so-called "4e movement" (enactive, extended, embodied and embedded; see Newen, De Bruin & Gallagher 2018) approaches to mind and cognition in general challenge the classical representational theory of mind. But even if this approach provided the correct account, this would not make intentional action descriptions inadequate on the agent level. The debate between representationalists and their critics is not about attributing mental states to agents but about how the cognition works. The latter includes the issue of what explains the applicability of folk-psychological attributions to human agents, but this is a different issue, which highlights the need for the distinction.

Furthermore, and even more importantly, the precise relation between psychological and agentive levels is more difficult to understand with directive mental states than with descriptive ones. Psychological-level descriptive representations and agentive-level beliefs can be thought of as being in a complex constitutive relation. But how the drives and motivational salience relate to agentive-level pro-attitudes is trickier. Motivational salience is a crucial explanatory component in the emergence of pro-attitudes, but it is difficult to see how it alone could have the right kind of propositional content. It is simply a causal factor, incentivising or aversive, that instigates behaviour. Motivational salience explains preferences in part but is not itself a preference with content. Furthermore, pro-attitudes are about particular goals, not behavioural tendencies towards or away from a type of behaviour in a type

of context, which motivational salience entails. The goals implied by a pro-attitude may be quite general and abstract, of course, such as world peace, being famous, or whatever goals moral values entail even prior to knowing these entailments. Folk psychology also accommodates moods and personality traits as more general and robust dispositional states. However, these are not the same as a tendency to be motivated by a certain type of things in certain contexts. This is also evident in how folk psychology is inadequate in capturing mental episodes such as depression. Depression has effects on individuals that make it difficult to rationalize their behaviour. The origin histories of depression cannot be fully understood in terms of folk psychology, either – that is, we cannot always give a rationalizing reason for being depressed, and it may be dangerously misleading when we try. Depression simply is not a reason-like state nor a collection of reasons or desires, and neither explaining depression nor explaining with depression is a rationalizing explanation (see Goldie 2007).

To sum up this part, there seem to be several philosophical and scientific reasons to distinguish the different frameworks conceptually, whatever their relation turns out to be. Without this revision, it looks like rationalizations of behaviour causally explain it, which seems to be both true (we *do* explain people's behaviour using reasons as if they were causes) and false (agentive descriptions do not refer to causal processes). If we make the distinction, the nature of the problem changes. The (rationalizing) agentive and (causally explanatory) psychological levels of description are also connected in various ways (for example, there is causally efficient rational deliberation that uses folk-psychological categories in reflection, and individual rationalizations have causal presuppositions) but folk psychology as a practice does not make the distinction or provide the tools to discuss how exactly the levels are connected. Consequently, folk-psychological explanations cannot be used directly in more sophisticated action explanation – philosophical, psychological, or evolutionary. I will take a closer look at this proposition now.

Rationality and rationalization

An essential source for philosophical difficulties (as well as the connectedness between psychological and agentive levels) is rationality. The agentive description attributes reasons to agents. The relationship between reasons to each other and the action is rational. However, the rationality of action seems to presuppose some sort of rationality in the causal processes that produce behaviour if agentive descriptions are given a causal explanatory role. Rationality itself cannot be a causal factor, but the causal processes must have systematicity in their functioning that exhibits behaviour that we perceive as rational. Furthermore, rational deliberation about the goals and means to achieve them is a part of human psychology, not just a property of action attributions. Attributing rationality to human psychology seems to be unavoidable.

There are, however, different concepts of rationality that may be applied to action and should not be conflated, especially if we are interested in their connection to causal explanation of behaviour. I will call the notion of rationality used in the philosophical theory of action *agentive rationality*. It is the idea that there is a reason for action. The action is rational when it is in accordance with the goals and beliefs of the agent in a way that can be expressed as giving the action a reason. There are both descriptive and normative elements in this: the action can be described as intentional by giving it a rationalizing conceptualization, but rationality is also evaluative in the sense that we consider action itself appropriate or not, given the reasons it was taken (see McGeer 2007; O'Brien 2019). Rationality may come in degrees in the sense that the action may be more or less appropriate, but *it is a qualitative property of an action that it can be given a reason*. It is about the intelligibility of behaviour as the action of an agent. The proposition that humans are rational agents is a categorical proposition about rationalization, both in its descriptive and normative dimensions. If humans are rational in this sense, rationalization is an adequate way to conceptualize humans and human behaviour. The normativity of rationality in this sense is what makes human action rational or *irrational*, while some other animals, for example, are not rational or irrational but *arational* (see Hurley & Nudds 2006). In contrast, the notion of rationality used in cognitive science, call it *cognitive rationality*, is a quantitative measure of cognitive capacity – but it is, likewise, also a normative notion. Rationality is measured against the *chosen optimality model,* which specifies what counts as rational, either in the *epistemic* (belief-formation) or *instrumental* (decisions about which course of action to take given the context) sense, and the *degree* of rationality and irrationality in human action and thinking is evaluated by comparing the performance to the model (see Stanovich 2011 & 2012).

