In Favor of Mentalism in Economics: A Conversation with Christian List

February 7 and 9, 2016*

Ludwig-Maximilians-University, Munich

Forthcoming in: Herfeld, Catherine: *Conversations on Rational Choice*, Cambridge: Cambridge University Press

Catherine Herfeld: Professor List, what comes to your mind when someone refers to rational choice theory? What do you take rational choice theory to be?

Christian List: When students ask me to define rational choice theory, I usually tell them that it is a cluster of theories, which subsumes individual decision theory, game theory, and social choice theory. I take rational choice theory to be not a single theory but a label for a whole field. In the same way, if you refer to economic theory, that is not a single theory either, but a whole discipline, which subsumes a number of different, specific theories. I am actually very ecumenical in my use of the label 'rational choice theory'. I am also happy to say that rational choice theory in this broad sense subsumes various psychologically informed theories, including theories of boundedly rational choice. We should not define rational choice theory too narrowly, and we definitely shouldn't tie it too closely to the traditional idea of *homo economicus*.

Catherine Herfeld: If you take rational choice theory to be this set of approaches that are applied across various different disciplines from mathematics, to economics, political science, and other social sciences, do you nevertheless think that these different approaches have something in common, which makes them a rational choice theory?

Christian List: Part of the way in which we demarcate the field may be historical. Many of the theories that fall under the label 'rational choice theory' can be traced back to certain overlapping origins. Game theory goes back to John von Neumann, Oskar Morgenstern, John Nash, and others. Decision theory goes back to Frank Ramsey, Bruno de Finetti, and Leonard Savage, and to the history of probability theory more broadly. And social choice theory goes back to Amartya Sen, Kenneth Arrow, and even earlier to Nicolas de Condorcet. Most of the approaches that now fall under the label of 'rational choice theory', broadly construed, got at least some inspiration from these origins, though they have developed in different ways.

Even theories of boundedly rational choice and psychologically informed theories were often prompted by difficulties with orthodox rational choice theory. Their starting point was often the perceived need to give up, modify, or amend some restrictive aspects of rational choice theory while preserving other aspects. Something else that many accounts of rational choice have in common is that they are committed to a certain kind of 'belief–desire' model of intentional agency.

Catherine Herfeld: What those approaches of Kenneth Arrow, John von Neumann, and John Nash also seem to have in common is that they deviated conceptually from traditional choice theories in being grounded upon of mathematical logic. They used set theory, probability theory, and the axiomatic method in formulating their account of rational choice. Do you think that those tools from

* The interview was slightly edited in March 2018.

mathematical logic is also what unifies them?

Christian List: It is true that mathematical methods are common to all of the different approaches that fall under the umbrella of rational choice theory. But they are not all axiomatic. While in the core areas of rational choice theory axiomatic approaches are quite central, some developments in, for instance, theories of boundedly rational choice or psychologically informed decision theory have not always proceeded by focusing on axiomatization. Instead they have sometimes just proposed certain decision principles and suggested that those decision principles are reasonably accurate models of human decision making under certain conditions. Axiomatizing those decision principles is not always the main focus.

Catherine Herfeld: You mentioned bounded rationality as an example of a psychologically more informed rational choice theory. However, one might question the degree to which those approaches count as a rational choice theory. Rather, some people would argue that those accounts are alternatives to rational choice theory in that we place much stronger constrains on rationality and thereby account for the cognitive and other limitations that people confront when reasoning rationally. Where should we start drawing a line between rational and non-rational choice theories, if it is not between rational and boundedly rational choice theories?

Christian List: We should not get too worked up about drawing boundaries. At the end of the day, how exactly we demarcate rational choice theory is much less relevant than what we take to be good theories of human decision-making.

Catherine Herfeld: Do you consider Prospect Theory to be rational choice theory?

Christian List: Prospect Theory is an example of a psychologically informed theory of choice. It is a contribution to the broadly defined field I have described, though it relaxes some of the assumptions of orthodox rational choice theory.

Kahnemann and Tversky, the original proponents of Prospect Theory, observed that decision makers often violate certain classical principles of expected utility maximization. In particular, people's choices are often not invariant under redescriptions of the options. Such effects are called 'framing effects'. For example, the same person may make different choices, depending on whether the options are described in terms of losses or in terms of gains, even when the material consequences are the same. A classical theory of expected utility maximization cannot accommodate such choice patterns. The question, then, is how we can modify the classical theory so as to account for those empirically observed patterns. Prospect Theory gives us a way of doing this. It introduces the idea that choices may depend on a reference point. This allows us to capture the fact that people have different attitudes to losses than to gains.

Prospect Theory also illustrates my earlier point that axiomatization is not always the main focus. Prospect Theory offers certain formal decision principles which describe how agents make their choices. But how we axiomatize those principles is secondary. What matters is whether the proposed decision principles explain the empirically observed patterns of choices.

Catherine Herfeld: However, Prospect Theory is mostly cited as a long-awaited alternative to rational choice theory that departs from the empirical observation that people are biased in their judgement and that they do not perform the behavior that expected utility theory would predict.

Christian List: It is true that Prospect Theory is a theory of certain non-classical patterns of choice. If one understands rationality in a relatively narrow, classical sense, then of course Prospect Theory

is a theory of non-rational choice. I tend to use the label 'rational choice theory' more ecumenically, so as to include psychologically informed theories like Prospect Theory too. As I said, I am much less concerned with drawing boundaries than with the question of what would be a good theory of human decision-making.

Catherine Herfeld: In your recent paper entitled *Mentalism versus Behaviourism in Economics: A Philosophy-of-Science Perspective*, you – together with your collaborator Franz Dietrich – argue for mentalism and against behaviorism in economics. You define mentalism as the view that mental states, such as preferences and beliefs, that are used in social scientific theories, 'capture real phenomena, on a par with the unobservables in science' (Dietrich/List 2016a, 249). You define behaviourism, in contrast, as the view that those mental states 'are nothing but constructs re-describing people's behaviour' (ibid.). More specifically, you take the position that although economists are not necessarily interested in mental states per se, they have to take them on board. This is because, as economics is a science, economists have to take the theoretical entities they actually postulate in their best theories as referring to something in the world. You claim that choice theories in economics make specific ontological commitments and entities that they evoke when they use, e.g., utility functions or binary relations, do actually exist. Those theoretical concepts refer to something that we would take to be mental states. And you take the position that as long as we do not have independent reasons to doubt that our best economic theories, such as for example expected utility theory, are false, we should take their ontological commitments at face value. This is an argument for psychological realism in economics. However, it has been shown repeatedly that those economic theories of choice – while maybe our best theories – do not have a lot of explanatory and predictive power. What would be an independent reason for rejecting economic theories, and expected utility theory in particular, if it is not their empirical power?

Christian List: Classical expected utility theory is of course too simplistic. One might argue that it has already been falsified by empirical observations of human choice behaviour. However, the more sophisticated and psychologically informed versions of decision theory, which might replace expected utility theory, still involve a commitment to mental states. Expected utility theory is just the simplest example of a formal theory of choice in which beliefs and desires play a certain role. But beliefs and desires continue to play important roles in more psychologically informed theories. Prospect Theory is not my own favorite theory, but it is an example of a theory that retains a role for preferences and beliefs, just as standard expected utility theory does, albeit in a revised form. It still involves a commitment to mental states.

Let me explain the background to our argument for mentalism in more detail. First, consider the natural sciences. Our theories in the sciences, for instance in physics, seek to explain certain observable phenomena. In order to do so, they often invoke a number of unobservables. In other words, they systematicize the observations by postulating some features of reality that we cannot observe themselves or that we can observe only indirectly. On the assumption that those features exist, we are then able to explain the relevant observations. Examples of unobservables are certain physical forces and fields or the tiny elementary particles of which all matter is composed. Even phenomena as familiar as gravity and electromagnetism are not by themselves observable. We can only observe their manifestations in the behavior of certain objects under their supposed influence. Similarly, very small elementary particles are not directly observable. They can be observed only indirectly, with the help of sophisticated instruments and experimental designs. When we say that we have 'observed' these particles, this usually involves a sequence of inferences, in which we rely

on various auxiliary hypotheses and theoretical assumptions.

The philosophy of science is divided between two camps. On the one hand, there are the 'instrumentalists', who believe that these unobservables are just instrumentally useful constructs that we invoke in order to make sense of our observations. On the other hand, there are the 'realists', who think that the unobservables are real features of the world. Many philosophers of science think that realism is more compelling because the distinction between observables and unobservables is very difficult to draw in practice and it keeps shifting all the time as our instruments and experiments get more sophisticated. Microscopy now allows us to observe a lot of things that we couldn't observe in the past. Moreover, if the postulated unobservables weren't real, the success of science would be a bit of a miracle. Why would postulating gravity or the Higgs boson be explanatorily useful if these unobservables weren't real? Wouldn't this be very surprising? As Hilary Putnam once put it, realism is 'the only philosophy that doesn't make the success of science a miracle'.

