Philos Stud
https://doi.org/10.1007/s11098-019-01290-4

Two theories of group agency

David Strohmaier[1]

Abstract Two theories dominate the current debate on group agency: functionalism, as endorsed by Bryce Huebner and Brian Epstein, and interpretivism, as defended by Deborah Tollefsen, and Christian List and Philip Pettit. In this paper, I will give a new argument to favour functionalism over interpretivism. I discuss a class of cases which the former, but not the latter, can accommodate. Two features characterise this class: First, distinct groups coincide, that is numerically distinct groups share all their members at all time. Second, we have access to the inner mechanisms of the groups agents, because members know what they have decided on. I construct a counterexample with these features allowing me to reject interpretivism about group agency in favour of functionalism.

Keywords Group agency · Functionalism · Interpretivism · Social ontology · Coinciding groups · Intentional stance

1 Introduction

The debate on group agency raises two questions: Are there group agents, that is groups with minds of their own? And if so, what makes the attribution of propositional attitudes, such as beliefs, intentions, and desires, to them correct?[1] Two theories answer these questions and dominate the current debate on group agency. On the functionalist side, we find the proposals by Huebner (2014) and

[1] In line with the literature on group agency (e.g. List and Pettit 2011; Tollefsen 2015), I will assume that all mental states of group agents, such as desires and beliefs, are propositional attitudes.

✉ David Strohmaier
davidstrohmaier92@gmail.com

[1] University of Cambridge, Cambridge, UK

Published online: 03 April 2019

Epstein (2015, 2017). On the interpretivist side, there are the accounts of Tollefsen (2015), as well as List and Pettit (2011).

In this paper, I will give a new argument to favour functionalism over interpretivism. There is a class of cases which the former, but not the latter, can accommodate. Two features characterise this class: First, they are cases of coinciding groups. Although the groups are numerically distinct, they share all their members at all times. As a consequence, such groups overlap in their behaviour. Second, in the instances of this class the members' knowledge of what they have decided provides access to the inner mechanisms of the group agents. The members know what the group intended, because they participated in the formation of the intention. I construct a counterexample with these features allowing me to reject interpretivism about group agency in favour of functionalism.

I start by introducing the distinction between functionalism and interpretivism, at the heart of which lies the difference between mechanisms on the inside and behaviour on the outside. In contrast to functionalists, who can take inner-mechanisms as well as outer-behaviour as determining the correctness of attributing propositional attitudes, interpretivists limit themselves to outer-behaviour. Consequently, interpretivists are often accused of overgenerating agents. After having discussed the interpretivist responses to this charge, I present my own argument against interpretivism. As I show, especially challenging problems arise from examples of coinciding group agents, undermining defence strategies devised by interpretivists. I end by suggesting how functionalist accounts of group agency can succeed where interpretivists fail.

2 The functionalist principle

Since the introduction of multiple realizability arguments, functionalism has become the standard account of propositional attitudes.[2] One can distinguish a variety of functionalisms,[3] but they share the functionalist principle as a criterion for whether an entity, be it a group or an individual, has a propositional attitude:

> (F) An entity has propositional attitudes, if and only if it has states with the appropriate functional profiles.

Functional profiles are the full specification of the causal roles[4] states of a system have to play to realise propositional attitudes. Ramsey sentences, that is sentences which describe the theory of propositional attitudes completely but replace mention of them with existentially quantified variables (cf. Lewis 1970), provide a way to

[2] For an introductory discussion of multiple realizability see Funkhouser (2007). For a recent critical take see Polger and Shapiro (2016).

[3] For Putnam's machine-state functionalism see the various papers collected in Putnam (1975), for psycho-functionalism see Fodor (1968) and Block and Fodor (1972). See also "Some Varieties of Functionalism" in Shoemaker (2003).

[4] Strictly speaking, a functional profile might also specify other roles, for example whether states ground each other. This subtlety is insubstantial for my argument.

specify such functional profiles. Functionalism of the kind I will discuss claims that propositional attitudes can be defined via Ramsey sentences.[5]

For a garden-variety functionalism, such Ramsey sentences would describe the causal roles of the propositional attitudes.[6] The Ramsey sentence for a belief that it rains might look something like this: ∃x (x responds to evidence of rain appropriately; together with other relevant states, x leads to carrying an umbrella,…).[7] In a complete theory of propositional attitudes, the other relevant states, such as the desire to stay dry, would also have to be cashed out using existentially quantified variables and appropriate description in the Ramsey sentence. In the end, no mention of propositional attitudes should remain in the sentence itself.

