Phenom Cogn Sci (2014) 13:217–238
DOI 10.1007/s11097-013-9296-0

Explaining social norm compliance. A plea for neural representations

Matteo Colombo

Published online: 18 January 2013

Abstract How should we understand the claim that people comply with social norms because they possess the right kinds of beliefs and preferences? I answer this question by considering two approaches to what it is to believe (and prefer), namely: representationalism and dispositionalism. I argue for a variety of representationalism, viz. neural representationalism. Neural representationalism is the conjunction of two claims. First, what it is essential to have beliefs and preferences is to have certain neural representations. Second, neural representations are often necessary to adequately explain behaviour. After having canvassed one promising way to understand what neural representations could be, I argue that the appeal to beliefs and preferences in explanations of paradigmatic cases of norm compliance should be understood as an appeal to neural representations.

Keywords Neural representations · Computational neuroscience · Representationalism · Dispositionalism · Social norm compliance · Explicit · Implicit · Tacit mental states

Introduction

Both philosophical and ordinary explanations of social norm compliance typically make reference to beliefs and preferences (or desires)[1] as the fundamental *explanantia* of norm compliance. The typical explanatory pattern for social norm compliance is that people comply with social norms because they possess the right kinds of beliefs (or expectations) and preferences (Bicchieri 2006; Elster 1989; Lewis 1969;

[1]Decision theorists tend to talk of 'preferences' instead of 'desires.' In what follows, 'preference' and 'desire' are used interchangeably, consistently with the accounts of norms I consider.

M. Colombo (✉)
Tilburg Center for Logic and Philosophy of Science, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands
e-mail: m.colombo@uvt.nl

Pettit 1990; Sugden 1986; Ullmann-Margalit 1977). In these accounts, what it takes to have beliefs and preferences such that they explain norm compliance is generally left unspecified. This kind of information about beliefs and preferences is important, however, because it bears, on the one hand, on the type of understanding provided by the explanatory relationship between beliefs/preferences and norm compliance, and, on the other hand, on what types of manipulations and control this relationship can afford.

The goal of this paper is to shed light on the explanatory relationship between beliefs/preferences and norm compliance by defending a variety of representationalism, which I call *neural representationalism*. The view I am going to defend is an epistemological thesis about the explanatory utility of invoking neural representations to explain certain cognitive phenomena and behaviour. If anti-representationalism is the view that adequate explanations of behaviour do not require representations of any stripe (see Chemero 2000 for a critical overview), then neural representationalism is inconsistent with this view. So, in defending neural representationalism, my target is epistemological anti-representationalism.

My argument, in essence, is as follows: The appeal to beliefs and preferences in explanations of paradigmatic cases of norm compliance should be understood as an appeal to neural representations because of the explanatory purchase of neural representationalism. In order to better highlight the explanatory fruits of neural representationalism, I contrast it with an alternative approach to belief and preference, namely: dispositionalism. If dispositionalism is true, then beliefs and preferences featuring in explanations of paradigmatic cases of norm compliance should be understood as behavioural dispositions; representations would be inessential to such explanations.

The paper is in two parts. The first is geared towards explaining what neural representations could be. The second part aims at establishing neural representationalism by laying out three arguments. The first section, "Belief, preference and norm compliance", sets the stage by rehearsing Bicchieri's (2006) account of social norms, which will be useful to contrast dispositionalism and representationalism as different ways of understanding what it takes to believe (and prefer). The second section, "Neural representations and belief/preference explanation in theoretical neuroscience", draws on work from theoretical neuroscience to explain what neural representations could be, and how beliefs and preferences can be understood in terms of neural representations. The second part of the paper begins with the section "Representational-hungry norm compliance", where it is shown that explanation of paradigmatic cases of social norm compliance requires an appeal to representations. This argument relies on Clark and Toribio's (1994) notion of a "representation-hungry" problem domain. "An objection. On Dreyfus's anti-representationalism" tackles the objection that paradigmatic cases of norm compliance do not consist in behaviour in representation-hungry domains by engaging with Hubert Dreyfus's (2002a, b) anti-representationalism. "Explanatory virtues of neural representations. Two arguments" articulates two further, independent arguments for *neural* representationalism. First, by understanding what it takes to have beliefs and preferences in terms of possessing certain neural representations, instead of certain behavioural dispositions, belief/preference explanations of norm compliance become especially apt to facilitate reliable control and manipulation of behaviour. Second, neural

representationalism yields non-trivial understanding of current practice in cognitive neuroscience, where explanation of adaptive behaviour such as norm compliance often involves probabilistic models.

Belief, preference and norm compliance

Bicchieri (2006) articulates one of the most prominent accounts of social norms, "one that explains norms in terms of the expectations and preferences of those who follow them" (Bicchieri 2006, p. 2). The basic idea is that "the very existence of a social norm depends on a sufficient number of people believing that it exists and pertains to a given type of situation, and expecting that enough other people are following it in those kinds of situations" (Ibid.). Social norms are social, for Bicchieri, because we prefer to comply with them only if we believe that most members of our society will do the same, and we believe that most members of our society expect us to follow that norm and may sanction us with a reward or punishment depending on our choice to follow or violate the norm.

What is important for my purposes is Bicchieri's claim that "the belief/desire model of choice […] does not commit us to avow that we always engage in conscious deliberation to decide whether to follow a norm. We may follow a norm automatically and thoughtlessly and yet be able to explain our actions in terms of beliefs and desires" (Bicchieri 2006, p. 3). Bicchieri's line of reasoning can be reconstructed thus:

> P1. Norm compliance behaviour does not generally involve deliberation.
> P2. Deliberation involves "beliefs and desires of which we are aware" (p. 6).
> C1. Norm compliance behaviour does not generally involve beliefs and desires of which we are aware.
> C2. If beliefs and desires feature in the explanation of norm compliance behaviour, then they do not generally feature as conscious states (i.e. states of which we are aware).
> P3. A dispositionalist account of beliefs and desires does not entail that beliefs and desires are conscious states.
> C3. If beliefs and desires feature in the explanation of norm compliance behaviour, then they can be conceived of as "dispositions to act in certain way in the appropriate circumstance" (p. 6).

The first premise points out that most of the time we follow norms thoughtlessly, with no deliberation, by relying on heuristics, of which we are unaware. For Bicchieri, heuristics can support norm compliance by activating default rules cued by contextual stimuli. From this perspective, "norm compliance is an automatic response to situational cues that focus our attention on a particular norm, rather than a conscious decision to give priority to normative considerations" (Ibid., p. 5).

Bicchieri then contrasts the heuristic route to behaviour with deliberation. "Deliberation is the process of consciously choosing what we most desire according to our beliefs" (Ibid., p. 6). If beliefs and desires are conscious states, then—Bicchieri maintains—they cannot play a role in the heuristic route to norm compliance, and they cannot generally play an explanatory role in norm compliance, as norm

compliance is generally automatic, effortless and unconscious. Beliefs and desires, however, need not be conscious states. Therefore, they can feature in our explanations of norm compliance even when behaviour is guided by heuristics. To qualify her position, Bicchieri embraces a dispositional account according to which beliefs and desires are *dispositions to behave* in certain ways under appropriate circumstances. She explains: "to say that someone has a belief or preference implies that we expect such motives to manifest themselves in the relevant circumstances" (Ibid.).