Given this, many philosophers of science advocate what they call a 'naturalistic ontological attitude'. They claim that if you want to figure out which entities and properties are real features of the world, then you should consult our best scientific theories of the relevant domains. If you want to know, for instance, whether the Higgs boson exists, you should consult our best theories of particle physics. And if these theories say it exists, then so be it. Of course, the resulting ontological commitments are fallible, because we may learn that the theories are false and they may eventually be superseded. In that sense, our ontological commitments must remain open to revision.

Now, let's return to economics. Let's set aside normative economics for the moment, and focus on economics as a positive science. One of the things all the different rational-choice-theoretic approaches have in common is that they try to explain the behavior of economic agents by attributing to them certain beliefs and preferences, perhaps also certain decision principles or reasoning processes. In short, they make 'mental state' attributions. These attributed mental states have certain behavioral manifestations. For example, a theory might attribute to a market participant a particular utility function – perhaps a function of profit or something more complicated – and a particular subjective probability function that assigns subjective probabilities to various states of the world. Then it might further attribute to this market participant the decision principle of expected utility maximization or something else. In this mode of explanation, the attributed beliefs and preferences, which are the agent's mental states, play the role of unobservables. We can't observe them directly; we can't look into other people's minds. But we still have good evidence for those attributions, to the extent that they explain those agents' observable choice behavior. It may well be that the most parsimonious way to explain someone's choices is to attribute to him or her a certain subjective probability function and a certain utility function. This attribution is warranted to the extent that it allows us to explain the agent's choices. In short, we attribute mental states to an agent in order to explain the agent's behavior.

We can now ask what the status of the attributed mental states is. One possibility is that they are just useful theoretical constructs. This would be analogous to the instrumentalist view in science. But this view suffers from some well-known problems, including the constantly shifting boundary between what is observable and what is not, and the fact that it makes the success of science a bit of a miracle. Like in the natural sciences, it is plausible in economics to adopt a realist view, according to which we should take the commitments of our best theories at face value, provided those theories are sufficiently well confirmed and we are reasonably confident in them. Of course, if our theory of

people's market behavior is poorly confirmed, then we should not take its commitments too seriously. It's the same in the natural sciences: if we have got a non-standard theory of particle physics and it performs poorly at predicting and explaining the relevant observations, then we should not take its ontological commitments too seriously.

Catherine Herfeld: Some people would argue that standard rational choice theories, such as expected utility theory and its variants, are very poorly confirmed and that this is because they refer to unobservables – such as utility or probability judgements – that do not exist. Especially if you defend realism in economics, why should we nevertheless use such theories and defend them as our best theories?

Christian List: I think that some reference to mental states will be indispensable if we wish to explain human behavior. Even if our current best theories are too simplistic and we need to find better ones, I still expect that some form of mentalistic explanation will be unavoidable. Future theories of choice will still attribute certain mental states and processes to the agents in order to explain their observable behavior.

Catherine Herfeld: Some economists and decision theories have argued that standard rational choice theories only make as-if statements about human behavior. Some revealed preference theorists would even argue that they only look at behavior and do not make any statements about mental states of individuals whatsoever but rather – in Paul Samuelson's case – to get rid of metaphysical statements about mental states altogether and thereby defend something along the lines of the view that you label Behaviorism. Why could we not just go with the as-if-interpretation?

Christian List: It is true that the most radical revealed preference theorists don't think that an agent's choice behavior genuinely 'reveals' this agent's preferences or mental states, but they rather take preferences as nothing but patterns of choices or choice dispositions. On that view, attributions of preferences are nothing but a fancy mathematical way of rewriting the informational content of the agent's choice function. To say that someone prefers x to y on that view is simply to say that this agent is disposed to choose x over y under appropriate conditions.

There are a number of problems with this approach. The first is that it is no longer clear how that theory could really offer us any explanation of anything. If preferences are just representations of choices, then preferences can no longer explain those choices because preferences are nothing over and above those choices. In a good explanation, the 'explanans' – the thing that does the explaining – should not be the same as the 'explanandum' – the thing that is to be explained. That principle seems to be violated if preferences are nothing over and above choices.

A further problem is this. If we try to define preferences in terms of choices, we run into the well-known difficulties that behavioral economists have identified, namely that peoples' choices often violate the standard conditions for representatility in terms of a preference orderings. The weak axiom of revealed preference or Richter's axiom of revelation coherence are standard conditions under which a choice function is representable in terms of a certain kind of preference relation. If people violate those conditions, then we cannot simply represent their choice functions in term of preference relations, at least not without significant theoretical complications.

In sum, there are at least two difficulties with a classical revealed-preference approach. First, representations of choice behavior in terms of preference relations do not always exist, because peoples' choices often violate some of the relevant conditions. Second, even when a representation

of choices in terms of a preference relation exists, it is not really explanatory if we treat those preferences as mere representations of those choices rather than as genuine mental states.

To give you an analogy, if a magnet attracts another object due to the relevant magnetic forces, we want magnetism to be the explanation for the attraction between those objects. Similarly, if two objects attract each other due to gravity, we want gravity to offer an explanation. We don't want to say that gravity *is* just the fact that they attract each other. Similarly, we don't want to say that my preference for apples over oranges is nothing but my tendency to choose one over the other.

Catherine Herfeld: One of the things that revealed preference theory in its radical version might allow us to say is that we do not really want to offer explanations of choices in economics. Rather, we want to make predictions. To do that, we only need behavioral evidence to justify certain behavioral principles that allow us to describe how agents behave. A description of behavior is sufficient because what we want to predict in economics are macro-level patterns. This is what economists are interested in. Gathering evidence of mental states is the task of psychologists. How would you respond this argument?

Christian List: To address this argument, let me draw a distinction between two different theses that a behaviorist might hold. The first is a thesis about the admissible evidence for economic theories. The second is a thesis about the ontological commitments of economic theories.

The thesis about evidence is that the admissible evidence in economics is restricted to choice behavior alone. According to this thesis, when we test economic theories, we should not take into account any evidence about mental states, such as self-reports about mental states or evidence that comes from neuroscience. We should only look at choice behavior. That's a kind of stipulative restriction of the evidence base in economics. The thesis about ontological commitments, by contrast, asserts that the ontological commitments of our economic theories should not include any mental states. If references to mental states feature in economic theory at all, these should be seen as nothing more than instrumentally useful constructs.

It seems to me that the first of these behavioristic theses, the one about evidence, is not completely unreasonable. One might argue that there is a division of labor between economics and psychology, and that psychologists should look at different kinds of evidence than economists. Perhaps one might say that in economics we want to look primarily at choice-behavioral evidence. That's not my own view, but it is a perfectly reasonable view, and no doubt Faruk Gul and Wolfgang Pesendorfer would hold some version of it. However, what I would argue is that even if our evidence in economics is restricted to choice behavior, then this does not imply that our ontological commitments should also be so restricted. It might be that, despite the stipulative restriction of our evidence to choice behavior alone, we still have good reasons to postulate certain mental states – genuine preferences and beliefs – if we wish to explain that evidence. That's what Franz Dietrich and I argue.

A good analogy comes from archeology. Think about what archeologists do. They investigate evidence from excavations. To put it bluntly, archeologists look at old pots and pans and various broken items that they dig out of the ground. Let us assume that the evidence base consists of everything they see, in particular old artifacts. But, of course, the ontological commitments of their theories go well beyond that evidence base. Archeologists are not just interested in describing pots and pans and other old items. They are interested in learning something about the cultures and civilizations in question. They want to know how those people in ancient societies lived, what their

cultures were, what their daily routines looked like, what their political organization was, what their religious rituals were, and so on. The archeologists' evidence base is obviously very limited. But they use that evidence base to test hypotheses about those ancient societies. In the end, they come up with some theories that best fit the observed evidence. Those theories might commit archeologists to various claims about those ancient societies, claims that go well beyond the observable evidence itself.

I think something similar happens in economics. We try to make sense of how agents behave in the marketplace and in various other settings. And it might happen that, in order to make sense of their observed behavior, we need to attribute certain beliefs and preferences, norms and convictions, and other intangible mental states to the agents. Without these mental attributions, we would not be able to make sense of the observed behaviour. It would then be in line with best scientific practice to adopt a realist view about those ontological commitments.