Proponents of group agency have a natural affinity with such a functionalism. If all there is to having a propositional attitude is having a state with a functional profile appropriately specifying causal roles, then why should groups not have propositional attitudes? Since this functionalism allows for the realisation of propositional attitudes in widely diverging ways, the realisation by states of a group becomes a live possibility.

Consider a coarse-grained functionalism, according to which "mental states are internal states of an agent that are caused by certain inputs to the system and cause both certain other internal states and certain behaviour outputs, where these causal dynamics *will be specified by common sense*." (Tollefsen 2015: 81, my emphasis) For example, we attribute the belief that the supermarket is at the corner to an individual, because that's where they go when they want to buy groceries. Given the desire to buy groceries, the candidate state for realising a belief about the supermarket location produces the behaviour we would expect from such a belief. The state plays the appropriate causal roles and thus exhibits the functional structure of a belief that the supermarket is located at the corner.

On such a functionalism, attributing minds to groups appears plausible. A corporation could have a state meeting the functional description of a propositional attitude. For example, if a corporation shows the behaviour of entering a market, this might be part of the functional description of having the belief that it can make profit in this market. Again, the state which serves as a candidate for realising a belief state produces the expected behaviour together with other states. The functionalist strategy appears readily extendable so that the main difficulty lies in specifying the functional profiles correctly to attribute attitudes to the appropriate groups and individuals, but not beyond.

Functionalist approaches to group agency have a long history going at least back to Block's China Brain (1978), purportedly a counterexample to functionalism, and

[5] One kind of functionalism not falling under this description might be Millikan's (2004) teleological functionalism. There might be further exceptions.

[6] However, a Ramsey sentence could also specify a realiser and for example define a certain propositional attitude as only being instantiated by a certain kind of neuron. This would be an unusual realiser-dependent functionalism, which I ignore in the following.

[7] In my example, the propositional content of the belief is written into the Ramsey sentence. One might try to avoid this, which has the benefit of allowing one specification for all beliefs. I chose the approach solely for the purpose of illustration.

Brooks' (1986) paper "Group Minds", which endorses the possibility of a city being the functional equivalent of a brain. Huebner (2014) is one of the recent proponents of a functionalist account of group agency. While he argues that the functional profile of propositional attitudes such as beliefs and desires are rather demanding, he suggests that some groups meet the requirements and become full agents. Epstein (2015, 2017) has also endorsed a functionalist theory of group agency.[8] I will argue that such approaches are to be preferred over the interpretivist accounts by throwing up a problem which the functionalists can solve, but the interpretivists cannot.

3 The interpretivist principle

List and Pettit (2011) in *Group Agency*, and Tollefsen (2015) in *Groups as Agents,* endorse Dennett's theory of the intentional stance in place of coarse-grained functionalism. To use the intentional stance "is to set aside non-intentional possibilities of explanation, to presuppose that the system under explanation is an agent, and to try to ascribe representations and motivations to it that make sense of its actions" (List and Pettit 2011: 23).[9] Thus, the intentional stance is a certain mindset with which a situation can be approached for explanation.

Consider an example adapted from List and Pettit (2011: 19–31): I look at a robot's behaviour and predict its future behaviour by ascribing propositional attitudes such as desires and beliefs. The robot moves towards some cylinders that are lying down and puts them upright. I ascribe perceptual beliefs and a desire for upright cylinders to the robot. To test my ascription, I topple a cylinder. The robot puts it upright again. Having acquired the belief that a cylinder is lying down, the robot satisfies its desire for upright cylinders by showing appropriate behaviour. My ascription results in successful predictions, which is explanatorily more powerful than a non-intentional description. Accordingly, the robot is an agent with perceptual beliefs and the desire to put cylinders upright.

In agreement with List and Pettit, Tollefsen describes interpretivism as "the view that, if we can successfully make sense of another being—understand and interpret its behaviour by using our folk psychology—it is an intentional agent" (Tollefsen 2015: 97). We interpret the behaviour of an object by ascribing propositional attitudes, and if the interpretation allows us to predict the object's behaviour, the success supports the ascriptions. In fact, if the predictions consistently prove correct, this renders the ascriptions true.[10]

[8] Epstein's version departs from the others, however. He argues that whether we should attribute agency to a particular group depends on its kind rather than the features specific to the particular group. That is, we would attribute group agency to a particular committee in virtue of it being a member of the kind committee (Epstein 2017).

[9] See also Dennett (1987: 15, 1991b).