Dispositionalism allows us to make good sense of the appeal to beliefs and preferences in explanations of norm compliance when norm compliance is the outcome of deliberation as well as when it is brought about by heuristics. Bicchieri is suggesting that dispositionalism is a natural way to conceive of beliefs and preferences as unconscious states featuring in explanations of norm compliance, as it does not commit us to maintain that they are necessarily conscious states.

Behavioural dispositions and representations

Dispositionalism can be construed as the conjunction of an ontological and an epistemological thesis. At the ontological level, dispositionalism claims that behavioural dispositions are fundamental to having beliefs and preferences, and that representations are "of only incidental relevance to the question of whether a being is properly described as believing" or preferring (Schwitzgebel 2006/2010). "For someone to believe some proposition *P* is for that person to possess one or more particular behavioral dispositions pertaining to *P*" (Schwitzgebel 2006/2010). In this sense, to believe and to desire are dispositions in the same way as being soluble and being fragile are. At the epistemological level, dispositionalism claims that invoking behavioural dispositions is often sufficient in order to make good sense of belief/-preference explanations of behaviour, and that adequate belief/preference explanations of behaviour do not require an appeal to representations of any stripe (see e.g. Schwitzgebel 2002; Vanderbeeken and Weber 2002).

Bicchieri (2006) leverages dispositionalism in order to make good sense of unconscious states featuring in explanations of norm compliance. Nonetheless, dispositionalism is not the only available option here. Appealing to representations, instead of behavioural dispositions, can also provide us with a perspicuous way to characterize unconscious states featuring in explanations of norm compliance.[2] Representationalism, furthermore, possesses conceptual resources, which are not available to dispositionalism, for distinguishing between conscious and unconscious states as well as between *explicit*, *implicit* and *tacit* mental states. An adequate account of what it is to believe (and desire) should provide us with the resources for distinguishing between these different types of mental states since these distinctions often play an important role in explanations of cognitive phenomena and behaviour (cf. Haugeland 1998, Ch. 7).

[2] Interestingly, some psychological theories seem to take into account some features of both representationalism and dispositionalism: e.g. Daniel Kahneman's dual-process theory, and the distinction between systems 1 and system 2 (cf. Kahneman 2003). I am grateful to an anonymous reviewer to draw my attention to this point.

Types of belief and preference

A cognitive system is said to have the *explicit* belief that *P* (or desire that *Q*) if it possesses cognitive states that carry the right sort of information, *and* this information is tokened in the system. If beliefs and desires are understood as representations, then one has the explicit belief that *P* (or desire that *Q*) if some representational structure with the right sort of content is stored in the cognitive system. For example, Mr. Pink has the explicit belief that everybody is leaving a dollar on the table at the restaurant if a representation with that content is tokened in his cognitive system.

Beliefs and desires are said to be *implicit* if the information that they carry is not actually tokened in the system, but is swiftly derivable from the information carried by explicit beliefs and desires in the cognitive system "via implications that the system could follow" (Haugeland 1998, pp. 142–3). In terms of representations, the distinction between explicit and implicit belief depends on whether the right representation is tokened in the system or not. Since swiftness is a matter of degree, "there will not be a sharp line between what one believes implicitly and what, though derivable from one's beliefs, one does not actually believe," even implicitly (Schwitzgebel 2006/2010). For example,[3] Mr. Pink desires to leave a big tip for the waitress, and believes that big tips impress waitresses. Mr. Pink holds those mental states explicitly, but he does not draw any implication. Nonetheless, the information that he desires to impress the waitress can be swiftly drawn via implications that the system could follow from the information carried by the mental states that he explicitly holds. We can say, then, that Mr. Pink desires, implicitly, to impress the waitress.

'Tacit' is used differently by different authors (cf. Dennett 1982; Engel 2005; Fodor 1968). Here, by 'tacit cognitions' I refer to a kind of competence built into the system and evinced from the behaviour emerging from the workings of the whole cognitive system. Tacit cognitions are neither explicitly tokened nor implied by explicit representations. For example, if people's performance in a number of perceptual tasks approximates Bayesian inference, it can be said that those people are sometimes *tacit* Bayesian observers. Any one component of their perceptual system need not map onto a single component of the Bayesian model. It is their perceptual system as a whole that performs Bayesian inference (cf. Knill and Richards 1996; Colombo and Seriès 2012).

'Tacit,' 'explicit' and 'implicit' are to be distinguished from 'conscious' and 'unconscious.' Conscious beliefs are those that occur when people consciously entertain them. In representational terms, when Mr. Pink is asked to leave one dollar for tip, he accesses and retrieves some of the relevant representations stored in his cognitive system. He thus consciously entertains the belief that the other guys are leaving one dollar for tip. Some mental states or processes are unconscious just in case they cannot be accessed. Thus, even if Mr. Pink tried to identify the types of algorithms implemented by neural activity in his brain when he learns a social norm, he could not have access to them. Identifying such processes takes deep, systematic investigation. Note also that there can be explicit beliefs that are inaccessible to consciousness. In representational terms, one has explicit beliefs that are inaccessible

[3] The following parallels an example in Haugeland (1998, p. 143).

to consciousness, if there are representations tokened in the system carrying the right sort of information that cannot be accessed or retrieved. Chomsky (1980), for example, argues for this possibility when he talks about the representation of a universal grammar in our head. These types of unconscious, inaccessible beliefs are not tacit since they would be actually tokened in the system.

The last distinction I would like to draw is between occurrent and dispositional mental states. We can say that Mr. Pink *dispositionally* believes that most people leave a tip in restaurants, if he has a representation with that content stored in his head, *but* that representation has currently not been retrieved for active deployment for reasoning or decision making. Given eliciting circumstances, when that representation is accessed and retrieved for active thinking or decision making, Mr. Pink *occurrently* believes that most people leave a tip in restaurants. "One needn't adopt a dispositional approach to belief in general to regard some beliefs as dispositional in the sense here described" (Schwitzgebel 2006/2010). To have beliefs and desires as behavioural dispositions is different from having representations that are dispositional, viz. non-occurrent. To say that most of our beliefs and desires are dispositional does not entail a dispositionalist view of what it takes to believe and desire. One can maintain that representations of some sort are essential to believe and desire, and still acknowledge that most of these representations are unconscious or dispositional.

It should now be clear that a dispositionalist account is not the only one that fits the heuristic route to norm compliance. Representationalism can make good sense of different types of beliefs, including the distinction between conscious and unconscious mental states. The reason why dispositionalism should be preferred to representationalism cannot be, therefore, that it makes good sense of belief/desire explanations of norm compliance that pick out unconscious mental states.

The next section lays down the background for my argument. It firstly clarifies one promising way to characterize what neural representations could be; secondly, it suggests how beliefs and preferences can be understood in terms of neural representations in light of a representative case study from theoretical neuroscience.

Neural representations and belief/preference explanation in theoretical neuroscience

Neurons carry information by generating patterns of action potentials, or spikes. Spike patterns carry information about internal and external variables. Cognitive capacities, including the capacity to comply with norms, are enabled by transformations of such patterns of neural activity.

If information is understood in terms of the statistical dependency between a source and receiver (Shannon 1948; for an excellent textbook on information theory see MacKay 2003), then to say that neural spike trains carry information is to say that neural activations are statistically dependent on internal and external sources of information. Neural activations not only are statistically dependent on some source but they also reliably correlate with their source. Neural activations and the sources, or variables, with which they correlate, can be said to constitute a *code*, viz. a mapping from some source alphabet to some target alphabet.