Catherine Herfeld: Let's grant the argument that, if we had a limited evidence base about people's mental states in economics, then what we do with applying economic decision theories is to formulate hypotheses that we want to test and corroborate by using data of observed choices. Because those theories turn out to be more or less corroborated by our choice data, our theory together with the theoretical entities it postulates can be taken at face value. However, there might be at least two ways in which this argument fails in economics: First, our decision theories are often not corroborated by the choice data. Second, the explanandum in economics might just not be choice, which is why the relevant data that corroborate our theories is data about stable relationships on the macrolevel, such as for instance demand-price relationships. However, if economic theories are used to formulate hypotheses about group-level behavior, and are often corroborated by average data of demand behavior, then data about individual choice is not something we need for corroboration. Maybe we could even remain agnostic about what that theory ontologically might commit us to on the psychological level.

Christian List: I agree that the explanandum in economics is often different from that in psychology. And I am certainly not suggesting that mental states are part of the explanandum in economics. Rather they are part of the explanans. We refer to those mental states to explain the observed behaviour, such as people's consumer choices or their strategic interactions.

Catherine Herfeld: Do you take the explanandum in economics to be individual choice?

Christian List: One of the explananda in economics is certainly individual choice. Other explananda may be certain collective patterns, or aggregate variables, like inflation and unemployment. We want to explain market transactions, for example, both at the individual level and at the aggregate one. Or we want to explain why all these voters vote for Donald Trump rather than Hillary Clinton. Those kinds of choice behaviors are often our explanandum in economics. We need not primarily be interested in the relevant agents' mental states. Nevertheless, it may turn out that the best explanation of those choice patterns is one that attributes certain mental states to the agents in question. And then I would argue that we should be realists about those mental states that feature in our best explanation.

In physics, I don't think that something like the Higgs boson, for instance, was ever meant to be the explanandum. It was actually postulated as a missing ingredient in the standard model of particle physics in order to make sense of certain things that the model was supposed to explain. It so happened that without the Higgs boson we wouldn't have an equally good explanation of certain

physical phenomena. Our best theory therefore led us to postulate this particular particle and subsequently, indeed, its existence was empirically supported.

Catherine Herfeld: Again, some economists might argue that what they want to explain in economics is macro-level social interaction. To do that, we could start with a description of behavior only, we then generalize over a large group of people and because individual deviations cancel each other out on average, our choice theories are a good predictor of group behavior while not requiring that they accurately explain individual behavior. Would you reject methodological behaviorism, i.e., the view that what we should represent in economic theories are not mental states but behavior, also in this case as a tangible approach? Or would you defend the position that we also have to go for mentalism in that case?

Christian List: To address this issue, we have to distinguish between microeconomics and macroeconomics. There are different ways in which we might draw that line. Roughly speaking, microeconomics gives explanations of certain behavioral phenomena and uses tools from decision theory or game theory to do so. Macroeconomics looks at the relationship between certain macroeconomic variables. There are some economists who think that we should always provide microeconomic foundations for macroeconomic theories and construct individual-level models to underpin those macroeconomic phenomena. But there are also other macroeconomists who take a more holistic view and think that macroeconomic explanations can be freestanding and there is no need for microfoundations.

I would say that the arguments for mentalism apply to microeconomic explanations in so far as those are intentional explanations and not just causal-mechanistic ones. It is crucial that the economic agents whose behavior we are explaining are viewed as intentional agents. We are taking what Daniel Dennett calls an 'intentional stance' towards them. By contrast, if you look at some of the phenomena that we study in macroeconomics – the Philips-Curve-relationship, for example – we are concerned with relationships between certain macro variables. Here, we might actually investigate those relationships as probabilistic causal relationships, where we don't necessarily engage in intentional explanation.

So, of course, if we are not engaged in intentional explanation, then the whole debate between mentalism and behaviorism does not even arise. The debate between mentalism and behaviorism really concerns the status of explanations that are of an intentional sort – that is, explanations of the behavior of intentional agents.

Catherine Herfeld: What would you make of the argument that many of the individual actors that populate microeconomic models are in fact not human beings that have a psyche? Those actors are businesses, corporations, governments, and banks, etc., which do not have a psyche, also in the sense that they do not have a brain that is the physical instantiation of the mind.

Christian List: One way in which we can draw the distinction between micro- and macroeconomics is precisely in terms of whether we are focusing on the behavior of intentional agents or whether we are focusing on the relationship between certain macro-variables. If we look at the theory of the firm, which is a core area within microeconomics, then we see that this theory actually models firms or corporations as intentional agents in their own right. In fact, the traditional theory of the firm models the firm as a utility-maximizing agent much like in expected-utility models of individual rational agency. In this sense, firms are sometimes modeled as rational individuals. In fact, such corporate actors might be even more rational in the orthodox sense than

human individuals. Perhaps the true *homo economicus* is not the human individual at all. Rather, it is something like a commercial corporation.

Some people are inclined to be instrumentalists with respect to the attribution of beliefs or preferences to firms or group agents. They might say that for some explanatory purposes, it is useful to employ the tools of game theory or decision theory to make sense of how British Petroleum behaves. But, they claim, British Petroleum does not truly have beliefs or preferences. The ascription of preferences and beliefs to British Petroleum is just a metaphor or an instrumentally useful construct. People who hold this view would be inclined to reject mentalism when it comes to group agents. Yet, I want to be consistent and embrace mentalism even in the case of group agents. This is basically the view Philip Pettit and I have defended in our book, *Group Agency* (List/Pettit 2011). We have argued that organized collectives can truly be intentional agents, with 'minds of their own', as Philip Pettit would say. If a collective entity is best explained by attributing intentional agency to it, then so be it. Again, the naturalistic ontological attitude would require that we should take the ontological commitments of our best scientific theories at face value. If our best social scientific theories treat certain collectives as agents, then we should accept the view that those collectives truly have beliefs and preferences.

I suspect that part of the reason why some people are reluctant to attribute beliefs, preferences, and other mental states to group agents is that they have a very strong intuition that group agents have no such thing as consciousness, while individual human beings do. If consciousness, especially phenomenal consciousness, were the mark of preferences, beliefs, and other mental states, then it would indeed be hard to defend the view that group agents can have those mental states. Most of us think that there is no group consciousness. I have given an argument in support of this commonsense view in my paper, "What is it like to be a group agent?" (List 2018). But it is important to distinguish between a thin notion of beliefs, preferences, and other mental states in a functionalist sense, which we can unproblematically attribute to group agents, and phenomenal consciousness in a richer, non-functionalist sense, which group agents do not have.

Catherine Herfeld: In defending mentalism, you seem to endorse the distinction between the mind and the brain. In defending the position that groups act intentionally and thus have preferences, beliefs, and other mental states, you need a concept of a group mind that has to be instantiated by something equivalent to the human brain. Do you also endorse the position that groups have brains or what does instantiate a group mind?

Christian List: I would indeed distinguish between the mind and the brain, and I think the distinction is conceptually very important. It is of course true that human beings and other biological animals have brains. In the human case, as in the case of a chimpanzee, our cognitive capacities are very much tied to our brains as the underlining hardware. As a conceptual matter, however, the brain and the mind need to be carefully distinguished. You can think of the brain as the biological hardware that implements the mind, but conceptually speaking, you could imagine that a mind may also be implemented by a different kind of hardware. In recent years, there have been rapid advances in artificial intelligence and computer science. Nobody knows how long it will take for computer scientists to come up with genuinely intelligent autonomous systems, whose cognitive capacities are comparable to those of biological animals. We are nowhere near this point right now, but some people think that sophisticated artificially intelligent systems could be developed during our lifetimes. Different people give you different estimates of how long it will take. Future sophisticated AI systems – for example, sophisticated robots with strong cognitive capacities and a

rich behavioral repertoire – will obviously have a very different hardware than biological organisms. Nonetheless it may well turn out that the best way to make sense of such systems is to attribute mental states to them. In that sense, an AI system might have a mind. It has beliefs about its environment, it has preferences about what it wants to achieve, and it displays the kind of goal-seeking behavior that we associate with intentional agency. But the mind of that system is computationally implemented in a very different hardware than a human brain. Once we recognize that a mind can be implemented – at least in principle – in different hardware systems, then we can also recognize that a group agent can in principle have a mind, which is implemented in a different hardware than the usual biological one. The hardware system in the case of a group agent is the particular organizational structure of the collective in question. It is social, rather than biological.

Of course, at this point, the hardware system that implements the human mind, namely the human brain, is the most sophisticated system – at least of a certain limited size – in the known universe. We don't have any other hardware system that is even remotely as sophisticated as the human brain. Compared to this, the hardware system of a group agent or the hardware of the kinds of robots that we have is very simplistic. Nonetheless, those systems can still implement rudimentary minds, which then give rise to certain forms of rational, goal-seeking behavior.

Catherine Herfeld: You accept that utility and probability functions and a binary relation with a set of constraints (the rationality axioms) imposed on it is an idealized way to think about mental states. Some people would, however, argue that preferences for example do not in fact formally represent mental states such as desires as postulated in a folk psychological belief-desire framework and that therefore formal theories of rational choice in economics are not equivalent with the folk-psychological approach to behavior in terms of beliefs and preferences. They are much thinner. While you think of your account as being psychologically more realistic than the traditional account, you draw upon similar concepts and thereby might confront the same challenge. How would you respond to such criticism?