[10] There are two comments to be made about my characterisation of interpretivism at this point: First, interpretivism typically also allows for an indeterminism where there is no deeper fact whether an ascription is correct or not (see Mölder 2010: 109–110). Second, I discuss a version of constitutive interpretivism, which makes interpretation from the intentional stance constitutive for having propositional attitudes, for distinguishing various forms of interpretivism see Mölder (2010: 75–85).

To use an example of a group, if the philosophy department's library team shows the behaviour of sorting books on its shelves, I attribute to it the desire to have its books well-ordered and the belief that the sorting behaviour helps to achieve this goal. These ascriptions allow me to predict that if I put a book from the shelf on the table, the team will put it back in the right place. The success of the prediction supports my ascriptions. Furthermore, if all such predictions deducible from the ascription are successful, this makes the ascription true.

The success of behavioural predictions determines the truth of the ascriptions, but since ordinary people miss facts about the behaviour of entities, interpretivists assume an interpreter who is idealised in that she has access to all behavioural facts about the entity in question and has the capacity to process this information. Thus, we can formulate the following interpretivist principle[11]:

> (I) An entity has propositional attitudes, if and only if an idealised interpreter would successfully predict the behaviour of that entity from the intentional stance by ascribing these attitudes.

The underlying intuition is that if an idealised interpreter of the behaviour can attribute propositional attitudes with success we will not ask for anything more. For support, interpretivists often just point to our folk psychological practices (see the Tollefson quote above). From this perspective, the predictive success is all we expect from a system with a mind. Hence, the interpretivists offer their own principle for the attribution of propositional attitudes, raising the question how it relates to the functionalist principle.

4 Functionalism and interpretivism: the outer-behaviour/inner-mechanism distinction

According to Tollefsen and in contrast to functionalism, interpretivism entails that "[p]ropositional attitudes are not internal states of a system but dispositional states of whole systems" (Tollefsen 2015: 110, see also Mölder 2010: 75). On this picture, the truth of an ascription of propositional attitudes does not depend on any facts about the internal life of the entity. In the case of the robot its movements count, but not the calculations by its processor unit.[12] In the case of the library team, moving the books onto the shelf counts, but not the internal deliberation about where to put them.

Functionalist accounts of mind typically do not care whether agents are made of carbon or silicon, but they often demand that internal states fulfil certain functional

[11] One might emphasise in addition that the intentional ascription has to provide more predictive power than other available approaches, but I leave this qualification out from the principle to keep it simple.

[12] As List and Pettit put it, "make-up—be it neural, electronic, or perhaps of another kind—may provide indirect evidence about what performance to expect in different situations, but the performance itself should dictate the representations and motivations we ascribe to the agent." (List and Pettit 2011: 28). At best the internal states can provide evidence, but they are not determining the truth of the ascription.

roles to realise propositional attitudes. Tollefsen denies such internal realisers any constitutive role in attributing propositional attitudes.

The distinction becomes slightly muddled in List and Pettit's work, who not only endorse Dennett's intentional stance theory like Tollefsen, but also consider themselves functionalists of sorts. Since their endorsement of Dennett's intentional stance is beyond doubt (see List and Pettit 2011: 11, 13, 23), we should understand List and Pettit as using an unusually broad notion of functionalism, which does not commit to the existence of internal states and is therefore compatible with interpretivism.[13] This reading is corroborated by List and Pettit's statement that they "gestured towards a functionalist account of mind in analyzing intentional states—beliefs and desires—in terms of the roles they play in directing the agent and guiding action" (List and Pettit 2011: 171).[14] List and Pettit do not mention any internal realization of mental states in their characterisation of functionalism, thus remaining compatible with an interpretivist construal.[15]

In the following I use "functionalism" for non-interpretivist versions of functionalism, that is versions of functionalism according to which the Ramsey sentence of propositional attitudes involve internal states. Thus, the separation of the two theories relies on a distinction between outer-behaviour and inner-mechanism. Interpretivism relies on a criterion for this distinction, which is surprisingly hard to pin down.

For individuals giving an approximate criterion for the outer-behaviour/inner-mechanism distinction proves easy enough:[16] Everything that stays within the skull is part of the inner-mechanism rather than the outer-behaviour. For example, Dennett contrasts his position with that of Fodor who holds "that the pattern of belief must in the end be a pattern of structures in the brain" (Dennett 1991b: 30, see

[13] While this notion of functionalism is in conflict with Tollefsen's way of carving up the theory space, it is by no means unprecedented. Dennett has described himself as a functionalist (e.g. Dennett 1991a: 31) without having given up on his interpretivism.