The two senses of rationality, and the notions of normativity accompanying them, are different. The philosophical analysis of folk psychology uses the agentive notion. It is supposed to capture something that is *constitutive* of agency. Its normativity is about the adequacy of action given its reasons, and the failure to be rational is a failure to be an agent and for the behaviour to be intelligible as human action (see O'Brien 2019). Cognitive rationality and its normativity are instrumental: there are models that we *choose* to represent optimal decision in a context, *given the aims of the agent*, and we compare the behaviour to this. Moreover, these models (and the concept of rationality) could be applied to non-intentional systems, too, such as those animals that we consider not to be intentional, and to Artificial Intelligence systems. Agentive rationality does not imply any specific model of cognitive rationality. Moreover, it cannot be assumed that agentive rationality implies any specific *degree* of cognitive rationality that would enable *agentive rationality itself* to be an explanatory factor for behaviour, for example (see also Henderson 1993; Ylikoski & Kuorikoski 2016).

The two notions of rationality are also related. The models of cognitive rationality are *meant* to be about what a rational agent would ultimately choose, given their goals. This implies a third notion of rationality, *normative rationality*: how one *should* reason and choose action, given the goals. This is a stronger notion of rationality than

the one used in rationalization of action – the assumption (rational) agency is not an assumption complete rationality. However, although this is a more demanding normative notion of rationality than the other two concepts in their normative component, the normativity of normative rationality is *instrumental:* it depends on chosen goals and acknowledged constraints on achieving these goals. This is the notion of rationality for fields such as Decision Theory and does not concern us here. However, the ability to be a rational agent in the agentive sense requires some cognitive capacities that explain it. Cognitive rationality is a measurement of how well some of the cognitive capacities function in certain tasks, and having these capacities is a partial explanation for why humans are agentively rational. These capacities are what we should be interested in when causally explaining human behaviour and how it fits with the causal understanding of human behaviour that humans are also (agentively) rational. I will refer to the agentive notion as "rationality" from now on, unless otherwise specified.

An essential feature of folk-psychological explanations is that they rationalize the behaviour into actions that have reasons behind them and goals to look forward to. Reasons (and their constituents, beliefs and desires, the "two directions of fit," as Elizabeth Anscombe (1967) put it, descriptive and directive) are connected to each other and to the action in rational relations: the propositional contents entail other propositional contents and are attributed to agents as holistic sets. At the same time, the attributions identify what action the behaviour is, and the identification of the behaviour as doing *x* is a part of the interpretation of which beliefs and pro-attitudes of the agent constitute their reason for action in the situation. That is, the action descriptions are a part of the same holistic net of semantic connections as the mental states that make behaviour intelligible. Some philosophers take this to mean that rationalizations cannot be causal explanations, since semantic entailments are not causal relations (Anscombe 1967; von Wright 1971; Sehon 1997 & 2005). Others think that this merely makes the ontology of action somewhat anomalous (Davidson 1970). It seems that folk-psychological practices require the attributions to have at least some causal counterfactual power: the point of persuasion and reasoning with a person, for example, is to change their underlying structure of desires and beliefs to affect their future behaviour. This is a causal intervention, not a matter of interpretation after the fact. Mental attributions should not be causal attributions, under some conceptual and metaphysical considerations, but they seem to function as if they were. Hence the attempts to reduce the rationalizing elements into something that also has psychological reality. (See Henderson 1993; Crane 1995; Mele 2000; Kokkonen 2011.)

Furthermore, folk-psychological practices seem to presuppose that of all the reasons we can attribute to the agent, give the agent's mental states, there is a *primary* reason among them that is why the agent actually did what they did. It determines what the action was about – it is not just an alternative description for the behaviour. How should we understand this? For a causalist like Davidson, the primary reason is the one that caused the action. Under the ascription view, this is a problem known

as *Davidson's Challenge*: a mere ascription is only about pattern fitting, it does not explain action (see Davidson 1963; Mele 2000; O'Brien 2019). For the causalists the problem is how the reasons can be causes. This problem can be broken into two parts. First, how can mental states in general (that is, agentive states under the description of folk-psychological conceptualization) be causes of physical behaviour? I call this the *Core Problem of Mental Causation*. Second, how could rationalization of action reliably capture states that are causally efficient for behaviour? In other words, how can *reasons*, identified by their modal and logical properties, be causal? I call this the *Hard Problem of Action Explanation*.

The hard problem of action explanation

The recently popular solution to the core problem of mental causation has been to use the contrastive theory of causal explanation and the manipulationist theory of causation as a framework to identify causal factors (Raatikainen 2007 & 2010; Menzies 2007 & 2008; Shapiro & Sober 2007; Woodward 2008). Folk-psychological descriptions make robust but imprecise claims about causal processes and behavioural dispositions of the agents on which the behaviour depends, with relevant counterfactual contrasts. These robust states are the states of the agents. There may be psychological states that implement these more or less directly and have causal relations with other psychological states and the behaviour, and similarly with neural states – but these are further issues. What matters is that the mental descriptions identify states that have intelligible contrast classes, and the difference between the explanatory state and its contrast class is a difference-maker between the explained behaviour and its contrast class. The *explananda* and *explanantia* need not be described on the same level, for as long as the framework identifies the correct dependence relation. In other words, reasons can be causes when having a reason is the adequate identification of a causal disposition.