Christian List: My preferred version of decision theory is not orthodox rational choice theory. My preferred theory – or I should better say 'framework' – is the one Franz Dietrich and I are currently developing. We call it a theory of 'reason-based choice' (Dietrich/List 2016b). At this stage, it's best viewed as an explanatory framework whose details still need to be developed further. The aim of this framework is to explain an agent's choices not merely by attributing a simple preference ordering to that agent, but by attributing to the agent a richer mental construct, namely what we call a 'reasons structure'. We retain the rational-choice-theoretic idea of explaining an agent's choice behavior by means of a certain mental state attribution. We just provide a richer account of what the relevant mental states are.

As we see it, the relevant mental states are not just preferences (and beliefs), but reasons structures. Classical rational choice theory can be viewed as a special case of our framework. And I think of this special case as an idealized case. Some features of that idealized special case must be relaxed or amended in order to explain choices more generally. In that sense, it is true that we give up some of the orthodox rationality assumptions, but we certainly do not give up the idea that there are mental states that explain an agent's choices.

Catherine Herfeld: Do you consider your reason-based rational choice theory as a step towards de-idealizing traditional rational choice theory?

Christian List: Yes, the aim is to arrive at a more psychologically realistic way of explaining

choices, while retaining as much of the elegance and parsimony of classical rational choice theory as possible.

Catherine Herfeld: In your theory, you attempt to develop an account of choice that draws upon the work in philosophy of action. One of the theory's virtues is that it generalizes standard rational choice theory and thereby shows that the standard theory is at best incomplete. What are the advantages of your theory over traditional classical approaches such as expected utility theory and what do you take the economist to get out of your theory?

Christian List: To explain the advantages of our theory, let's first remind ourselves of how traditional revealed preference theory works. For illustrative purposes, let's set aside uncertainty or incomplete information, and let's just focus on a very simple case, namely an agent's choice between several options, where the agent has complete information about what those options and their consequences are.

An agent's choice behavior in this case can formally be expressed by a choice function. The choice function assigns to each choice context the set of those options that the agent would choose in that context. If the set of chosen options in a given context is singleton, then there is a unique option that the agent would choose in that context. If the set of chosen options is non-singleton, then the interpretation is that these separate options are tied for choice and maybe the agent needs to randomize or employ some tie-breaking criterion to pick one of those options. So, in short, a choice function specifies which option, or options, the agent would choose in each context.

Now, traditional revealed preference theory seeks to explain the agent's choice function by ascribing to the agent a preference ordering – or more generally a preference relation which need not always be a complete ordering – and then hypothesizing that the agent will choose a most preferred option in each context. This explanation is possible if, and only if, the agent's choice function satisfies certain classical rationality conditions. We can then indeed find a preference ordering that will represent that choice function. But we are likely to encounter the two difficulties that I have already mentioned. First, there are some realistic choice functions that cannot be explained in terms of a preference ordering over the options, because they violate the relevant rationality conditions. Second, even when there is a preference ordering over the options that rationalizes a given choice function, this is not necessarily a genuine explanation of the agent's choices.

Let me give you an example of a choice function that cannot be naturally explained in terms of a preference ordering over the options. This example was given by Amartya Sen. He asked us to imagine a polite dinner party guest who is offered a choice among several pieces of cake. Politeness requires the guest not to choose the largest piece of cake, because that would be greedy. Now, in one situation, the guest is offered a choice between a big, a medium, and a small piece of cake and chooses the medium piece because it is the biggest of the three that she can politely choose. In another choice situation, the big piece is already gone. The only available pieces of cake are the medium and the small ones. Now, applying the norms of politeness again, our guest chooses the small piece and no longer the medium one. This pattern of choice – choosing the medium piece over the big and the small ones in one context and choosing the small piece over the medium one in another – violates some of the classical rationality conditions, in particular contraction consistency. The option that was chosen from the bigger menu – namely the medium piece of cake – is no longer chosen from the smaller menu even though that option is still available. Indeed, we cannot rationalize that choice function in terms of a preference ordering over pieces of cake. Of course, we

could try to ascribe to the agent a more complicated preference ordering, perhaps a preference ordering over pieces-of-cake-in-a-particular-situation. But that can easily end up becoming tautological. We would have to re-individuate the options in a very fine-grained way, so that the medium piece of cake in the first situation no longer counts as the same option as the medium piece in the second. In the limit, options might be tied to one and only one choice context, so that the agent can never encounter the same option more than once.

What Franz Dietrich and I suggest is that we should not view options as uninterpreted primitives. Instead, we should recognize that agents perceive the options through the lens of certain motivationally salient properties. An agent considers each option. The option potentially has a vast number of properties. Most of those properties are irrelevant for the agent, but some properties are motivationally salient. We suggest that when an agent makes a choice in a particular context, he or she looks at all the feasible options through the lens of those properties that are motivationally salient in that context. He or she then chooses an option which he or she thinks offers the most choice-worthy bundle of salient properties.

Formally speaking, such a reason-based explanation has two components. First, we introduce what we call a 'motivational salience function', which is a function that assigns to each choice context the set of those properties that are motivationally salient for the agent in that context. Those are the properties that the agent pays attention to in that context. Second, we introduce what we call a 'fundamental preference relation', which we sometimes also call a 'weighing relation'. This is a binary relation over property bundles – combinations of properties. It ranks property bundles in terms of how choice-worthy they are for the agent. In each context, the agent now looks at the options through the lens of the motivationally salient properties in that context. Specifically, for each option, the agent considers that option's set of motivationally salient properties in the given context. In effect, the agent perceives each option as the bundle of motivationally salient properties offered by it. Options are then ranked on the basis of how the agent's fundamental preference relation ranks the relevant property bundles. The agent chooses an option whose bundle of motivationally salient properties is ranked most highly.

Catherine Herfeld: How do you go about explaining human choice with your account?

Christian List: According to the reason-based account, we explain an agent's choice behavior, not by attributing to the agent simply a preference ordering over the options, but by attributing a 'reasons structure', which is our term for the pair consisting of a motivational salience function and a fundamental preference relation (or weighing relation), as just explained. The goal is to find a reasons structure that entails the agent's choice function.

Let me give a simple example. Suppose I am a consumer and I go to the supermarket to do my shopping. I see all these different consumer goods there, for example, different yogurts. Each yogurt may have lots and lots of different properties. It might be a cherry yogurt and it has some sugar content, and a certain fat content, the label is blue, the brand is such and such, the price is such and such, the best-before date on the label is the 5th of April, and so on. You can easily see that there may be thousands of properties that this particular yogurt has. As a consumer, I will initially focus on just a few of those properties. Maybe it's just a very small number. Maybe I am in a rush and all I care about is getting a cherry yogurt that is fat-free. Then 'cherry flavored' and 'fat free' are the only motivationally salient properties for me. Alternatively, I might be in my environmentally friendly mode and also care about whether the yogurt was sustainably produced, or whether it is free from chemical additives. This example illustrates that in different situations – that is, in

different choice contexts – I might focus on different properties.

For a reasons structure to explain my choices, it would have to specify which properties I care about in each context, and it would have to specify how choice-worthy I find different bundles of properties relative to one another. If it specifies both of these things in such a way as to accommodate my choices in all the different contexts, then the reasons structure is choice-behaviourally adequate. It entails, and thereby explains, the choices that I make.

An important feature of our framework is that it captures different ways in which an agent's choice behaviour can be context-dependent. Specifically, there are two dimensions of context-dependence. First, my motivational salience function may specify different properties as motivationally salient in different contexts. If I am in a rush, I may care about fewer properties than if I am in my reflective mode, or in my environmentally friendly mode, where these modes are triggered by different contexts. We call that phenomenon 'context-variant motivation'. To explain the second dimension of context-dependence, note that, so far, the properties of the yogurts I have mentioned are all examples of what we call 'intrinsic properties' or 'option properties'. If a yogurt is fat-free, for example, then it has that property independently of the context in which you consider it. A fat-free yogurt remains fat-free no matter whether you are looking at it in the supermarket, in your fridge at home, or in the breakfast room in the hotel. If it has that property in one context, then it has that property in all the other contexts too. But there can also be what we call 'relational properties', such as the property of being the cheapest yogurt. That clearly depends on the range of available yogurts. In one context, a particular cherry jogurt may be the cheapest, while in another it is not. Or, having the lowest sugar content among the available yogurts is a relational property. Another example is the property of being the yogurt which is placed in the middle of the shelf. All of these are properties that are related to the context. An option doesn't have them by itself, but only in relation to the context in which it is available.