[14] List and Pettit also refer to the David Lewis papers "How to Define Theoretical Terms" and "An Argument for the Identity Theory" in this context. It is not entirely clear what to make of this reference. For example, in "How to Define Theoretical Terms", Lewis (1970) expresses the hope that we can make sense of theoretical terms such as those for propositional attitudes without multiple realizability. That would not bode well for group agency. The importance of Ramsey sentences is compatible with interpretivism, since the latter only puts restrictions on what the sentences can contain: no reference to internal states. "An Argument for the Identity Theory" Lewis (1966) argues for the identity of every experience with physical state. He thinks of neurochemical states in particular. This paper is harder to square with interpretivism of the Dennetian kind, but if one takes behavioural states instead of neurochemical states as the realisers, it becomes possible.

[15] Tollefsen reads List and Pettit as suggesting "that the formation of group judgements […] somehow realizes group beliefs" (Tollefsen 2015: 81). According to Tollefsen, List and Pettit endorse a coarse-grained functionalism in which the functional profile includes internal mechanisms. I consider this a misreading, which is invited, first, by List and Pettit's unusually broad notion of functionalism and, second, by their extensive discussions of the internal mechanisms of group agents in the later parts of their books. But this discussion does not concern the constitution of propositional attitudes.

[16] This matter is complicated by the debate on the extended mind (see Clark and Chalmers 1998), which is also discussed by Tollefsen as supporting group agency (2006, 2015). In the following, I will leave this complication aside.

also page 42). We might make some exceptions, but as a general heuristic the skull- or intracranial-criterion will do.

This criterion, however, proves unsuitable for drawing the distinction with regard to groups. Groups do not have a skull of their own. Neither List and Pettit, nor Tollefsen, offer an explicit criterion suited for groups, leaving it to their readers to judge where inner-mechanism ends, and outer-behaviour starts. But while explicit limitations are lacking, the dialectic of the debate imposes restrictions on how to construe the outer-behaviour/inner-mechanism distinction.

For interpretivism about group agency to be interesting, enough events need to fall into the inner-mechanism category. An over-extended category of outer-behaviour renders the difference to functionalism negligible. Interpretivism about group agency would not be of interest if it used all information about documents and discussion within the group as the basis for attributing propositional attitudes.

In the case of individuals, any neural behaviour falls into inner-mechanism rather than the outer-behaviour box. Otherwise Dennett's attempt to distinguish himself from Fodor's position would fail. But if individuals are roughly the analogue to neurons for group agents, does then fall everything that is individual behaviour into the inner-mechanism box too? One should not push the analogy between neurons and individuals beyond its limits. While Tollefsen (2015: 106–107) suggests that an interpretivist can ignore some individual behaviour, for example they could ignore the actions of individual managers to predict the Ford Motor Company's response to an increase in gas prices, ruling out all individual behaviour goes too far.

Consider the library team sorting the books again. If we ignore the behaviour of all members of the team, no group behaviour would remain. The library team showed the behaviour of sorting books, but so did the relevant individuals. At least for groups exhaustively constituted by individuals we want to allow that an event can be both outer-behaviour of the group and individual behaviour. However, this individual behaviour should not be internal to the group, but rather relate it to the outside. In the book sorting case the individual members engage with the books, which are external to the library team. A deliberation about book sorting, however, would remain internal, because the members only engage with one another.

Let me then propose the following criterion: A behaviour of a constituent[17] of a group is an instance of an outer-behaviour of the group if and only if the behaviour engages with an entity external to the group. Only those outer-behaviour events of constituents of the group are outer-behaviour events of the group which involve non-constituents as well.

Without limiting which individual behaviour counts as outer-group behaviour in such a way, the interpretivist approach to group agency is not interestingly different from its functionalist rival. After all, the functionalist will often identify the internal states of group agents realising propositional attitudes with states of individuals and their behaviour. The constituents of these states can then no longer serve as the

[17] I use the notion of a constituent rather than a member to allow that more than individuals might constitute a group and contribute to the behaviour of the group.

based for applying the intentional stance on pain of rendering interpretivism and functionalism effectively equivalent.