Furthermore, the contrast classes of explanation may be different when referring to the causal process on different levels. In fact, given that we attempt to explain behaviour that is specified with a goal, a folk-psychological description (including reasons and intentions) may be a *more* adequate way of identifying the contrast class than an alternative explanation on a different level (see Raatikainen 2010). This seems to solve the causal explanatory part of the problem regardless of what the relationship between agentive states and the underlying causal processes may be. Moreover, it grants autonomy to causal explanations on different levels and fits the general pluralistic approach adopted here. Some issues remain untouched with this solution, however. These include problems such as the ontological relation between the objects of the different descriptions. More importantly, this solution does not touch the issue about the role of rationality and rationalization itself (the Hard Problem): how can we discover causes of behaviour by rationalizing action (or rather: can we do so, and how can we justify this practice)?

The problem has two components. First, how is it possible that humans are natural beings whose behaviour is a part of the causal structure of the world but follow the dictates of rationality at the same time? (*The role of rationality in naturalism*.) Second, is rationalization a form of (causal) explanation? There are only two possible solutions to the first part: some sort of anomalous monism (Davidson 1970), or that humans are not actually as rational as rationalization practices presuppose. There are good empirical reasons to think that humans are not fully rational when it comes to *cognitive* rationality (see Kahneman, Slovic & Tversky 1982; Kahneman 2011; Gigerenzer 2007; Stanovich 2011 & 2012). As discussed above, the agentive notion of rationality is a different notion, but there is a substantial connection between the two notions in *explaining* rationality. If rationality itself does not have causal powers (and it follows from the naturalistic premises that it does not) there must be something in the psychology that explains this. Rational deliberation is a part of how mind works, and it has causal consequences, but the empirical research seems to imply that it plays a limited role in cognition. This also implies limitations on the extent of agentive rationality in humans. (See also Henderson 1993; Ylikoski & Kuorikoski 2016.) Partial rationality (whether it is because of deliberation or something else) may be enough to justify the interpretative practices, however, and it is not an unsolvable problem for a naturalistic view of humans. Humans have complex cognitive systems adapted to survive flexibly in complex, changing environments. A part of this process has been the decoupling of representations from what is immediate, and this has also created a need to represent states of affairs as related to each other and make inferences between them (Godfrey-Smith 1996; Sterelny 1999 & 2003). In other words, humans have evolved psychological processes that are causal (and implemented by neural processes) but deal with representational states in a way that is partially rational, since the psychological mechanisms have been selected for having rational outcomes. But this rationality is relative to selected tasks and their proper contexts, and even there it is limited by how reliably rational the outcomes the underlying biological structure can produce. There is no selection for universal rationality. Even if there was, an organ such as the brain could not produce universal rationality through causal operations. Then again, we are not universally rational. This solution also makes the rationality of human behaviour and psychology (to the extent that it is rational) an *explanandum* itself – *rationality* (the entailments between propositional attitudes) does not explain rationality of behaviour. *Rationality is a part of descriptions of the behaviour to be explained*.

There are explanations for partial rationality. Folk psychology is a crucial part of our social practices and the cognitive skills related to them, and it evolved to be functional for the many different needs of our many kinds of social interaction (Byrne & Whiten 1988; Bogdan 1997; Corbalis & Lea 1999; Tomasello 2009; Emery 2012; Devaine *et al* 2014). The need for effective mindreading for various social activities to be possible, entails selection pressures on our behavioural tendencies too, as well as the "control structures" in our cognition, to be more in accordance with the kind of rationality that we use as a guide in folk psychology (Sterelny 2015). Furthermore,

the folk-psychological practices, the language related to them (see Gopnik & Meltzoff 1997; Zawidzki 2013), and the agent-based narrative structure we learn in childhood (see Hutto 2008) affect our thinking. They do have not only mindreading but also *mindshaping* functions – they are an extra-genetic form of inheritance to shape our behaviour and its underlying psychology to be in line with folk-psychological assumptions, as suggested by Matteo Mameli (2001) (see also Zawidzki 2013; Sterelny 2015). Moreover, folk psychology has regulative and justificatory functions in social interaction (Andrews 2015a & 2015b; see also McGeer 2007; Zawidzki 2013). All this makes rationality understandable from a naturalistic point of view as far as it is limited, but rationality as such does not play an explanatory role in why we think and act rationally.

This still leaves us with the second problem: How could rationalizing with a reason itself be an explanation? One possible solution would be to revise the non-causalist stance on attribution of mental states by proposing that psychological states (in the narrow sense) are references to causal states, but rationalizations are about agentive states. I have already alluded to something like this as the first approximation. But making this distinction alone would cut the connection between rationalizations and causal explanations, and rationalizing attributions seem to work as attributions of causal factors, as discussed earlier. We could go even further: to reason to rationalize action in the first place is only because it captures something causal that is useful to us. Moreover, it would leave us with Davidson's Challenge: the notion of primary reasons, the intended reasons for action, require some further explanation if intention is not causally effective (see Mele 2000; O'Brien 2019).