The neural code specifies functional relationships between properties of neural activity and properties of internal or external variables. Although it is controversial what the precise mappings constituting neural coding are (Dayan and Abbott 2001 review most of the options in detail), and by which neural mechanisms such mappings can be carried out (cf. deCharms and Zador 2000; Pouget et al. 2003), *neural representations* can be said the constituents of the neural code. Neural representations can be individuated by encoding and decoding mappings between two alphabets constituting the neural code. Accordingly, neural representing can be described as a two-stage encoding and decoding process.

Neural *encoding* refers to the transduction of some external stimulus (or internal variable) by the system resulting in the spiking of one or more neurons. Neural encoding specifies the functional dependence of some neural property on some property of an external stimulus (or internal variable). Action potentials are the basic units of the encoding alphabet. Neural *decoding* refers to the extraction of information about some external stimulus (or internal variable) from neural spiking. It specifies which property of the external stimulus (or internal variable) s is read-out by the system from the spiking of one or more neurons. Physical properties of internal variables or external stimuli are plausibly the basic units of the decoding alphabet. The estimate $\hat{s}$ yielded by the decoding processing is used by the system to generate behaviour.

To get to grips with the concept of a neural representation as encoding–decoding processing, consider perceptual visual beliefs. Visual neurons code physical properties with their activity in response to stimuli. The action-potential firing rates of neurons in the primary visual cortex reliably co-vary with and selectively respond to properties such as spatial location, orientation and direction of motion of visual stimuli (Hubel and Wiesel 1962). Neural encoding provides a transduction from visual stimuli to neural responses. Given a visual stimulus, neural encoding determines how neural activation in a certain brain area transduces the stimulus in function of some non-neural parameter like spatial orientation. Neural decoding provides a read-out of some visual stimulus from neural spiking. Decoding determines how the visual information carried by activity in a certain neural population yields a perceptual visual belief and is used by the rest of the system to generate behaviour (Colombo 2010 provide further details about this notion of neural representing and its relationship with intentionality).

This way of understanding neural representing allows for the existence of representational hierarchy: complex, abstract representations might be encoded at higher levels in the hierarchy, computed in function of low-level representations *and* of the estimations of a top-down *generative model*—where, here, a generative model is a top-down mechanism aiming to predict the upstream flow of sensory input. Neural representations do not ground only simple, transient, low-level and task-specific states like perceptual beliefs. Transformations of higher-level representations, grounding more complex states and behaviour, might be implemented along a cascade of encoding–decoding operations carried out by cortical processing (Eliasmith 2003; Rust and Stocker 2010).

To relate the notion of a neural representation just canvassed to explanations of norm compliance, let us now consider a case study. This case study will also help us identify how beliefs and preferences could be best understood so as to fit explanatory practice in theoretical neuroscience.

The Trust Game: a case study

The Trust Game is a sequential interaction between two agents, typically governed by a social norm. There are two players in the Trust Game: an investor and a trustee. The investor has to decide how much money out of an initial endowment to send to the trustee. This amount is multiplied by some factor—e.g. three—and then the trustee has to decide how much of the money received to send back to the investor. Both investor and trustee know that the game terminates after a given number of rounds.

The standard game-theoretic prediction for a single, anonymous interaction between two narrowly self-interested, rational players is for the investor to send nothing, as the investor should anticipate that the trustee will not reciprocate. Experimental results are inconsistent with this prediction: investors typically send a significant amount of the initial endowment, and most trustees reciprocate (Camerer 2003).

Ray et al. (2009) aimed to account for these as well as for other findings concerning the neural substrates of social decision making in the Trust Game (King-Casas et al. 2005; King-Casas et al. 2008) within a Bayesian modelling framework. According to their account, each player's cognitive system is comprised of a *generative* and a *recognition* model. The generative model is a probabilistic mapping aimed to predict the other player's observed sequence of decisions in the game so as to determine a decision policy about what to do at a given time. The recognition model is the inverse of a generative model: it is aimed at recognizing what sort of player one was facing, given her observed sequence of decisions in the game.[4] The generative model predicts sensory data (i.e. the opponent's sequence of choices) from hidden causes (i.e. the opponent's cognitive and volitional profile); the recognition model infers causes from data. Adaptive interaction is accounted for by each player's cognitive system acquiring a recognition model that is effectively the inverse of their generative model. This acquisition is driven by the difference between the opponent's observed choices and the choices predicted on the basis of the generative model, viz. by a prediction error. This prediction error can be used to update high-level representations embedded in the generative model (Friston and Stephan 2007).

One noteworthy feature of Ray and colleagues' account of Trust Game interaction is the separation between the computation of social utility (or value) and probabilistic inference, which might correspond to distinct encoding–decoding goings-on implemented by discernible patterns of neural activity. As Ray and colleagues explain: "the separation between the vagaries of utility and the exactness of inference is attractive, not the least by providing clearly distinct signals as to the inner workings of the algorithm that can be extremely useful to capture neural findings (Ray et al. 2009; see also Gershman and Daw 2012). As this separation lends itself to an explanation of behaviour in terms of preferences about payoffs in the game and beliefs dynamics about other players, reference to such representations is relevant to the question of whether an agent complying with a social norm is properly described as believing and preferring. Thus, when two players interact in a Trust Game, some of their high-level

[4] Different types of agents in Ray et al. (2009) account of the Trust Game were defined by the extent to which they were averse to unequal outcomes and by their level of strategic thinking.

neural representations carry information about the other player's profile, some carry information about the action to implement given the current state of the game. With their concerted transformations, these neural representations are responsible of generating adaptive social behaviour.

More precisely, beliefs here correspond to probability distributions over the other player's possible cognitive and volitional profile; preferences correspond to probability distributions over actions. Although there may be no clean separation between belief and volitional systems, the two—and hence belief and preference—can be distinguished based on differences in neural spiking in discernible populations of neurons, which might implement different types of probability distributions and transformation of probability distributions.

Taking stock: I have distinguished between representationalism and dispositionalism as two ways of understanding belief/preference explanation. Both dispositionalism and representationalism can account for unconscious mental states, which can feature in explanations of norm compliance. Neural representations could be probabilistic encoding–decoding cascades implemented by neural spike trains. If what it takes to have a belief (or a preference) is to have certain neural representations, then beliefs and preferences could be understood as certain kinds of probability distributions encoded by neural activity. With this background in place, I now argue for a representationalist approach to explanations of norm compliance.

Representational-hungry norm compliance

My first argument for why paradigmatic cases of norm compliance should be explained by appeal to representations has two premises.

> P1. Internal representations give us unique explanatory leverage regarding agents' behaviour in "representational-hungry" problem domains.
> P2. Paradigm cases of social norm compliance consist in behaviour in "representational-hungry" problem domains.
> C. Therefore, internal representations give us unique explanatory leverage regarding paradigm cases of social norm compliance.

The argument is deductively valid. Premise 1 involves the notion of "representational-hungry" problem domain, which is elaborated by Clark and Toribio (1994). As Clark and Toribio define it, a problem domain is "representational-hungry" just in case "one or both of the following conditions apply:

1. The problem involves reasoning about absent, non-existent, or counterfactual states of affairs.
2. The problem requires the agent to be selectively sensitive to parameters whose ambient physical manifestations are complex and unruly (for example, open-endedly disjunctive)" (Ibid., p. 419).