Once we recognize that the motivationally salient properties an agent cares about may include not only intrinsic properties but also context-related properties, we can see that there is a second dimension of context-dependence, namely focusing on relational properties. We call that phenomenon 'context-related motivation'.

What I have just explained is our taxonomy of two different forms of context-dependent choice. One form of context-depedence is context-variance, which means that an agent cares about different properties in different contexts. Another form of context-dependence is context-relatedness, which means that the agent cares not only about intrinsic properties of the options, but also about certain context-related properties.

Catherine Herfeld: How exactly is your account different from revealed preference theory?

Christian List: First of all, classical revealed preference theory does not focus on properties or on reasons for choices; it just focuses on preferences over options, which are largely fixed. There are some exceptions, such as the characteristics-based approach of economists such as Lancaster and Gorman. Secondly, classical revealed preference theory does not recognize the forms of context-dependent choice that our framework can capture. However, we can view classical revealed preference theory as a special case of our framework, namely the special case in which both forms of context-dependence are absent. The options are always viewed through the lens of the same properties irrespective of the context we are in. For instance, an option that is perceived as a fat-free cherry yogurt in one context is also perceived as a fat-free cherry yogurt in another. In classical

revealed preference theory, options are not perceived differently depending on the context. In this way, classical revealed preference theory assumes away context-variant motivation, and it also assumes away context-related motivation. We can therefore view it as a limiting case of our framework. However, our approach has the advantage of taking the context of choice into account. We know from the well-established work on framing effects that different contexts may make different properties, features, or characteristics of the options salient for decision makers. We think this supports the claim that, in the real world, people's choices are often context-dependent in one of the ways our framework captures.

Similarly, if we find Amartya Sen's example of the polite dinner party guest plausible, or if we accept the claim that social or moral norms matter in many situations, then we have good reasons to think that context-related motivation and caring about relational properties are perfectly real phenomena. If this is so, then I would suggest that decision theory would be better off recognizing those phenomena.

Catherine Herfeld: How would you go about testing your reason-based rational choice theory?

Christian List: If we want to test a reason-based explanation of an agent's choice behavior, one possibility is simply to aim at choice-behavioral adequacy. Choice-behavioural adequacy means that the reasons structure that we attribute to the agent implies the empirically observed choice function. This test is really no different than the test for any classical-rational-choice theoretic explanation. Choice behavior is the ultimate criterion of adequacy of a reason-based explanation in this case.

A second possibility is to test a reason-based explanation by making predictions. Our approach not only allows us to ask whether a reasons structure correctly entails the agent's observed choices up to now, but it also allows us to make predictions of future choices: choices that the agent would make in new contexts for which we do not yet have any observations. Such predictions are possible because reasons structures can be naturally extended from smaller domains of previously observed choice contexts to larger domains of additional choice contexts. We can then ask what choice behavior the attributed reason structure would lead us to predict when extended to some larger domain of contexts. That gives us a natural way of performing an enhanced choice-behavioral test for a reason-based explanation.

In general, an empirical researcher using our approach would go through the following steps. First, observe a given agent's choice behavior. Then, come up with a hypothesis as to what the agent's reasons structure might be, and make sure that this hypothesis correctly explains the observed choice behavior. Next, extend this reasons structure to a larger domain of previously unobserved choice contexts and make predictions as to how the agent would choose in those new contexts. Finally, try to get some evidence through observations or experiments in order to see whether those predictions are correct or not.

Note that, so far, I have only spoken about choice-behavioural tests. Psychologists might be able to look at evidence other than choice behavior, and use this to test certain hypotheses about people's motivating reasons. To give just one example, psychologists could ask people to give verbal reports about their reasons. Of course, verbal reports need not always be reliable, but if we gather enough psychological data, then we may be able to identify some reasons structures as more psychologically adequate than others. In addition, there might be some psychological evidence about the level of cognitive complexity that we can realistically ascribe to an agent. Perhaps there is evidence that an agent is not simultaneously able to focus on more than a certain number of

properties. Maybe focusing on more than three or four properties at once is too hard. If so, then this would allow us to reject any reasons structures that deem too many properties motivationally salient. In this way, there could be psychological ways of testing reason-based explanations that go beyond choice-behavioural tests.

Catherine Herfeld: Why is this a useful approach to rational choice?

Christian List: Among other things, as I have already noted, it can accommodate the phenomenon of context-dependent choice. The properties that are motivationally salient may vary from context to context, which is in line with what empirical evidence from the framing literature suggests. Furthermore, among the motivationally salient properties, there can be relational properties. Those are properties which options do not have intrinsically, but only in relation to the choice context, such as the property of not being the largest piece of cake. What Franz and I are suggesting is that people who follow social norms often care about relational properties. A particular option, such as the medium piece of cake, may be polite in one context, but impolite in another. By recognizing that relational properties may be motivationally salient, we can make sense of the resulting norm-following behaviour. Traditional rational choice theory cannot easily accommodate this.

Catherine Herfeld: Your account seems to be empirically quite demanding. People might have different motivational salience functions and weighing relations. In order to explain or predict their choices, we would be required to have information not only about their psychological makeup but also about the context within which this choice is or will be taken. As such, does your account confront the same difficulties as traditional rational choice theory, which was challenged because it either lacks empirical support for the utility function or – if such support is not given – suffers from poor empirical power?

Christian List: Of course, we do not – strictly speaking – know what someone's motivational salience function is, just as we do not antecendely know what someone's preferences are or what their beliefs are. Rather, when we try to explain an agent's choice behavior, we make certain hypotheses about this agent's mental states. In the traditional theory, we make hypotheses about what this agent's beliefs or preferences are. If we are lucky and these hypotheses entail the right behavioral predictions, then we have some tentative support for them. Let us say we attribute to a firm a utility function that is strictly increasing and concave in profits. Then we look at whether this attribution of utilities to the firm allows us to explain its behavior in a satisfactory way. If it does, we have some support for this attribution. That is basically how classical rational choice theory works. In the reason-based framework, similarly, if we want to explain an agent's choices, we need to formulate some hypotheses. We need to come up with some hypotheses about what this agent's reasons structure is. Of course, we look for parsimonious hypotheses. We do not want to make complicated attributions. And so we must ask: what is the simplest hypothesis about the agent's reasons structure that would allow us to explain and predict this agent's choice behavior? Once we have identified the simplest hypothesis that does this job, then we have good reasons to accept this hypothesis, at least tentatively.

Catherine Herfeld: However, it appears that your account is more demanding than the classical framework in the sense that you need to have to have more empirical knowledge about the specific context in order to come up with certain hypotheses.

Christian List: I agree with you that if we want to analyze the role played by the context in detail, then we may need more information about those contexts than merely the set of feasible options in

each context. In that sense, reason-based explanations sometimes require a richer evidence base than classical rational choice theory. However, we can still use our framework without going beyond the evidence base of the classical theory. Take Sen's example of the polite dinner party guest again. Suppose we are given only the agent's choice function and no other empirical data. We know that the agent chooses the medium piece of cake from the option set 'big, medium, and small', and he or she chooses the small piece of cake from the option set 'medium and small'. And perhaps we also know the agent's choices from some other option sets. That's the choice function, and – let's suppose – we have no other empirical data. To explain the given data, the classical theory would then try to come up with a binary relation over the options – a preference relation – such that by attributing it to the agent we can rationalize the agent's choice function. In Sen's example, this would not work, unless we re-describe the options in some complicated way. If we use the reason-based framework, by contrast, we can explain the data, namely by invoking the idea that the politeness property is motivationally salient for the agent and that the agent prefers the property combination 'small and polite' over the property combination 'medium and not polite'.

Of course, the motivational salience of the politeness property is not something we can observe directly. Rather, it is part of the explanans that we introduce in order to make sense of the observed data. That the agent cares about politeness is a hypothesis we make. But if this hypothesis gives us a good explanation of the observed choice behaviour, then this would at least tentatively corroborate the hypothesis. So, in principle, we do not need to look at richer evidence than classical rational choice theory does. That said, richer evidence is often helpful, because if we look at additional psychological evidence, we may be able to come up with better explanations. But even if you have only choice-behavioral evidence, you can still try to give reason-based explanations of the observed choices.

Catherine Herfeld: Another objection that might be raised against your account is that it is tautological in the same way as classical theories can be. An infinite number of things can potentially be motivationally salient for an agent in a certain situation and those can also hughly differ across agents. However, if you do not specify empirically the motivational salience function for an agent – for example, that politeness is motivationally salient in the case of Sen's polite dinner guest – but just postulate what is motivationally salient, it appears that the concept of motivational salience does not offer more substantial explanations of this choice pattern. How would you cope with this challenge if not by offering substantially more empirical support for your hypothesis than classical rational choice theorists do?