A causal presoppositionalist account of rational action explanation

Consider the following option. It is not the *reasons* the agent has that are manipulated in an interaction, but something that *having the reasons depends on* (that is, something causal that can be described on the psychological and/or neurophysiological level). If the connection between the reasons and their underlying conditions is sufficiently robust, reasons identify causal relations, albeit under an imprecise description. Reasons are attributed by rationalization, and they depend on psychological processes. This would be a form of anomalous monism that is not anomalous, given there are explanations available for why the two are correlated. However, this is not a sufficient solution. Describing intentional action involves ascribing an *intention* to the agent, not just rationalizing reasons for action that can be interpreted for the agent: some reasons express what the agent intends to do, and these intention references are clearly meant to capture something causal (Davidson 1978; Bratman 1987; Mele 1992 & 2009). And as Elizabeth Anscombe (1967) (albeit a non-causalist herself) pointed out already, we also seem to have direct knowledge of our own intentions. Our knowledge of all the factors that play roles in why we do what we do may be fallible, but the experience of intending to do something specific is direct,

not a process of interpretation. Within the causal interpretation, we identify some of our reasons as causes. Furthermore, we do not just act and interpret the action; we reason about our goals and the means to achieve them, and this reasoning seems to make some causal contribution to producing behaviour. Hence, there seems to be a connection between agentive rationalizations and causal psychological processes.

It is, however, one thing to say that we have more intuitive understanding of ourselves as agents than a mere interpretation and another thing completely to say that this understanding involves direct observations of the causal processes that guide our behaviour. We are only conscious of a part of our cognitive processes and motivations for action. Cognitive and social psychologists distinguish two kinds of processes in mind (the so-called *dual process* and *dual system* theories of cognition): Type I (or System 1) and Type II (or System 2). Type I processes are *automatic*; they are fast, reactive, non-conscious, associative, heuristic, and effortless. Type II processes are *analytic*; they are slow and effortful but controlled and deliberative. (See Kahneman, Slovic & Tversky 1982; Evans and Over 1996; Bargh & Chartland 1999; Stanovich 1999 & 2011; Kahneman 2011; Gigerenzer 2007; Frankish & Evans 2009; Evans & Stanovich 2013.) These processes (or systems) are jointly activated, and they give a rise to more complex cognitive operations, but only some processes are conscious, and we are (indirectly) aware of only some of the non-conscious processes. We have no access to all the processes that influence our thinking, even our conscious thinking. When people are asked about the reasons for their actions, they do not identify an *effective motivation* behind them, but describe a *state with a goal*, and this may be just as much a rationalization after the fact as if they were explaining another person's action, even in highly deliberative contexts such as making a moral judgment (Haidt 2001; see also Nisbett & Wilson 1977; Nisbett & Ross 1980; Bargh & Chartrand 1999).

As mentioned earlier, the notion of rationality used in cognitive science is different from the one used in the analysis of folk-psychological conceptualizations, although there are substantial connections. There are two properties of the two-level cognitive system that are consequential for the issue at hand. First, the analytic processes that we are conscious of and constitute our deliberation are the ones we identify as our thinking and decision-making in our cognitive phenomenology. We experience other states too, such as emotions, and we are usually aware that we have other psychological motivating factors, but reasoning is what we consider to be our "actual" thinking and we have an impression that it is responsible for our decision-making. We can disregard the normative, gradual notions of cognitive rationality for a while and concentrate on some of the qualitative aspects of the analytic processes. First, they process propositional contents: this part of cognition is closest to what folk-psychological rationalization presumes human thinking to be like. Second, our thinking and decision-making involves the non-conscious processes as well, even while we deliberate, and they have inputs into the deliberation. When we deliberate, we become aware of the products of non-conscious processes as our own thoughts (even if we do not have access to the processes producing them), and they become a

part of further deliberation. Third, the agentive rationalizing attributions to agents (as whole persons) do not distinguish between these two kinds of processes.

If folk psychology is pluralistic both in its mechanisms but also in its reference, this extends to self-reflection. When we reflect our motives and decisions, we attribute rationalizing intentional states (desires, beliefs, reasons) to ourselves according to folk psychology. However, the sources for these states include both the deliberative process and the other processes that participate in guiding our thinking and behaviour. If this is the case, the object of reflection on our own mental states is a combination of deliberative conscious states, products of non-conscious processes that we are aware of and that we interpret in folk-psychological categories, and quasi-theoretical assumptions about ourselves that are folk-psychological postulates. Our self-understanding is fallible regarding these differences. Even if our self-attributions of mental states are correct in terms of folk psychology in the moment of action, and even if they are based on epistemically reliable self-observations, our *justificatory* self-rationalization does not necessarily identify the *causal* processes of how we came to the decision correctly. Furthermore, we are not necessarily correct in our self-attributions either, and our self-observations are not always reliable.