Of course, difficulties remain for applying the distinction and a full interpretivist account would have to address them. For example, my criterion leaves open what counts as an engagement or involvement of non-constituents. The interpretivist should sharpen the criterion further such that neither too much nor too little ends up counting as external behaviour. Moving air or radiating heat are not enough engagement with external entities for the behaviour of a constituent of a group to count as an instance of an outer-behaviour of the group. Otherwise all behaviour by group members, including internal deliberations between group members, would count as outer behaviours of the group. Our neuronal activities also create external traces in an MRI and interpretivism does not consider them behaviour.

For the purpose of the present paper, we will gloss over the challenge of specifying the criterion further, which is only charitable towards the interpretivist. If it turned out that there is no principled criterion for distinguishing outer-behaviour from inner-mechanisms, then this would just settle the issue in favour of functionalism. My counter-example will apply independently of interpretivists resolving this difficulty.

5 The overgeneration worry

While only taking outer-behaviour into account distinguishes interpretivism from functionalism, it raises the worry that interpretivism hopelessly over-attributes agency. The worry suggests that this limitation on the basis of interpretation leads interpretivism to attribute agency to too many groups (as well as other entities). Although my counterexample will not concern such an *overgeneration* of agency attributions, it also turns on interpretivism limiting the basis of interpretation to outer-behaviour. Hence, the overgeneration worry and the responses to it have bearing on my counterexample. We will first consider the general version of the overgeneration worry as it has already received ample discussion in the literature and then turn to the special case of interpretivist accounts of *group* agency.

A variety of authors have suggested that interpretivism wrongly attributes mental properties to systems which lack them. Lycan (1987: 5) proposed the tinfoil man against behaviourist theories of mind,[18] Peacocke (1983) brought in Martian marionettes, and Block (1981) has his Blockheads. These are all variations on a theme. I focus on the Martian marionettes, since Peacocke explicitly directs the example against Dennett's theory.

Peacocke invites us to imagine a human body without a brain; instead a radio transmitter controls the nerves. A computer on Mars calculates the behaviour and manages to make the body, its marionette, behave like an ordinary human. As Peacocke observes, the marionette "is voluminously and reliably predictable via the

[18] Bryce Huebner (2014: 90) has pointed out that the tinfoil man example speaks against attributing agency to groups based only on the intentional stance since it speaks against interpretivism in general.

intentional strategy, as voluminously and reliably as for any normal human being" (Peacocke 1983: 205). Interpretivism gives the wrong conclusion for the Martian marionette. It finds an agent where there is only a puppet. In this case, looking only at the outer-behaviour appears to give the wrong answer. A functionalist would look under the skin and consider internal states. The interpretivists cannot accept this solution and has to come up with another one.

Bruno Mölder, whom Tollefsen (2015: 97) quotes as a recent defender of interpretivism, has responded to the marionette-type counterexamples. He admits that looking at predictive success is insufficient, but suggests that folk psychology imposes restrictions on the possible objects of the intentional stance. According to Mölder "[i]t is part of folk psychology that people have beliefs whereas tables and lecterns do not" (Mölder 2010: 193). The Martian marionette does not fall into the range of objects covered by folk psychology, because it does not take objects with empty heads seriously. If folk psychology only covers a limited range of objects, then it constrains our interpretations and rules out Martian marionettes. To avoid this kind of response, my counterexample to interpretivist accounts of group agency will concern mundane examples of groups.

Turning from the general case of over-attribution to the special case of overgenerating group agency, Mölder's solution falls short. Many troubling instances of overgeneration of group agency do not rely on sci-fi scenarios and empty heads, but rather concern actual groups. For many groups, we can raise the question whether we will not always have some predictive success using the intentional stance towards it. Might we not predict some outer-behaviour of the global human population using intentional vocabulary? For example, we can roughly predict the behaviour of humanity by ascribing to it the intention to increase the global temperature. Nonetheless, humanity does not form a group agent, and certainly not one intending global warming. Interpretivist accounts of group agency have to provide criteria for excluding such attributions of propositional attitudes to groups.

List and Pettit (2011: 24–25) introduce strict rationality criteria, for example that attitudes must track facts, that attitudes be internally coherent, and that actions must follow attitudes. Using these criteria List and Pettit generate much fewer ascriptions of propositional attitudes. For example, they would no longer attribute such attitudes to humanity as a group since its behaviour does not allow such rational interpretation. For most intentions we might ascribe to humanity, we will find violations of rationality. While we have some predictive success attributing the intention to increase the global temperature, some of humanity's behaviour, such as a reduction of coal plants, conflicts with it.