Clark and Toribio argue persuasively that internal stands-in, or representations, are necessary to successfully tackle representational-hungry problem domains. Representations give us unique explanatory leverage about agents' behaviour in such domains. I take P1 for granted, and focus on P2. I argue that conditions 1 and 2 apply

to paradigm cases of norm compliance. If paradigm cases of norm compliance consist in behaviour in "representational-hungry" problem domains, then representations give us unique explanatory leverage regarding many cases of social norm compliance.

The first case I consider is the type of interaction taking place in the Trust Game. To trust someone implies some degree of uncertainty: you take the risk of betrayal. You repay another person's trust even though it may go against your interest to maximize your profit. When you trust strangers you do not know whether they are motivated only by a selfish desire to take advantage of you. Finding out the preferences and beliefs of other agents in a Trust Game requires an ability to *anticipate* their actions and to reason *counterfactually*. Condition 1 then applies to this case. Anticipation and counterfactual reasoning, as argued by Clark and Toribio (1994), require the use of inner resources. Ray and colleagues' generative model is one type of inner resource that enables agents to behave appropriately even when they have little evidence about the type of player with whom they are interacting. Trust Game-types of situations are therefore "representational-hungry."

Consider another situation. There is this social norm in football: When a player goes down injured, the ball should be kicked out of play to allow that player to receive treatment. If the ball is kicked out of play by the opponents, a further norm is to return the ball to them. These norms have never been formalized in the rules of the game, but furious reactions are likely to ensue if somebody fails to comply with them.

Imagine that you are playing an important football game. You notice that a football player from the opponent team looks as though he is injured. You have the ball and you can set up a teammate for a goal. The decision to pass the ball to your teammate or to throw it out to allow the opponent player to receive treatment takes fractions of seconds. It is very likely to be unconscious and driven by heuristics. Nonetheless, *counterfactual* reasoning and *anticipation* play an important role in this occasion. You need to make a rapid judgement concerning the actual state of the opponent. You need to find out whether the opponent is actually injured. You need to judge what could happen if you played on and the opponent is actually injured; you need to anticipate the reactions of the opponents. Therefore, abilities for counterfactual reasoning and anticipation, even if unconscious and driven by heuristics, seem essential to your decision to comply with the norm.

The same problem domain in football requires that the player who is to make a decision is "selectively sensitive to parameters whose ambient physical manifestations are complex and unruly" (Ibid.). Imagine that the ball has been kicked out of play because a player went down injured. It is time for a throw-in. It is known that you give the ball back if an opponent player deliberately kicked the ball out of play because a teammate of yours was injured. One condition for the player to comply with this norm is that he is sensitive to abstract, relational properties such as "fair play," "reciprocity," and "cheating." The physical manifestation of relational properties such as "fair play" is typically "complex" and "unruly" since whether a pattern of physical features in a social situation counts as "fair play" depends on other features obtaining or failing to obtain in that situation, and on the learning trajectory of the agents involved. In fact, we do not seem to rely on invariant general rules when we identify a certain pattern as "fair play."

In order to track such types of properties, one needs to rely on representations. Given his previous experience in the world of football, the football player has

developed a capacity to track those abstract properties across situations. Clark (2000) calls this capacity *representational re-coding*, whereby complex, abstract relations are re-coded into simple, usable objects. Given a diverse array of perceptual input, representational re-coding allows one to compress that array into an item whose content corresponds to an abstract property. The item can be stored in memory and retrieved for further processing without the need to store and retrieve all of the diverse perceptual inputs underlying it. Before the throw-in, under those circumstances, the player's sensitivity to such properties as "fair play" is important in order to explain his behaviour. Such sensitivity depends on representational re-coding. Since the idea of an internal representation is essential to this kind of re-coding, it follows that the idea of internal representation is essential to explaining the player's behaviour. The problem domain that our football player faces is an instance of a "representational-hungry" problem.

If my accounts of norm compliance in the Trust Game and of two common social norms of fair play in football are correct, then at least some paradigm cases of social norm compliance consist in behaviour in "representational-hungry" problem domains. It follows, from the argument stated at the beginning of this section, that representations give us unique explanatory leverage regarding at least some paradigm cases of social norm compliance. The arguments developed below, in the section "Explanatory virtues of neural representations. Two arguments", will make the case for neural representationalism.

An objection. On Dreyfus's anti-representationalism

Contrary to what I have just argued, according to Hubert Dreyfus, paradigm cases of norm compliance do *not* consist in behaviour in "representational-hungry" problem domains (Dreyfus 2002a, b). Dreyfus claims that "*some* central cases of intelligent behavior do not involve mental representation" (Dreyfus 2002b, p. 414). Social norm compliance falls among those "central cases of intelligent behaviour." For Dreyfus, paradigm cases of norm compliance consist in *non*-representational-hungry behaviour.

Dreyfus (Ibid., pp. 417–8) asks us to consider a situation in an elevator. The elevator stops at the seventh floor and two people step in. The people already in the elevator shuffle and move around until they are at appropriate distance from the others. This is a paradigmatic case of social norm compliance. According to Dreyfus, the situation just described is not hungry for representation. Rather, it is an instance of "skillful coping," which amounts to a spontaneous responsiveness to the demands of a situation. Skillful coping does not require either deliberation or attention, and, importantly, does not involve the representation of goals. If norm compliance is typically an instance of "absorbed skillful coping," so argues Dreyfus, then we should *not* explain norm compliance with recourse to representations.

Representations after all?

Dreyfus draws on Merleau-Ponty's explanatory notions of *intentional arc* and the tendency to achieve *maximal grip*. "The *intentional arc* names the tight connection between body and world" (Dreyfus 2002a, p. 367). It describes a relationship

between an agent's skills and the world: when an agent acquires a skill, becoming an expert in doing something, the skill manifests itself spontaneously given certain solicitations of a situation. The intentional arc does not depend on representations stored in the head; rather, this notion underwrites dispositionalism, as skills underlain by the intentional arc are finer and finer behavioural dispositions to respond to cues in the world. This kind of body-world relationship develops courtesy of extensive interaction with other agents, objects and situations. "*Maximal grip* names the body's tendency to respond to these solicitations in such a way as to bring the current situation closer to the agent's sense of an optimal gestalt" (Ibid., pp. 367–8). Maximal grip describes the process whereby an agent comes to "see" how to be drawn by environmental solicitations to realize a particular goal without representing the goal. In order to give flesh to these two notions, Dreyfus borrows from neural networks modelling and from Walter Freeman's (1991) attractor theory of brain dynamics. He claims: "neural networks exhibit crucial structural features of the intentional arc," and Freeman's account might underlie maximal grip (Dreyfus 2002a, p. 413).

This reliance on neural networks and dynamical system theory is indicative of how Dreyfus conceives of representation. He associates the notion of representation with the "classicist" idea of strings of symbols tokened in a system, which are isomorphic to propositional attitudes (e.g. Fodor and Pylyshyn 1988; Newell and Simon 1972). But this is just one possible way of understanding representation. Representational-hungry problem domains do not require these types of structures, and, in this respect, it is telling that many connectionists as well as Freeman (1991) appeal to representations in their explanations of cognitive phenomena and behaviour.