Christian List: If you impose absolutely no restrictions on which properties you may postulate as potentially motivationally salient, then you can in principle 'explain' everything in our framework. That is true. But likewise, in classical rational choice theory, if you impose no restrictions on how you may individuate the options, the classical theory becomes tautological and you can explain everything. In the classical theory, if you are prepared to re-describe the options such that each option can occur in one and only one context – maybe you put a particular date-and-time stamp stamp on each option – then trivially any pattern of choices can be viewed as maximizing preferences over those very finely individuated options. This would make the classical theory tautological too.

No theory, whether we take classical rational choice theory or our reason-based theory, can avoid the tautology-charge without certain auxiliary assumptions or auxiliary hypotheses. The classical theory becomes non-tautological only relative to certain auxiliary assumptions about what the

options are and how they should be described or individuated. Likewise, our theory becomes non-tautological only if we are prepared to make certain auxiliary psychological assumptions about which properties may plausibly become motivationally salient and which properties may not. That is exactly how it is when we engage in intentional explanation. We have to rely on certain auxiliary hypotheses.

When Donald Davidson discussed the methodology of intentional explanation, he famously proposed the principle of charity, which is a principle that is meant to guide our choice of psychological hypotheses in the interpretation of people's behavior. He recognized that the empirical evidence about someone's behavior often underdetermines the explanation. More than one explanation may be compatible with the same evidence. We need to introduce some further constraints in order to choose one explanation over others. Davidson's principle of charity is an attempt to guide us in that choice of explanation. It tells us to go for explanations that render people as rational and reasonable as possible. We should accept explanations that attribute irrationality or unreasonableness to people only if we have exhausted all the more charitable alternatives.

Likewise, we need to rely on certain auxiliary hypotheses and invoke certain methodological principles of charity in order to come up with good reason-based explanations of an agent's choice behavior.

Catherine Herfeld: The audience for your approach consists of social scientists, among others. Could you give an example of an economics problem that would benefit from your reason-based approach to rational choice?

Christian List: The intended audience is an interdisciplinary one. This includes not only economists, but also philosophers, psychologists, cognitive scientists, and computer scientists working on AI. Our approach is applicable quite broadly. I have already given the examples of framing effects, on the one hand, and norm-following behaviours, on the other, where decision makers focus on properties that options have in relation to the choice context. The kind of context-related choice illustrated by the example of the polite dinner-party guest is involved in many forms of norm compliance and ethical behaviour. Whether a particular option is norm-conforming or morally permissible often depends on the context in which you choose this option. Some patterns of behavior are acceptable in some contexts, but not acceptable in others. Insofar as social and moral norms play an important role in many situations, I expect our framework to be applicable quite broadly.

Also, think about problems in political economy, the area of political science in which we use methods from microeconomics to study the behavior of politicians or voters. Having a realistic account of which properties of the options those agents care about and how those properties might depend on the context seems quite central for many applications.

Catherine Herfeld: Some economists would argue that the classical expected utility approach, while being empirically inadequate at the level of individual behavior, approximates economic behavior reasonably well, especially on the group level. Given that it also satisfies additional epistemic virtues, such as consistency with other exisiting approaches in economics, simplicity, and predictive power, we should just keep it until we have a better alternative. In which way is your account a useful alternative for economics, beyond improving standard decision theory?

Christian List: As I have argued, our approach can explain non-classical choice behaviours that are not adequately accommodated by classical rational choice theory. Why does this matter? The *homo*

economicus model might be a reasonably good account of the behavior of certain kinds of market participants, but it is extremely simplified and idealized. Maybe it captures the behavior of firms or commercial corporations better than it captures the behavior of individual human beings. But even if the *homo economicus* model still has a role to play in economic explanations, we must recognize that it is rather idealized, even from the perspective of core areas such as the theory of the firm. How closely firms approximate that model is – at the end of the day – an empirical question. We have good reasons to expect that even firms sometimes display deviations from classical rationality, because firms, like individuals, will be subject to informational and computational limitations and will occasionally also be subject to certain forms of reasoning failure. Therefore, some of the rationality failures that are well documented in the case of individuals can potentially occur at the level of firms too.

Furthermore, there is evidence to suggest that even highly sophisticated market participants – such as financial traders – sometimes deviate from the kind of behavior that we associate with the *homo economicus* model. In recent years, there has been a growing interest in the subject of behavioral finance, where scholars study the behavior of such market participants. Interestingly, even those sophisticated agents sometimes deviate from how a traditional *homo economicus* would behave, even though you would expect – based on their profession – that they are trained to act in line with the *homo economicus* model.

So, the domain of applicability of a psychologically more realistic version of decision theory, such as the one that Franz and I are trying to develop, is potentially very broad.

Catherine Herfeld: In your work, you seem to defend the position that the psychological level is the adequate level of of explanation in economics. Your argument against going to the neural level to explain behavior is that it might go too far. The explanation would be very and unnecessarily complex and you care about having simpler explanations. You seem to think there being a trade-off between the huge effort for offering a very detailed and complex explanation and the rather low benefit of getting the respective information. You argue that such a complex explanation offers us information that we do not really need in the social sciences. Do you consider neural evidence, which is in large parts what neuroeconomists are engaged in, to be of any use in the social sciences?

Christian List: Neuroscience can complement economic decision theory and psychology in the study of human decision-making. As a complementary research program, there is absolutely nothing wrong with neuroeconomics. It is an interesting and potentially promising development. What I would object to, however, is the attempt to reduce psychological-level explanations – such as those in decision theory and microeconomics – to neuroscientific explanations. Similarly, I object to the idea that we might be able to dispense with microeconomics or decision theory altogether and turn microeconomics into a kind of subdiscipline of neuroscience, where instead of talking about preferences, beliefs, and various other intentional-level notions, we give an account of economic behaviour solely in terms of neurobiological processes. That, I think, is not a very promising approach.

Catherine Herfeld: Could you elaborate further on your argument against reduction?

Christian List: For different explanatory purposes, we tend to adopt different levels of description. There are some phenomena that are best explained at a microphysical level. If we want to understand the behavior of elementary particles, for instance, then clearly the microphysical level – the one associated with particle physics – is the appropriate one. But if we look at more

macroscopic systems, the attempt to understand those systems at the level of the particles of which they are composed is not very promising. Even the attempt to reduce chemistry to physics seems not very fruitful. Philosophers of chemistry have made arguments to the effect that even simple chemical properties such as acidity are not easily type-reducible to corresponding physical properties. Likewise, when we move from chemistry to biology, we often need to adopt a different level of description than the chemical one. In biology, there is the level of the cell and its constituents, on the one hand, and there is the level of the organism, or the level of the ecosystem, on the other. In a field such as behavioral ecology, it would be very difficult to explain animal behavior if we tried to understand each animal as nothing but a complex chemical system consisting of gazillions of cells. What these considerations show is that it is quite standard in the sciences to move to a higher level of description for explaining many phenomena.

For similar reasons, I would argue that even though it is important to understand the biological brain and how it gives rise to its remarkable cognitive capacities, the sheer complexity of the brain is such that by focusing on the neural level alone we would not be able to come up with parsimonious and illuminating explanations of human behaviour.

Let me give you an example from the social domain. Close to my apartment in London, there is a bus stop and there are multiple bus lines that stop at this particular bus stop. There is the number 19 and there is also the number 38. Aside from the difference in the number, those two buses look the same, and sometimes the driver of number 19 might also drive number 38, and vice versa. What's more, the routes of the two buses partly overlap as well. So, initially, when you take one of the buses, it looks like they are on the same route. Only at some point, after several stops, the routes diverge: number 19 takes a different turn than number 38. How do we explain this?

Clearly, there is a regularity here which requires an explanation. It can be observed day after day, in a more-or-less exceptionless manner. Bus number 19 takes one route while bus number 38 takes another. First of all, we must figure out that it has something to do with the relevant driver and what instructions he or she has received. If we were to look at the engine or the steering wheel, we would clearly go wrong and would not be able to pinpoint the difference between the two buses. Suppose we've figured out that the different routes have something to do with the driver. If you then do sophisticated brain scans of the drivers to identify how their brains differ, you would end up with an unbelievably complex explanation, which would involve a lot of extraneous details. But once you adopt an agential or psychological level of description, where the one driver has been instructed to take route 19 and has the appropriate goals and intentions, while the other driver has a different set of goals and intentions, associated with route 38, then that gives us a perfectly satisfactory explanation of the regularity. Similarly, in the case of many choice-behavioral phenomena, the right level of explanation is the psychological one, not the one associated with neuroscience. We must recognize that – in many cases – it is people's mental states – their beliefs and preferences – that best explain their choices, not the details of the hardware implementation within the brain.

Catherine Herfeld: I understand the pragmatic argument here, but a reductionist would probably object that the neurological explanation is ultimately a more fine-grained, a more detailed explanation and thereby offers us real understanding about the phenomenon, which is what we ultimately want.