However, reflection is not mere rationalization. Sometimes we explain our own behaviour with non-rational causes, such as anger, sorrow, or intoxication. But the point here is that sometimes we also misidentify having non-rationalizable psychological processes as having reasons. Conscious reasoning (as a part of cognition) and interpretation using the theory of mind (on the agentive level) are confused in the simplified image of rational agency, and they should be distinguished. Moreover, although we experience intending and identify it correctly as the motivational state that triggers action, the content (the reason, or a plan) we accompany it with may sometimes nevertheless be an interpretation within a folk-psychological conceptual framework, not an experience of a deliberated state. There are also problems with prediction of one's own behaviour: people are notoriously bad at predicting their future actions based on their current self-perceived states of minds – although it is not clear whether this is because of misinterpretation of one's own motives or underestimating the situational factors that are not present in the context of prediction (Poon, Koehler & Buehler 2014).

Having an intention (in the sense of intending) does, however, presuppose that there is at least one causal factor that is identified in experiencing intending. Psychologically speaking, we experience motivational forces, aversive and incentivising saliences that guide our behaviour. A successful agentive explanation does not need to specify these processes precisely to be a form of causal explanation. But a successful agentive description must include a reference to the *existence* of such factors. The identification of an intention in the context of action involves attributing a reason that adequately describes the agent's relation to the world in a robust way in the context, given both her epistemic states and active motivational forces. Moreover, even if we think the agent *knows* what they are doing (or what their intention is), this does not require them to know all the psychological processes involved. On the

other hand, when I reflect my own motives, or an external observer of the situation wonders what it is that I am doing, the observation or speculation (depending on which one is doing it) targets the *psychological states*, but this may still use the same *goal-directive semantics*.

Rational deliberation is a *part* of our cognitive capacities. It is a part of the causal makeup of mind, not just a passive reflection of cognition. But reason, in this sense, is not a determinative factor. At the same time, the object of rationalization is the action, not a partial factor of it: we use folk psychology to represent our own holistic agentive states, such as beliefs and desires, in metacognition (whether in conscious deliberation or in automatized processing, which also has metacognitive functions). We do not represent just the reasoning part of our cognition, although this is the part we mostly identify our thinking with, and we tend to conflate the two – the contents of reasoning and the contents of the holistic states. Folk-psychological categorizations affect how we deliberately plan our actions, but once again, this is causal influence of folk psychology on our cognition – it does not make agentive and psychological states the same. The same applies the other way around; not all behaviour needs to be produced by deliberation alone to be rationalizable in the sense of agentive rationality. Much of the unconscious, automatized processing has a positive function in reaching the chosen goal of action (see Bargh & Chartrand 1999; Gigerenzer 2007; Marewski, Gaissmaier & Gigerenzer 2010).

Re-thinking mental causation (again)

We can summarize the outcome of the discussion in this paper so far in the following propositions about rationality and action: (1) Reasoning (as rational deliberation) is a cognitive process that participates in the causal production of behaviour. (2) People are conscious of this part of their own cognitive processes, while the other processes manifest only in the products of these processes. (3) The non-conscious parts of cognition are often instrumental to the chosen goals, and therefore they participate in producing the action that we rationalize *without being a part of the rational guidance* of the action on the cognitive level. (4) People rationalize both their own and others' actions within the folk-psychological framework and this rationalization has more to do with justification and evaluation than causal explanation, but it functions both ways. (5) People conflate the *rationalization* they apply to their action and the *experienced intention* that triggers this action *whether it is the outcome of rational deliberation or some other process*. It can be either. To the degree that non-rational processes are instrumental to chosen goals, this does not make a difference in understanding the action as guided by reasons. But giving only reasons in the causal explanation misidentifies the causes. (6) People are aware of non-rationalizable causes such as emotions and use them in the folk-psychological explanations as well, and these explanations are not rationalizations. Emotions, personality characteristics, reasons and other factors are not separated as being on

different levels; but there is only one folk-psychological "level". (7) Sometimes, people have no idea what motivated them, and their rationalization of their own action is simply incorrect.

If these conclusions are accepted, agentive, rationalizing descriptions do not refer directly to psychological processes with causal powers, but they *presuppose* that there are causal processes that are responsible for the action in order for the folk-psychological practices to work (see O'Brien 2019). In these practices, agentive attributions of reasons and attributions of psychological states proper that underlie the agentive states are mixed into a heterogeneous category, and the distinction between them would not make a difference. The connection between agentive ascriptions and the underlying psychology is strong enough to allow rational arguments, persuasion and other folk psychology-based practices to enter the cognitive system and influence behaviour. In philosophical and scientific scrutiny, however, different levels of description need to be acknowledged. Slices of the causal process that result in behaviour can be described on any level, although they do not make the same causal explanatory claims (since they have different contrast classes) and sometimes there is no rationalizing action explanation at all – that is, when the behaviour under scrutiny is irrational. For psychological and philosophical purposes, however, the two levels should be kept apart.