Clark (2002), commenting on Dreyfus's (2002a), raises exactly this worry: attractor states in dynamical systems and high-dimensional weight spaces of neural networks can be understood "as new powerful kinds of internal representations" (Clark 2002, p. 386). A characterization of neural representations in terms of encoding–decoding mappings also makes good sense of the way neural networks learn, and with the way brains use attractor dynamics. It can be said, then, that social situations like the one in the elevator described by Dreyfus can still be hungry for genuine representations, although non-classicist representations.

Dreyfus (2002b) has two complaints in response to Clark's (2002) worry. He claims that the use of *any* notion of representation in reference to the processing carried out by neural networks and dynamical systems is unwarranted. In those contexts, the notion of representation is too weak "to do the job of showing that *particular* brain states are correlated with *particular* items in the world, let alone that they have content, that is, that they *represent* such particular items under an aspect" (Dreyfus 2002b, p. 420). So, the first complaint concerns the quality of the correlation between neural activation and physical features in the world; the second concerns how neural activations can represent external stimuli under an aspect—e.g. seeing a carrot under the aspect "nourishment."

I believe that Dreyfus's concerns are unjustified. They can be successfully answered in light of the characterization of a neural representation as encoding–decoding cascade. The input stimulus to a neural network is in fact encoded by a certain pattern of activation. From a given activation, the system decodes information about the input, thereby using it to display relevant behaviour and cognitive phenomena.

Although neural networks do not store particular rules for dealing with particular inputs, they give the same or similar outputs to same or similar inputs after training. Encoding–decoding mappings consist of probability distributions, which reliably specify systematic relationships between particular neural activations and particular features in the world. Hence, the notion of representation in terms of encoding–decoding cascade seems actually to be strong enough "to show that particular brain states are correlated with particular items in the world" (Ibid.). Let us now consider the context of brain dynamics.

Dreyfus (2002b, p. 420) recognizes that Freeman himself claims that the brain uses attractors to *represent* causes in the sensorium (but see Freeman and Skarda 1990). Dreyfus, however, asks us to resist representation talk in this case. He points out that "when the rabbit smells and successfully eats a carrot, it forms a new attractor, and that attractor, in an appropriate context, will henceforth cause the rabbit to go for a carrot, this is just a complex physical event" (Ibid.).

Here, Dreyfus is describing an example where a particular brain state can be reliably correlated with a particular feature of the world. The attractor in the rabbit's brain is capable of standing in for the carrot—in situations, for example, where the carrot is not here and now, the rabbit is hungry, and the rabbit is directed towards carrots. In this case, "what makes one want to use representation talk" is *not* as Dreyfus's claim "that the complex event of the system relaxing into an attractor basin is isomorphic with the agent's experience of being drawn towards an equilibrium" (Ibid.). Rather, it is the fact that an appeal to representation is justified, minimally, when an entity stands-in for some possible state of affairs. Since the attractor in the rabbit's brain stands-in for the carrot *and* is consumed by the system to generate behaviour courtesy of decoding, representation talk can be justified also in the context of systemic dynamics.

Dreyfus's second complaint concerns content. He wonders how representations in neural networks and system dynamics can represent "particular items under an aspect." The concern is that, for example, the attractor in the rabbit's brain cannot *represent* carrots as nourishment.

The notion of a neural representation canvassed above can also handle this concern (cf. Colombo 2010). Decoding determines the relevance of the encoding for the system, as it specifies how the information carried by neural activations is consumed by the system to produce behaviour. Within the cognitive system of the rabbit, which might comprise a top-down generative model, the representation of a carrot can be associated with the representation of a high-level property like *nourishment*. Properties such as *edible* and *dangerous* can also be encoded along the cascade of encoding–decoding processes implemented by activity in cortical layers. Encodings of such properties might depend on encodings–decodings of low-level physical variables such as *displacement*, *mass*, *orientation* and so on (Eliasmith 2003, p. 502). So, there is no reason why the attractor in the rabbit's brain cannot be justifiably said to represent carrots under the aspect nourishment.

Dreyfus has not established that the mechanisms that might underlie the intentional arc and maximal grip are representation-free. *Even if* the intentional arc and maximal grip are in place in situations where people comply with norms, those situations can still be representational-hungry.

Shuffling in the elevator. Systemic dynamics and causal couplings

Why do people in the elevator shuffle until they get to an appropriate distance? For Dreyfus, this is an example of "spontaneous absorbed coping." Dreyfus's explanation is that after repeated interaction with others in elevators, people have acquired a disposition to respond appropriately to the solicitations of that kind of situation. Nobody can specify the appropriate distance to maintain. Nobody is trying to get to that distance. People in the elevator are drawn to get there by responding to the whole (elevator—person A—person B—person C—…) situation. They do not respond to particular features of the situation. They do not represent the person who is stepping in as a separate feature. Dreyfus explains: "the embodied agent doesn't *think of* doing what is solicited either. He just lets himself be drawn to lower a tension and straightway finds his body doing what feels appropriate, without needing to, or being able to, represent some desired goal" (Dreyfus 2002b, p. 420).

Dreyfus's explanation is couched in terms of a perception-based, fine-grained behavioural disposition in an extended body-environment system. This explanatory framework has two features: (1) embodied agents respond to the *whole* situation, (2) embodied agents coping with their situation do not have any representation of their *goal*. Dreyfus's main motivation for (1) is causal coupling. That agents are coupled with their surroundings means that the agents continuously affect and are affected by what surrounds them. Coupling can be taken as a reason in support of the arbitrariness of distinguishing brain-centered cognitive systems from the environment where they are embedded (e.g. Beer 2008). Causal coupling, moreover, would constitute a reason to doubt that the situation in the elevator is representational-hungry.

To understand the interactive complexity underlying skilful coping—so runs Dreyfus's argument—we should adopt a "wholist" perspective. According to Dreyfus, the situation in the elevator is best explained in terms of the dynamics of the whole system (elevator—person A stepping in—other persons in the elevator) evolving towards an adaptive equilibrium. This would suggest that paradigmatic cases of norm compliance may not involve either a behavioural or a neural ability, but systemic dynamics. If they essentially involve systemic dynamics, then it is mistaken to view such situations as involving specific representational components.

I am not persuaded by this argument. From an epistemological standpoint, even in cases of skillful coping there can be good, independent reason to ask about information-processing components representing specific features of a situation. A brain mechanism of the capacity to respond to certain external solicitations—e.g. a person stepping in the elevator—can be taken to be coupled to the whole body-environment because we have a representational pre-understanding of its role: we have a pre-understanding of the type of information the mechanism could use and manipulate. Without this kind of understanding, it would be problematic to identify where to apply the dynamicist analysis—whether at the level of brain–body–environment system, or of body–neuromechanical interactions, or neural interactions. We would lose an independent rationale for understanding why we should (de)couple possible components of the system in certain ways rather than others.

The second feature of Dreyfus's account of the situation in the elevator is (2) that the embodied agent coping with her situation does not have any representation of her *goal*. Dreyfus starts with the following puzzle. During skill acquisition, agents

modify their behaviour in function of their results. When the action results in failure, then something needs to be revised. But in order to adjust one's behaviour in function of failure and success, some representation of a goal seems to be necessary. This representation specifies a target state that determines appropriate adjustments in the agent's behaviour. If this is so, then it seems that all skilful action requires goal representation. If one is acting skillfully, then there is something she is trying to do. If there is something she is trying to do, then she is pursuing a goal. Hence, goal representations seem to be necessary for skillful action.