Christian List: Sometimes more fine-grained explanations can indeed be more illuminating than more coarse-grained ones. Suppose, for example, you want to explain why a person's behavior changes under the influence of alcohol. Then it makes sense to invoke the biology of the human

organism. Biologists are able to explain how alcohol affects brain functions, and this sheds light on how exactly a person's agential capacities are diminished under the influence of alcohol. Here we have a good example of how invoking the neural level can help to explain human behavior. In contrast, there are also cases in which providing further details of the hardware implementation of a cognitive process would be a distraction. It would not add anything useful from an explanatory perspective. In the example of the buses, we can really give a fully satisfactory explanation of why the two buses take different routes simply by pointing to the different intentional states of their drivers. Describing the drivers' brain processes would be redundant and would distract us from the features that matter.

Indeed, suppose you change the drivers, but keep the relevant instructions in place. Our intentional explanation is completely invariant under this change and continues to work just as well. With the new drivers, the two buses will still travel along their designated routes. However, the new drivers' brains will be ever so slightly different. Therefore, many details of the neural story would have to be updated if we wanted to give a full neural-level account of why the new drivers equally take the appropriate bus routes. This shows that the brain is the wrong level of explanation here.

Let me give you another example of why higher-level explanations are sometimes preferable to lower-level ones. Suppose your Microsoft word processor has a bug, and it systematically crashes under certain conditions. Say, whenever you simultaneously switch on the spell check and the tracking of changes, it crashes and it does so systematically. For example, it does this no matter what other software is installed, what the time of the day is, and which documents you are editing. And the issue persists even when you install the same word-processing software on another computer. The right level of explanation for this bug is obviously the level of the software, not the level of the hardware. There is a mistake in the software that gives rise to the crash. If you tried to give a more fine-grained explanation by focusing on the flow of electrons through the underlying computer chips, you might be able to identify some complicated configuration of electron flow that is associated with the crash of the computer. But a different flow of electrons could equally be associated with the same crash under different conditions. And the flow of electrons would most definitely be different if the computer were different. Intuitively speaking, the hardware level is the wrong level of explanation here. It is too fine grained. It would give you too many extraneous details, and these would impede rather than aid our understanding of the phenomenon we are trying to explain.

Analogously, I suggest that, when we seek to explain human decision-making, the psychological level, rather than the neural one, is often the right level of explanation. It captures the relevant phenomena in a way that conveys neither too little nor too much information.

Catherine Herfeld: You mention in your article 'Mentalism vs. Behaviorism' that economists reject the theory of choice by, for example, Gerd Gigerenzer not because it is not grounded upon the utility-maximization principle but rather because economists are unsure whether such theories offer the best explanations of empirical phenomena economists intend to explain. This is interesting because Gigerenzer's work for example heavily relies upon empirical results from psychology etc. and seems to explain a lot of human behavior much better than traditional choice theories, precisely because it dispenses with those demanding assumptions of complete information, unlimited computational power, etc. that traditional rational choice theory rests upon. How, do you think, we can justify that rational choice theories are our best theories if they often do not offer good explanations?

Christian List: Let me clarify a few things. I am certainly not saying that approaches that are based on the assumption of maximization are necessarily better approaches than various non-maximization-based alternatives. In fact, we have perfectly good reasons to believe that in many circumstances people satisfice rather than maximize when they make decisions, or that they use certain heuristics, as Gigerenzer shows. So, maximization is definitely not a universally confirmed pattern. I also think that Gigerenzer's work is extremely interesting, and he has accumulated some compelling evidence for the use of simple heuristics by human decision-makers in a number of realistic choice situations. I find all this not only interesting but also relevant to our understanding of how people make decisions.

What I would argue, though, is that Gigerenzer has not yet given us a fully satisfactory *theory* of decision making. He gives us a number of important data points that a good theory would have to accommodate, but it's not clear to me that he has fully systematized this into a single unified theory. Gigerenzer provides good evidence that people employ simple heuristics in certain decision situations. When a doctor has to make a quick decision in a triage situation and has to figure out whether a patient is likely to have a heart attack, he or she might focus on a certain relatively short list of criteria and then go through these criteria quickly in some hierarchical order to distinguish between high-risk and low-risk patients. This might be a perfectly good heuristic for a quick decision – and one that strikes a decent balance between speed and accuracy.

Equally, there is evidence that people sometimes employ the recognition heuristic. When asked which two of two cities is bigger, for instance San Diego or San Antonio, people tend to choose the city whose name they recognize, rather than the city whose name they don't recognize. And that may also be a perfectly reasonable choice strategy. But the discovery of these – and other – examples does not yet amount to a general theory of decision making. Gigenzer gives us a great list of examples of heuristics that real people use when they make decisions. But ideally, we would like to subsume all these examples under the umbrella of a general theory that identifies the underlying mechanisms.

I would argue that the heuristics-based forms of decision making that Gigerenzer has identified can be represented in the reason-based framework that Franz Dietrich and I propose. The different criteria on which a user of Gigerenzer's heuristics focuses can be viewed as the motivationally salient properties of the options. The way in which the agent compares the options on the basis of those criteria can then be encoded in the fundamental preference relation or weighing relation. Suppose a doctor uses a heuristic involving three criteria to sort patients into potential heart attack cases and less urgent cases. She first looks at criterion one, and if that criterion is met, she puts the patient into the high-risk category, whereas if it's not met, she moves on to criterion two. Again, she looks at whether that criterion is met. If it is, she puts the patient into the high-risk category, and if not, she moves to criterion three. If none of the criteria is met, the patient is classified as low-risk. That can easily be represented in terms of a reasons structure in which the motivational salience function picks out three properties as motivationally salient, namely those corresponding to the three criteria. The fundamental preference relation is then a lexicographic one. It compares combinations of properties as follows. Initially, just the first property is considered. If that property is included in combination x but not in combination y, then x is ranked above y. If there is a tie with respect to the first property, then the second property is considered, in the same way. If there is also a tie with respect to the second property, the third property is considered, and so on. In this way, a Gigerenzer-style 'take-the-best' heuristic can be represented in our framework.

More generally, our framework gives us a way of subsuming several 'piecemeal' decision principles under the umbrella of a more general mechanism, namely that of reason-based choice.

Catherine Herfeld: Again, I understand that it is a valuable goal to unify all our observations and approaches under one single and general theory. However, as people have pointed out, it might not adequately capture the specific characteristics of gut feelings and heuristics when thinking about them in terms of a reason-based or rational choice, which not necessarily but plausibly would rest on the idea that you have certain reasons, you consciously deliberate about them, and then make a choice, whereas heuristics seem to capture exactly the opposite of what 'choice' entails.

Christian List: The label 'reason-based rational choice' makes our approach sound more rationalistic than it is. I would argue that the role played by gut feelings can actually be captured in our framework. Crucially, the reason-based framework allows that properties can make a difference in two distinct ways. One, as I have already explained, is that a property can be motivationally salient so that an agent focuses on this property and evaluates options based on it. This is a fairly consciously aware mode of using properties to make choices. But properties can also influence choices in a more subconscious way. Namely: the properties of a context can affect which other properties become motivationally salient for the agent. It might be, for example, that if a context has one particular property, then this subconsciously leads the agent to focus on certain properties when assessing options in that context.

Let me give you a simple example. Suppose I'm normally quite health conscious when I decide what to eat. I try to avoid foods that are too fatty or that are too sugary, as well as foods that are unhealthy in other ways. So, suppose that in normal circumstances, all of these health-related properties are motivationally salient for me. When I choose between different food items, I look at them through the lens of those properties. If some food has the unhealthy properties, I stay away from it, and if it has the healthy properties, I am drawn towards it. But it might also be that when I enter a restaurant in which there is a strong smell of fried food, my health-conscious attitudes go out of the window. There is this appetizing smell of chips and other deep-fried delicacies, and I develop a craving for fatty food and forget about my health-oriented motivations. In this context, the health-related properties are no longer motivationally salient for me, and I focus on various flavor properties alone: the flavor of fat, sugar, saltiness, and so on. Now in that context, my choice behavior will be very different from my normal one. I suddenly choose unhealthy food items over healthy ones. What has happened is that a property of the context, namely the smell of fried food, has influenced my choices by making different properties motivationally salient for me. But all this may be entirely subconscious. I need not be consciously aware of this. Indeed, that's the sort of thing that often happens in the case of framing effects. And we can naturally explain this in the reason-based framework. In short, the framework is nowhere near as rationalistic as the name might suggest at first sight.

Catherine Herfeld: Much of your work involves axiomatization of some kind. In this regard, you have a lot in common with other decision theorists, such as Itzhak Gilboa, for example. At the same time, large parts of the justification of your account seems to be empirical. You even want to apply your general account to actual behavior as well as empirically test it. Here, your view seems to profoundly deviate from decision theorists such as Itzhak Gilboa, who argues that because most decision theories are axiomatic theories, rational choice theory is a purely theoretical undertaking. Why do you nevertheless think that using the axiomatic method is fruitful to formulate empirical theories?