As I mentioned above, the problem of mental causation breaks into two sub-problems that I referred to as the Core Problem and the Hard Problem. The classic Davidsonian problem of mental causation was about the relationship between the physical (causal) domain of regularities that humans as natural entities follow and the rational level of action explanations that refer to reasons. This relationship has three steps in total. The first step is how biological design emerges from physical regularities. This is well understood, but it is worth noting that the causal basis in mental causation is not physics and regularities in its processes, but rather neurobiology which is constituted by physical processes but also has evolved functional structure. The second issue is how the psychological level (cognitive and conative) descriptions are related to neurobiology. We may have some understanding about this on empirical grounds, as I alluded above. The third step is how rationalizable states emerge from the psychological mechanisms and processes – and, again, we have some idea how this works, once we keep the two levels distinct and do not conflate them into one level of "mental descriptions". As for the problem of mental causation, I will make the following conjectures. First, the solution to the Hard Problem (that is, how could rationalizations reliably capture causal states) is that they do not. Not reliably – but often enough to make folk-psychological practices useful most of the time, for the reasons discussed above. Second, the solution to the Core Problem (that is, how can mental states be causes) is that when the folk-psychological statements refer to the psychological level proper, we can understand them as functional descriptions of brain processes and mechanisms that get their semantics partially from folk psychology, but when they do not refer to them (given the answer to the Hard Problem), they do *not* refer to causes but make a presupposition of an existence of

a causal structure with the right kind of effects. Rather than identifying the cause, as in the first case (in the accordance with the standard solution), they identify an effect that, nevertheless, gives us information about what the action is about. Given the pluralistic nature of folk psychology, there is not *the* solution to the problem of mental causation.

If the account put forward here is correct, the standard solution to the problem of mental causation, put forward by Raatikainen and others, is not wrong, but it is imprecise and does not always capture the correct causal structure. However, it gives the correct ontology on what kind of events mental causation events are. It also captures the logic of causal explanations of action in the cases in which reason-based explanations are causal. It is also worth noting that introducing the standard solution was a groundbreaking shift in the discussion on mental causations and the account given here is also building upon it, by adding detail and incorporating research and discussions on folk psychology in other fields.

References

Andrews, Kristin (2012): *Do Apes Read Minds? Toward a New Folk Psychology*, Cambridge, Ma: MIT Press.

Andrews, Kristin (2015a): 'Pluralistic folk psychology and varieties of self-knowledge: an exploration', *Philosophical Explorations* 18(2): 282–296. URL = https://doi.org/10.1080/13869795.2015.1032116.

Andrews, Kristin (2015b): 'Folk psychological spiral: explanation, regulation, and language', *The Southern Journal of Philosophy* 53(S1), Spindel Supplement: 50–67. URL = https://doi.org/10.1111/sjp.12121.

Anscombe, Elizabeth (1967): *Intention*, 2nd edition (2000 reprint), Cambridge, MA: Harvard University Press.

Apperly, Ian (2010) *Mindreaders: The Cognitive Basis of Theory of Mind*, New York, NY: Psychology Press.

Bargh, John A. & Thanya L. Charthrand (1999): 'The unbearable automaticity of being', *American Psychologist* 54(7): 462–479. URL = https://doi.org/10.1037/0003-066X.54.7.462.

Batson, C. Daniel (2011): *Altruism in Humans*, Oxford: Oxford University Press.

Bogdan, Radu J. (1997): *Interpreting Minds*, Cambrige, Ma: MIT Press.

Bratman, Michael (1987): *Intention, Plans, and Practical Reason*, Cambridge, MA: Harvard University Press.

Byrne, Richard W. & Andrew Whiten (eds.) (1988): *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes and Humans*, Oxford: Oxford University Press.

Call, Josep, & Michael Tomasello (2008): 'Does the chimpanzee have a theory of mind? 30 years later', *Trends in Cognitive Science* 12(5): 187–192. URL = https://doi.org/10.1016/j.tics.2008.02.010.

Corballis, Michael C., & Stephen E. G. Lea (eds.) (1999): *The Descent of Mind: Psychological Perspectives on Hominid Evolution*, Oxford: Oxford University Press.

Crane, Tim (1995): 'The mental causation debate', *Proceedings of the Aristotelian Society* (Aristotelian Society Supplementary), 69: 211–236. URL = https://doi.org/10.1093/aristoteliansupp/69.1.211.

Davidson, Donald (1963): 'Actions, Reasons and Causes', *Journal of Philosophy* 60(23): 685–700. URL = https://doi.org/10.2307/2023177.

Davidson, Donald (1970): 'Mental Events', in Lawrence Foster and J. W. Swanson (eds.), *Experience and Theory*, London: Duckworth.

Davidson, Donald (1978): 'Intending', in Yirmiaku Yovel (ed.), *Philosophy of History and Action*, D. Reidel and the Magnes Press, 41–60.

Dennett, Daniel C. (1987): *The Intentional Stance*, Cambridge, Ma: MIT Press.

Dennett, Daniel C. (1991): 'Two contrasts: folk craft versus folk science, and belief versus opinion', in John Greenwood (ed.): *The Future of Folk Psychology: Intentionality and Cognitive Science*, Cambridge: Cambridge University Press, 135–148.