Dreyfus resolves the puzzle by rejecting the first conditional. It is not always the case that if one is acting skillfully, there is something she is trying to do. Dreyfus claims that "[i]n general, we don't have to *try* to comport ourselves in socially acceptable ways" (Dreyfus 2002b, p. 418). We experience such kinds of situations "as drawing the movements out of us" (Dreyfus 2002a, p. 380). In "the *experience of acting*" the direction of causation is not from a represented goal to the world. It is the world itself that initiates certain of our bodily movements drawing us towards appropriate actions: No goal state is pursued in norm compliance. Success in complying with a social norm is assessed as experience of *optimal gestalt*.

Dreyfus's argument is as follows: For some skillful actions such as some cases of norm compliance we experience the situation as drawing the appropriate action out of us. If this is so, then for some skillful action we do not experience our goals as causing our action. Rather than experiencing goals, we experience that the direction of causation goes from the situation to the action itself. Therefore, the representation of goals is not involved in some skillful actions such as social norm compliance.

I think this argument is a *non-sequitur*. Assume that we do not experience norm compliance as caused by the pursuit of a goal. Assume also that sometimes we cannot formulate the goal that we may pursue in certain contexts. For example, we cannot tell what the socially appropriate distance to maintain in an elevator is. From these, it does not follow that the representation of a goal is not involved in norm compliance. It only follows that the representation of a goal in certain instances of norm compliance is not explicit and conscious. In such cases, the representation of the goal may be tacit, unconscious, or dispositional as characterized in the section above "Belief, preference and norm compliance".

The explanatory leverage given by goal representations in the case of norm compliance has to do with both the anticipatory and evaluative nature of goals. On the one hand, goals indicate potential future states of affairs towards which we are driven. They govern our behaviour towards the realization of that state. On the other hand, goals indicate valuable states of affairs. They allow us to evaluate the current state of affairs in function of the target state. When a person steps in a crowded elevator, goal representations provide us with a natural explanation of why people start to shuffle until they reach a certain position. Each agent may have the goal of keeping a socially appropriate distance from the others. The current state is confronted with that goal. If the state fails to fit the goal, a prediction error ensues and some adjustment is required. The goal representation enables the agent both to anticipate what might happen if the target state fails to be reached and to evaluate that certain possible states are "bad" whereas others are "good."

Explanatory virtues of neural representations. Two arguments

The last section argued that Dreyfus's (2002a, b) argument against representationalism is unsuccessful. The types of cases considered by Dreyfus can appropriately be described as being representational-hungry. Yet, as argued by Ramsey (2007), it is always *possible* to treat a system as representational or a situation as involving representations. The challenge is to specify what the positing of representations could give us in terms of non-trivial explanatory purchase. I now address this challenge by comparing representationalism with dispositionalism, focusing on requests for explanation that aim for manipulation and control of social norm compliance.

Neural representationalism: manipulation and control

Suppose that dispositionalism is the right way to think about belief and preference. Suppose that beliefs and preferences understood as behavioural dispositions enter an explanatory relationship with norm compliance behaviour. What is the type of control and manipulations of norm compliance behaviour that such a relationship can facilitate? To what extent does understanding beliefs and preferences as behavioural dispositions facilitate the control and manipulation of norm compliance?

There is good evidence that social norm compliance can be affected by what an agent expects others would do in a similar situation. An agent's tendency towards complying with a norm is also affected by what one believes others think she ought to do in that type of situation. Given some social situation, agents' beliefs and expectations can be manipulated by providing them with information about other agents' judgements and behaviour in the same type of situation. By manipulating agents' beliefs and expectations in this way, their tendency to comply with a given norm in that situation can change (Bicchieri and Xiao 2009).

Assume that beliefs and preferences are dispositions to behave in certain ways under appropriate circumstances. Suppose that in a Trust Game the information provided to the players causally affect their disposition to reciprocate. Since preferences are dependent on beliefs in Ray et al.'s (2009) account of the Trust Game,[5] the provision of a certain type of information about the type of trustee causes the investor to be disposed to prefer, for example, to invest nothing. The investor's beliefs are manipulated by the provision of a certain type of information, which affects the investor's preferences about payoffs. If to prefer something to something else is just the disposition to do what realizes the former thing rather than the latter, given the right triggering circumstances, we should say that information about the trustee's cognitive and volitional profile causes the investor to be disposed to have a disposition to invest nothing. Although it seems odd to say that one is "disposed to have a disposition," on a dispositionalist understanding of beliefs and preferences that is the way we should explain the investor's decision to invest nothing. The investor's preferences would be second-order states elicited by beliefs.

[5] Recall that in their account beliefs about your cognitive and volitional profile influence my preferences about payoffs in the game. Interestingly, also in Bicchieri's (2006) account of norm compliance preferences are dependent on beliefs. On her model, an agent's preferences are conditional on his or her own beliefs regarding other people's actions and expectations. So one *prefers* to follow a norm if he or she *believes* that certain conditions occur.

Accordingly, a possible explanation of the investor's behaviour would run as follows. The investor prefers to invest nothing in the game because she expects that the trustee will not reciprocate. She expects that the trustee will not reciprocate because she has received a certain type of information about her type. The 'because' is causal in both statements. The first 'because' connects two dispositions: an expectation and a preference. The second connects a disposition, viz. an expectation, and a piece of information. In the second statement, we can individuate the cause as a physical, triggering process, viz. the transmission of messages about the trustee's history of plays. This way of individuating the cause enables us to manipulate or control it so as not to trigger the disposition: for example we can destroy the message before it reaches the investor, or we can modify it by adding noise.

If the type of interest that motivates the request for explanation here is for reliable control and manipulation of the explanandum behaviour, two questions are left unanswered by this dispositional account: Why, or in virtue of what, does that message about the history of the game cause the investor to have a certain expectation? Why, or in virtue of what, does the expectation cause the investor to have a certain preference? The answers to these questions are clearly important if we want to intervene causally on the investor's expectations and preferences.

As a parallel case, consider this question: "Why did your mug break when Courtney dropped it?" You can answer: "Because it was fragile and Courtney dropped it." This can be a perfectly adequate explanation, unless we want to know what we should do in order to prevent the mug from breaking when dropped. It does not facilitate us to individuate how and where we should intervene if we wanted to manipulate or control the effects of dropping the mug.

Another possible answer to the question above is: "The mug broke because it has such and such features and structure; and, given this structure, these features and the impact with the floor when Courtney dropped it, the mug broke." This explanation places us in a better position to manipulate and control the effects of dropping the mug. For example, it provides information that would allow for manipulating the atomic structure of the mug in order to control the effects of dropping it. Although there is no appeal to fragility, or to other dispositions, this is a satisfactory explanation, which can facilitate reliable control and manipulation of the explanandum.

It can be complained, however, that this explanation affords impractical manipulation and control: routinely, when ordinary people wish to control the effects of dropping a mug, they intervene on fragility by protecting the mug with some packaging material, rather than on its underlying structure or features. Hence, the disposition-free explanation is in this case practically irrelevant for manipulation and control.