Christian List: The idea of a reason-based explanation can in principle be spelled out without using the axiomatic method at all. In principle, you could just introduce the notion of a reasons structure and define what it means for a reasons structure to entail a choice function. Then you could call a choice function 'reason-based explicable' if and only if there exists some reasons structure that entails it. You can do all this without using the axiomatic method. But the axiomatic method is useful for showing that the notion of reason-based explicability is systematically related to some other conditions. We can show, for instance, that if, and only if, the choice function satisfies certain conditions, then it can be explained in this particular way.

To illustrate the usefulness of the axiomatic method, consider an example from social choice theory, rather than individual decision theory. Take voting methods such as single transferable vote or alternative vote. Those voting methods are being used in some real-world elections, for instance in Ireland and in Australia. At first sight, they look attractive, and many reasonable people favour them. Indeed, alternative vote was being proposed as the alternative to plurality rule in a referendum in the UK a few years ago.

Now when you study them axiomatically, it turns out that those voting methods violate an important axiom, namely monotonicity. It can happen that additional support for a previously winning candidate turns this candidate into a loser. So, a candidate who won the previous election may rise in some voters' preference rankings while all other preferences remain equal, and yet this candidate may lose the next election, despite having gained additional electoral support. That's a violation of monotonicity of the voting method. Once we see that a voting method violates this axiom, this should ring some alarm bells and prompt us to ask whether there is something conceptually wrong with this voting method. Without the axiomatic method, we might have overlooked this problem.

Generally, by axiomatically characterizing a voting method, we come to a better understanding of its properties. Similarly, if we consider a particular decision principle for an individual agent and we axiomatically characterize it, this can give us a much better understanding of the nature of this decision principle. This is why the axiomatic method is a useful tool, even though we shouldn't fetishize it.

Catherine Herfeld: However, while the axiomatic method has often been appreciated precisely for allowing to investigate a formalized system by way of its properties, it has also been argued that the main value of axiomatization is the guarantee of internal consistency of such a system and the possibility to reach a generality by only formulating those properties of structurally similar phenomena that are highly abstract. Do you think that axiomatic theories also have any empirical value?

Christian List: The axiomatic method is a very good tool for characterizing the logical space of possibilities, such as the logical space of possible decision methods. We can then ask where the decision methods that are being used in practice – in the real world – are located in this logical space. And we can axiomatically compare those real-world decision methods with the decision methods that would be recommended by certain normative criteria. We can ask: what would a normatively appealing decision method look like, and how would it differ from the decision methods that are being used in practice? (For a non-technical overview of my thoughts on the idea of 'cartography of the logical space', though in the context of social choice rather than individual choice, see my 2011 paper, 'The Logical Space of Democracy'.)

Catherine Herfeld: Transitivity is one axiom that is seen as a crucial property of preferences that

are considered to be rational. What do you consider to be the most reasonable axioms that should charatcterize a rational choice?

Christian List: Let me answer this question from the perspective of the reason-based framework. Franz Dietrich and I do not assume from the outset that an agent's fundamental preference relation must be transitive. Rather, we can use our framework to investigate under what conditions the fundamental preference relation is transitive and under what conditions it may fail to be transitive. It is an empirical question whether someone's choice behavior is best explained by a reasons structure with a transitive fundamental preference relation or by one without transitivity.

One reason why transitivity is an important property is this. If a binary relation – whether over options or over bundles of properties – is transitive, then it is a candidate for being interpreted as a 'betterness relation': a relation that expresses betterness comparisons between options, or between property bundles. Of course, it might capture only subjective or agent-relative betterness judgments. But I accept John Broome's point that transitivity is a necessary condition that any betterness relation must satisfy. Of course, not every transitive preference relation can reasonably be interpreted as expressing betterness judgments. But if we have a preference relation that is not transitive, then it is not even a candidate for being interpreted in this way.

Catherine Herfeld: Although you mostly think of your account as a positive account, you also mentioned that theories of rational decision-making can be useful normative accounts. In fact, decision theorists frequently argue that the preference axioms imposed on the preference relation belonging to a rational actor *should* be satisfied. What do you think is the usefulness of rational choice theories as normative theories?

Christian List: Our reason-based framework in its basic form is indeed concerned with the explanation of choices. So, the initial purpose is a positive one. However, we think that the same formal framework can also be employed for normative purposes. Specifically, the framework can be used to give a taxonomy of different moral theories and their implied decision principles (Dietrich/List 2017). If we adopt this normative interpretation, a reasons structure is no longer a pair consisting of a motivational salience function and a fundamental preference relation. Rather, it is a pair consisting of a normative relevance function and a suitably interpreted weighing relation. The normative relevance function specifies which properties of the options are normatively relevant in each context. And the weighing relation compares bundles of properties in terms of their choice-worthiness from a normative perspective. Then one can show that a consequentialist theory (and its associated decision principle) can be naturally represented by a reasons structure in which no context-related properties are deemed normatively relevant, while a non-consequentialist theory can be represented by a reasons structure in which some context-related properties are deemed normatively relevant. So, the distinction between consequentialism and non-consequentialism corresponds to one of our two dimensions of context-dependence.

Furthermore, the distinction between universalism and relativism corresponds to the other dimension of context dependence. A universalist theory (and its associated decision principle) is representable by a reasons structure in which the same properties are normatively relevant in all contexts, while a relativist theory is representable by a reasons structure in which different properties are relevant in different contexts. In this way, our framework provides useful tools for mapping out the space of different normative theories.

Catherine Herfeld: Does your reason-based account have concrete policy implications?

Christian List: In its present form, it is still a relatively abstract theoretical framework, but it can provide some conceptual resources for more applied debates. Take, for instance, the debate about nudging. Cass Sunstein and Richard Thaler have famously argued that policy makers and institutional designers should take into account the findings of psychologists and behavioral economists when designing decision environments in which people make choices. In particular, they have suggested that decision environments should be structured in such a way as to 'nudge' people into making better choices. Those arguments have prompted a lively debate about whether nudging is an acceptable form of paternalism and about whether there can be such a thing as liberal paternalism. For the purposes of our research program, Franz and I do not need to take a stand in this debate – though the debate is important. We think, however, that we can provide some useful tools for explaining the patterns of choice that scholars such as Sunstein and Thaler have in mind when they discuss nudges. The resulting lessons for institutional design will depend on our background assumptions and on our views in moral and political philosophy.

Catherine Herfeld: What do you think are new trends in research about rational choice theory?

Christian List: There are lots of new trends, and I can't enumerate all of them here. I am particularly interested in certain interdisciplinary developments, including the development of connections between economics, psychology, philosophy, and computer science. And I would like to see a mentalistic approach to decision theory thrive. With respect to the research program of reason-based choice theory, I would be thrilled if some empirical researchers such as behavioral economists or psychologists got interested in it and engaged with it in their empirical work. I would be excited to see empirical work on whether reason-based explanations of choice outperform more traditional rational-choice-theoretic explanations, and on whether reasons-based explanations offer greater scope for predicting choices than traditional rational-choice-theoretic explanations do. I myself am not an empirical researcher, so I am not in the best position to design experiments to address these questions, but there is the potential for doing this.

Catherine Herfeld: Are you after truth?

Christian List: Yes, certainly. If one is engaged in the activity of science, then one should be engaged in the quest for truth. But often the truth is elusive and difficult to identify. And sometimes we are faced with competing explanations of the phenomena we are studying, and theoretical virtues such as parsimony, elegance, and explanatory power may be our only criteria for choosing between competing explanations. But ultimately, we are looking for true theories. The fact that I embrace mental state realism rather than instrumentalism reflects my realist, rather than anti-realist, leanings in the philosophy of science. One of the distinguishing characteristics of a scientific realist is precisely that he or she thinks that science seeks to arrive not merely at instrumentally useful theories but at true theories.

References

Dietrich, Franz, and Christian List (2016a) "Mentalism versus behaviourism in economics: A philosophy-of-science perspective", *Economics and Philosophy*, 32(2): 249-281.

Dietrich, Franz, and Christian List (2016b) "Reason-based choice and context-dependence: An explanatory framework", *Economics and Philosophy*, 32(2): 175-229.

Dietrich, Franz, and Christian List (2017) "What matters and how it matters: a choice-theoretic representation of moral theories", *The Philosophical Review*, 126(4): 421-479.

List, Christian (2011) "The Logical Space of Democracy", *Philosophy and Public Affairs* 39(3): 262-297.

List, Christian (2018) “What is it like to be a group agent?” *Noû*s, in press
List, Christian, and Philip Pettit (2011) *Group Agency: The Possibility, Design, and Status of Corporate Agents*, Oxford (Oxford University Press).