Devaine Marie, Guillaume Hollard & Jean Daunizeau (2014): 'Theory of Mind: Did Evolution Fool Us?"' PLOS One 9(2): e87619. URL = https://doi.org/10.1371/journal.pone.0087619.

Duval, Céline, Pascale Piolino, Alexandre Bejanin, Francis Eustache & Béatrice Desgranges (2011): 'Age effects on different components of theory of mind', *Consciousness and Cognition* 20(3): 627–642. URL = https://doi.org/10.1016/j.concog.2010.10.025.

Emery, Nathan (2012): 'The evolution of social cognition', in Alexander Easton & Nathan Emery (eds.), *The Cognitive Neuroscience of Social Behaviour*, Hove and New York: Psychology Press, 115–156.

Evans, Jonathan St. B.T. & David E. Over (1996): *Rationality and Reasoning*, Psychology Press.

Evans, Jonathan S.B.T., & Keith E. Stanovich (2013): 'Dual-process theories of higher cognition: advancing the debate', *Perspectives on Psychological Science* 8(3): 223–241. URL = https://doi.org/10.1177/1745691612460685.

Fodor, Jerry (1981): *RePresentations: Philosophical Essays on the Foundations of Cognitive Science*, Cambridge, Ma: MIT Press.

Frankish, Keith & Jonathan Evans (eds.) (2009): *In Two Minds*, Oxford: Oxford University Press.

Gigerenzer, Gerd (2007): *Gut Feelings: The Intelligence of the Unconscious*, New York: Viking Penguin.

Godfrey-Smith, Peter (1996): *Complexity and the Function of Mind in Nature*, Cambridge: Cambridge University Press.

Godfrey-Smith, Peter (2006): 'Mental Representation, Naturalism and Teleosemantics', in David Papineau & Graham Macdonald (eds.), *New Essays on Teleosemantics*, Oxford: Oxford University Press, 42–68.

Goldie, Peter (2007): 'There are reasons and reasons', in Daniel D. Hutto & Matthew Ratcliffe (eds.), *Folk Psychology Reassessed*, Dordrecht: Springer, 103–114.

Gopnik, Alison & Andrew Meltzoff (1997): *Words, Thoughts and Theories*, Cambridge, Ma: The MIT Press.

Haidt, Jonathan (2001): 'The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment', *Psychological Review* 108(4): 814–834. URL = https://doi.org/10.1037/0033-295x.108.4.814.

Henderson, David (1993): *Interpretation and Explanation in the Human Sciences*, Albany: State University of New York Press.

Henry Julie D., Louise H. Phillips, Ted Ruffman & Phoebe E. Bailey (2013): 'A meta-analytic review of age differences in theory of mind', *Psychology and Aging* 28(3): 826–839. URL = https://doi.org/10.1037/a0030677.

Hurley, Susan & Matthew Nudds (2006): 'The questions of animal rationality: Theory and evidence', in Susan Hurley & Matthew Nudds (eds.), *Rational animals?,* Oxford: Oxford University Press, 1–83.

Hutto, Daniel D. (2007): 'Folk Psychology without Theory or Simulation', in Daniel Hutto & Matthew Ratcliffe (eds.), *Folk Psychology Reassessed*, Dordrecht: Springer, 115–135.

Hutto, Daniel D. (2008): *Folk Psychological Narratives: The Sociocultural Basis of Understanding Reasons*, Cambridge, Ma: MIT press.

Hutto, Daniel D. & Erik Myin (2013): *Radicalizing Enactivism: Basic Minds Without Content*, Cambridge: MIT Press.

Hutto, Daniel D. & Erik Myin (2017): *Evolving Enactivism – Basic Minds Meet Content*, Cambridge: MIT Press.

Kahneman, Daniel (2011): *Thinking, Fast and Slow*, London: Macmillan.

Kahneman, Daniel, Paul Slovic & Amos Tversky (eds.) (1982): *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge: Cambridge University Press.

Kim, Jaegwon (2005): *Physicalism, or Something Near Enough*. Princeton: Princeton University Press.

Kokkonen, Tomi (2011): 'Mielen teoria, selittäminen ja ymmärtäminen', *Tiede & edistys* 36(4): 277–290. URL = https://doi.org/10.51809/te.105048.

Kokkonen, Tomi (2021): *Evolving in Groups: Individualism and Holism in Evolutionary Explanation of Human Social Behaviour*, Doctoral Thesis, Helsinki: University of Helsinki. URL = http://hdl.handle.net/10138/333344.

Marcus, Ruth B. (1990): 'Some revisionary proposals about belief and believing', *Philosophy and Phenomenological Research* 50: 132–153. URL = https://doi.org/10.2307/2108036.

Marewski, Julian N., Wolfgang Gaissmaier & Gerd Gigerenzer (2010): 'Good judgments do not require complex cognition', *Current Directions in Psychological Science* 11(2): 103–121. URL = https://doi.org/10.1007/s10339-009-0337-0.

McGeer, Victoria (2007): 'The Regulative Dimension of Folk Psychology', in Daniel D. Hutto & Matthew Ratcliffe (eds.): *Folk Psychology Re-Assessed*, Dordrecht: Springer, 137–156.