Nonetheless, if we are interested in the behaviour of a functional mug—that is, if we are interested in the behaviour of a mug that does not break easily *and* from which we can drink, this complaint is misguided. Mugs and other fragile objects are ordinarily protected with packaging material when they are shipped or transported. Under those circumstances, mugs are not functional as drinking cups. We have to unwrap them, for mugs to behave as drinking cups. Mugs in fact routinely break when they are used as drinking cups. Thus, a fragility-based explanation of the mug behaviour affords interventions that can control for the effects of dropping the mug, if only at the expense of functionality. Those types of interventions would prevent us

from using the mug as a mug. Instead, explanations of the mug behaviour based on its underlying structure and features would afford interventions that can control for the effects of dropping the mug, while respecting its functionality as a cup, from which we can drink.

If we are interested in information that affords manipulation and control, disposition-based explanations of the behaviour of the mug have another limitation. Typically, dispositions are ascribed as global properties of a whole system. So, the fragility of a glass is not ascribed to any discernible part of the glass. The underlying structure of the glass can be inhomogeneous, instead. The categorical properties of the glass, which describe its discernible features, need not be ascribed to the whole system. Hence, an explanation of the glass behaviour based on categorical properties can provide us with information about where to intervene on the mug in order to control or to obtain certain effects.

Analogously, on a dispositionalist view, beliefs and desires are ascribed as global dispositions of a whole system. Because of this, a belief/desire explanation would not facilitate us to identify *where*—other than at the whole system level—we should intervene in the cognitive system to make a specific difference in its behaviour. Instead, if beliefs and desires are understood as *neural* representations, then we can identify particular signals and specific neural structures, on which we can intervene to make specific differences in the behaviour of the cognitive system.

Recall Ray and colleagues' case study. An explanation of the players' behaviour couched in terms of behavioural dispositions would have difficulties, in comparison to an explanation couched in terms of neural representations, to provide us with an answer to the question why, or in virtue of what, a player manipulates another player's beliefs. An answer to this question will enable us to individuate where and how to intervene in order to cause certain effects. For the discernible neurocomputational signals featuring in their account "mandate probing, belief manipulation and the like" (Ray et al. 2009). Let's make the point vivid.

Suppose that two human agents are playing a Trust Game. The agents' preferences are grounded in neural representations in the form of utility (or value) neurocomputational signals. Their beliefs are grounded in neural representations in the form of inferential schemes embedded in a generative model. Assume that we appeal to these neural representations to explain why a player is playing fair. Identifiable neural populations are the vehicles of the two distinct representations of utility and inference, as in Ray et al. (2009) account. If we knew the details of the algorithms in which such representations of utility and inference are embedded, then it would be possible to estimate in real time the neural computations carried out by a player complying with a norm of fairness in the Trust Game—*granted* that those algorithms and the player's brain activity during the Trust Game are computationally equivalent (Churchland and Sejnowski 1992). The information carried by these neural computations could then be read-out, manipulated, and fed back into the player's brain so that it would make a difference in the neural computations underlying the player's behaviour. An explanation of the player's behaviour couched in terms of neural representations would thus provide us with information that allows for formidable control and manipulation of social norm compliance in that social interaction.

Admittedly, this looks like a science fiction scenario. Research in computational neuroscience and brain–machine interface, however, is beginning to make the scenario just described feasible, at least for simple behaviour. Kawato (2008a) illustrates

with a number of real cases how the combination of computational models, brain–network interfaces—which non-invasively estimate neural activity and read out the information carried by neural activity—and decoding algorithms can foster what can be called *manipulative neuroscience*.

Kawato (2008b) reports on a project where a monkey's brain activity could control a humanoid robot across the Pacific Ocean. In this project, the pattern of activity of certain populations of neurons encoded in a monkey's motor cortex was recorded while the monkey was engaging in a motor task in a lab in the USA. The kinematic features of the monkey's motions were decoded from neural firing rates and sent via an internet connection in real time to a robot located in Japan. Courtesy of this signal, the robot could execute locomotion-like movements similar to those performed by the monkey. Another instance of manipulative neuroscience is the remote radio control of insect flight. Sato and Maharbiz (2010) review studies where insects in free flight are controlled courtesy of implantable interfaces. Courtesy of an implant for neural stimulation of an insect's brain coupled with low power radio systems, the insect can be put into motion, stopped and controlled while it is in flight. In light of this type of research, manipulative neuroscience "has already moved beyond mere science-fiction fantasy in the domain of sensory reconstruction and central control repair as exemplified by artificial cochlear and deep brain stimulation" (Kawato 2008b, p. 139). Hence, it does not seem to be a mere whim of fantasy to expect that non-trivial choice behaviour in social contexts can be manipulated in similar ways.

Neural representationalism: explanatory practice in cognitive science

Another type of argument makes neural representationalism appealing. The notion of a neural representation yields non-trivial understanding of current explanatory practice in manipulative neuroscience as well as in fields of cognitive science, where the use of generative models is crucial for explanation.

If we consider work such as Kawato's (2008b) and Ray et al.'s (2009), it seems that in order to understand current explanatory practice in cognitive science, invoking neural representations is often necessary. Practical manipulations and control of agents' behaviour leverage encoding–decoding mappings between a neural alphabet and a physical alphabet. Identification of neural representations enables one to dissect them into components at lower levels, or to recombine them in ways sensitive to the information they carry. Furthermore, identification of neural representations could facilitate us to guide agents' behaviour in the absence of the properties those representations are about, as in the cases reviewed in the previous section.

In manipulative neuroscience as well as in other fields of cognitive science, generative models play an ever more important role in explanations of behaviour and cognitive phenomena (Tenenbaum et al. 2011). If generative models play this crucial role in some accounts of behaviour and cognitive phenomena, then the notions of *prediction* and *assumption* should be central to such explanations. This is because such notions yield non-trivial understanding of the formalism underlying the typical explanatory pattern based on the workings of hierarchical generative models implemented by cortical neural processing.

The insight given by such types of explanations is that an agent's behaviour is explained by the complex interaction between her assumptions—implemented by

"high-level" cortical activity—about which causes in the world give rise to her sensory signals, and her predictions about incoming sensory signals, mediated by distinct neural subpopulations (Friston and Stephan 2007). The combination of assumptions and predictions implemented by the brain enables the agent to update her knowledge of the environment and to behave on the basis of such knowledge. It seems obvious that neural assumptions and predictions are genuinely representational constructs in this type of (broadly Bayesian) explanatory pattern. For example, investors playing the Trust Game would rely on assumptions about the type of trustee they are facing, and on predictions about which choices the trustee will make, given assumptions about her cognitive and volitional profile. By engaging with the trustee and observing her choice, the investor's neurocomputational system can update her knowledge about the other player, and, on this basis, interact adaptively. Our assumption-laden, prediction-driven neural representation of the world is geared towards facilitating our successful interaction with it.

Conclusion

This paper has argued for a variety of representationalism called neural representationalism as a way to characterize belief/desire explanation of paradigmatic cases of norm compliance behaviour. Three arguments were put forward: first, paradigm instances of norm compliance take place in representational-hungry problem domains; second, neural representationalism facilitates manipulation and control of behaviour; third, neural representationalism yields non-trivial understanding of current explanatory practice in cognitive neuroscience. Hence, belief/desire explanations of paradigmatic cases of norm compliance should be understood in terms of neural representations.

Acknowledgments I am sincerely grateful to Andy Clark, Dave Des Roches-Dueck, Angelica Kaufmann, Julian Kiverstein, Suilin Lavelle, Ray Debajyoti and Mark Sprevak for their generous feedback on previous versions of this paper, and/or for fun discussion of specific ideas in the paper. A special thank you to two anonymous reviewers of this journal for their constructive comments and helpful suggestions. This work was supported by a grant from the Deutsche Forschungsgemeinschaft (DFG) as part of the priority program "New Frameworks of Rationality" (SPP 1516). The usual disclaimers about any error or mistake in the paper apply.

References

Beer, R. D. (2008). The dynamics of brain-body-environment systems: a status report. In P. Calvo & A. Gomila (Eds.), *Handbook of cognitive science: an embodied approach* (pp. 99–120). San Diego: Elsevier.

Bicchieri, C. (2006). *The grammar of society: the nature and dynamics of social norms*. New York: Cambridge University Press.

Bicchieri, C., & Xiao, E. (2009). Do the right thing: but only if others do so. *Journal of Behavioral Decision Making, 22*(2), 191–208.

Camerer, C. (2003). *Behavioral game theory: experiments on strategic interaction*. Princeton: Princeton University Press.

Chemero, A. (2000). Anti-representationalism and the dynamical stance. *Philosophy of Science, 67*, 625–647.
Chomsky, N. (1980). Rules and representations. *The Behavioral and Brain Sciences, 3*, 1–61.
Churchland, P. S., & Sejnowski, T. J. (1992). *The computational brain*. Cambridge: MIT Press.
Clark, A. (2002). Skills, spills, and the nature of mindful action. *Phenomenology and the Cognitive Sciences, 1*, 385–387.
Clark, A. (2000). Making moral space. A reply to Churchland. In: Campbell R, Hunter B (Eds.), *Moral epistemology naturalized*: *Canadian Journal of Philosophy*, Supplementary VolumeXXVI, 307–312.
Clark, A., & Toribio, J. (1994). Doing without representing? *Synthese, 101*, 401–431.
Colombo, M. (2010). How 'authentic intentionality' can be enabled: a neurocomputational hypothesis. *Minds and Machines, 20*(2), 183–202.
Colombo, M., & Seriès, P. (2012). Bayes in the brain. On Bayesian modelling in neuroscience. *The British Journal for Philosophy of Science, 63*, 697–723.
Dayan, P., & Abbott, L. (2001). *Theoretical neuroscience*. Cambridge: MIT Press.
deCharms, R. C., & Zador, A. (2000). Neural representation and the cortical code. *Annual Review of Neuroscience, 23*, 613–647.
Dennett, D. (1982/83). Styles of mental representation. *Proceedings of the Aristotelian Society*, *New Series*, LXXXIII, 213–26.
Dreyfus, H. (2002a). Intelligence without representation: Merleau-Ponty's critique of mental representation. *Phenomenology and the Cognitive Sciences, 1*, 367–383.
Dreyfus, H. (2002b). Refocusing the question: can there be skillful coping without propositional representations or brain representations? *Phenomenology and the Cognitive Sciences, 1*, 413–425.
Eliasmith, C. (2003). Moving beyond metaphors: understanding the mind for what it is. *Journal of Philosophy, C*(10), 493–520.
Elster, J. (1989). Social norms and economic theory. *Journal of Economic Perspectives, 3*(4), 99–117.
Engel, P. (2005). Tacit Belief. In W. Østreng (Ed.), *Synergies: interdisciplinary communications* (pp. 98–100). Oslo: Center for Advanced Study.
Fodor, J. A. (1968). The appeal to tacit knowledge in psychological explanation. *Journal of Philosophy, 65*, 627–640.
Fodor, J. A., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: a critical analysis. *Cognition, 28*, 3–71.
Freeman, W. J. (1991). The physiology of perception. *Scientific American, 264*, 78–85.
Freeman, W. J., & Skarda, C. A. (1990). Representations: who needs them? In L. McGaugh & N. M. Weinberge (Eds.), *Brain organization and memory: cells, systems, and circuits* (pp. 375–380). London: Oxford University Press.
Friston, K., & Stephan, K. E. (2007). Free energy and the brain. *Synthese, 159*, 417–458.
Gershman, S. J., & Daw, N. D. (2012). Perception, action and utility: the tangled skein. In M. Rabinovich, K. Friston, & P. Varona (Eds.), *Principles of brain dynamics: global state interactions*. Cambridge: MIT Press.
Haugeland, J. (1998). *Having thought: essays in the metaphysics of mind*. Cambridge: Harvard University Press.
Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology, 160*, 106–154.
Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *American Psychologist, 58*, 697–720.
Kawato, M. (2008a). From "Understanding the brain by creating the brain" towards manipulative neuroscience. *Philosophical Transactions of the Royal Society B, 363*, 2201–2214.
Kawato, M. (2008b). Brain controlled robots. *HFSP Journal, 2*, 136–142.
King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to know you: reputation and trust in a two-person economic exchange. *Science, 308*, 78–83.
King-Casas, B., Sharp, C., Lomax-Bream, L., Lohrenz, T., Fonagy, P., & Montague, P. R. (2008). The rupture and repair of cooperation in borderline personality disorder. *Science, 321*, 806–810.
Knill, D. C., & Richards, W. (Eds.). (1996). *Perception as Bayesian inference*. New York: Cambridge University Press.
Lewis, D. K. (1969). *Convention: a philosophical study*. Cambridge: Harvard University Press.
MacKay, D. J. C. (2003). *Information theory, inference and learning algorithms*. Cambridge: Cambridge University Press.
Newell, A., & Simon, H. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Pettit, P. (1990). *Virtus Normativa*: rational choice perspectives. *Ethics, 100*, 725–755.
Pouget, A., Dayan, P., & Zemel, R. S. (2003). Inference and computation with population codes. *Annual Review of Neuroscience, 26*, 381–410.
Ramsey, W. M. (2007). *Representation reconsidered*. Cambridge: Cambridge University Press.
Ray, D., King-Casas, B., Montague, P. R., & Dayan, P. (2009). Bayesian model of behaviour in economic games. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems* (Vol. 21, pp. 1345–1352). Cambridge, MA: MIT Press.
Rust, N. C., & Stocker, A. A. (2010). Ambiguity and invariance: two fundamental challenges for visual processing. *Current Opinion in Neurobiology, 20*, 382–388.
Sato, H., & Maharbiz, M. M. (2010). Recent developments in the remote radio control of insect flight. *Frontiers in Neuroscience, 4*(199), 1–12.
Schwitzgebel, E. (2006/2010). Belief. *Stanford Encyclopedia of Philosophy*. URL = http://plato.stanford.edu/entries/belief. Accessed 14 Jan 2013.
Schwitzgebel, E. (2002). A phenomenal, dispositional account of belief. *Nous, 36*, 249–275.
Shannon, C. (1948). A mathematical theory of communication. *Bell Systems Technical Journal, 27*(279–423), 623–656.
Sugden, R. (1986). *The economics of rights, cooperation and welfare*. Oxford: Blackwell.
Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: statistics, structure and abstraction. *Science, 331*, 1279–1285.
Ullmann-Margalit, E. (1977). *The emergence of norms*. Oxford: Oxford University Press.
Vanderbeeken, R., & Weber, E. (2002). Dispositional explanations of behavior. *Behavior and Philosophy, 30*, 43–59.