Mele, Alfred (1992): *The Springs of Action*, New York: Oxford University Press.

Mele, Alfred (2009): *Effective Intentions: The power of conscious will*, Oxford: Oxford University Press.

Menzies, Peter (2007): 'Mental Causation on the Program Model', in G. Brennan, R. Goodin, F. Jackson, & M. Smith (eds.): *Common Minds: Themes from the Philosophy of Philip Pettit*, Oxford: Oxford University Press, 28–54.

Menzies, Peter (2008): 'Exclusion problem, the determination relation, and contrastive causation', in J. Hohwy & J. Kallestrup (eds.): *Being Reduced—New Essays on Reduction, Explanation and Causation*, Oxford: Oxford University Press, 196–217.

Newen, Albert, Leon De Bruin & Shaun Gallagher (2018): *The Oxford Handbook of 4E Cognition*, Oxford: Oxford University Press.

Nisbett, Richard & Lee Ross (1980): *Human Inference: Strategies and Shortcomings of Social Judgement*, Englewood Cliffs: Prentice Hall.

Nisbett, Richard & Timothy Wilson (1977): 'Telling more than we can know: Verbal reports on mental processes', *Psychological Review* 84(3): 231–259. URL = https://doi.org/10.1037/0033-295X.84.3.231.

O'Brien, Lilian (2019): 'Action Explanation and its Presuppositions', *Canadian Journal of Philosophy* 49(1): 123–146. URL = https://doi.org/10.1080/00455091.2018.1518629.

Poon, Connie S. K., Derek J. Koehler & Roger Buehler (2014): 'On the psychology of self-prediction: Consideration of situational barriers to intended actions', *Judgment and Decision Making* 9(3): 207–225. URL = https://doi.org/10.1017/S1930297500005763.

Raatikainen, Panu (2007): 'Reduktionismi, alaspäinen kausaatio ja emergenssi', *Tiede & Edistys* 32(4): 284–296. URL = https://doi.org/10.51809/te.104902.

Raatikainen, Panu (2010): 'Causation, exclusion, and the special sciences', *Erkenntnis* 73(3): 349–363. URL = https://doi.org/10.1007/s10670-010-9236-0.

Ryle, Gilbert (1949): *The Concept of Mind*, Chicago: University of Chicago Press.

Sehon, Scott (1997): 'Deviant Causal Chains and the Irreducibility of Teleological Explanation', *Pacific Philosophical Quarterly* 78(2): 195–213. URL = https://doi.org/10.1111/1468-0114.00035.

Sehon, Scott (2005): *Teleological Realism: Mind, Agency, and Explanation*, Cambridge, Ma: MIT Press.

Shea, Nicholas (2018): *Representation in cognitive science*, Oxford: Oxford University Press.

Stanovich, Keith E. (1999): *Who is Rational? Studies of Individual Differences in Reasoning*, Lawrence Erlbaum.

Stanovich, Keith E. (2011): *Rationality and the Reflective Mind*, Oxford: Oxford University Press.

Stanovich, Keith E. (2012): 'On the Distinction between Rationality and Intelligence: Implications for Understanding Individual Differences in Reasoning', in Keith J. Holyoak & Robert G. Morrison (eds.), *The Oxford Handbook of Thinking and Reasoning*, Oxford: Oxford University Press, 343–365.

Sterelny, Kim (1999): 'Situated Agency and the Descent of Desire', in Valerie Gray Hardcastle (ed.), *Biology Meets Psychology: Constraints, Conjectures, Connections*, Cambridge: MITPress.

Sterelny, Kim (2003): *Thought in a Hostile World: The Evolution of Human Cognition*, Oxford: Blackwell Publishing.

Sterelny, Kim (2015): 'Content, Control and Display: The Natural Origins of Content', *Philosophia* 43(3): 549–564. URL = https://doi.org/10.1007/s11406-015-9628-0.

Tomasello, Michael (2009): *Why We Cooperate*, Cambridge, Ma.: MIT Press.

von Wright, Georg Henrik (1971): *Explanation and Understanding*, Cornell University Press.

Wellman, Henry M. & David Liu (2004): 'Scaling of Theory-of-Mind Tasks', *Child Developing* 75(2): 523–541. URL = https://www.jstor.org/stable/3696656.

Woodward, James (2003): *Making Things Happen: A Theory of Causal Explanation*, Oxford: Oxford University Press.

Woodward, James (2008): 'Mental Causation and Neural Mechanisms', in Jakob Hohwy & Jesper Kallestrup (eds.), B*eing Reduced: New Essays on Reduction, Explanation, and Causation*, Oxford: Oxford University Press.

Ylikoski, Petri & Jaakko Kuorikoski (2016): 'Self-interest, norms, and explanation', in Mark Risjord (ed.): *Normativity and Naturalism in the Philosophy of the Social Sciences*, New York: Routledge, 212–229.

Zawidzki, Tad (2013): *Mindshaping: A New Framework for Understanding Human Social Cognition*, Cambridge, Ma: MIT Press.