In addition to the rationality criteria, Tollefsen (2015: 102, 108) reminds us that Davidson suggested linguistic intelligibility as an interpretivist requirement. An agent must engage in linguistic behaviour which we can interpret. Since we cannot attribute any rational linguistic output to humanity, it is no agent. In effect, Tollefsen narrows further what counts as a successful predictive interpretation from the intentional stance: It must include successful predictions which are either based on or concern linguistic behaviour.

List and Pettit, and Tollefsen have managed to narrow down the number of group agents they postulate. Independently of whether these restrictions succeed in blocking the overgeneration of group agents, they limit potential counterexamples for my purposes. The group agents I present have to meet List and Pettit's rationality restrictions and engage in linguistic behaviour as demanded by Tollefsen. Still, even with these limitations, interpretivist accounts cannot evade the problems resulting from *coinciding* group agents.

6 Towards the counterexample

The interpretivist principle implies that any difference between the mental lives of two group agents can be explained with reference to the outer-behaviour of the groups by an idealised interpreter. This consequence of the interpretivist principle creates a problem for certain cases of coinciding group agents.[19]

Consider two coinciding groups, which are completely constituted by human individuals. Perhaps no such individualist constitution holds for Tollefsen's example of the Ford Motor Company. The manufacturing plants, paper contracts, or its computer systems might also partially constitute the company,[20] but that is not the case for all groups. To keep things simple, we consider the committees of a philosophy department and stipulate that they are exhaustively constituted by their individual members.[21]

The philosophy department creates these committees by randomly assigning its members to them. It throws all the names of the faculty in a big urn and picks five from it to establish one committee. To form the next committee, it throws the five names back in and draws again. Following this procedure, the department fills up one committee after another. In our example and by pure chance, the teaching committee ends up with the same members as the public engagement committee.

As stipulated, the committees are not constituted by anything other than the individuals and their constituents. In such a case, each event that is an outer-behaviour of the teaching committee is also an outer-behaviour of the public engagement committee. This overlap in behaviour follows from sharing the material constitution and the limited notion of behaviour at play here. The shared constitution base is stipulated by the example, but the notion of behaviour follows from interpretivism: On an interpretivist account, behaviour must precede any attributions of mentality since it serves as the basis for such attributions from the

[19] Coinciding groups have been endorsed by most account of group ontology, including Gilbert (1992: 220–221), Uzquiano (2004), Sheehy (2006), Ritchie (2013, 2015), Epstein (2015, 2017), and Thomasson (2016).

[20] Brian Epstein's way of handling differs from that of other authors. While he states that "[a] group is a thing constituted by and only by individual people" (Epstein 2015: 133), he considers many social entities which are often treated as groups not to have the required individualist constitution. For example, he asserts that corporations and universities are not groups (see Epstein 2015: 133).

[21] This individualistic constitution is not strictly a requirement for the counterexample, as along as other constituting entities are not involved in the particular decision processes we discuss in the course of the example. The assumption only serves to simplify the discussion.

intentional stance. Thus, the description of the groups' behaviour cannot depend on any ascriptions of mentality. It has to be *outer*-behaviour and what remains is the criterion we have introduced above: A behaviour of a constituent[22] of a group is an instance of an outer-behaviour of the group if and only if the behaviour engages with an entity external to the group. Clearly, if two groups share all constituents, in this case their individual members, then their outer-behaviour coincides. It follows that when the teaching committee shows the behaviour of sending out a text about pedagogical methods, the public engagement committee shows the same behaviour.[23]

Given that the two groups exhibit the same behaviour, the question arises which of the two committees had the intention to distribute the text. Assume that it was the teaching committee, not the public engagement committee, which intended to send out the text. How can an interpretivist explain this difference in propositional attitudes between the groups given their behavioural coincidence?

While it becomes difficult for interpretivists to pry the mental lives of two coinciding group agents apart, they can still respond to this example. The behaviour of the two groups might be indistinguishable, but the features of the behaviour clearly indicate which committee is acting in sending out the text. The email might be signed by the teaching committee or it might be sent from the email account of the teaching committee. While the behaviour is the same, its features allow for differences in propositional attitudes between the groups from an interpretivist perspective.

Many cases do not allow this response, however, because the behaviour lacks any disambiguating features. For example, one of the groups might intend to engage in some clandestine action that leaves not such identifying behavioural traces. Assume that the teaching committee forms an intention to play a prank on the department. The group decides to hide alarm clocks, which will disrupt lectures. Clearly, in this case the behaviour will not carry the signature of the teaching committee. The outer-behaviour consists only of the members of both groups buying alarm clocks and stowing them away in hidden corners. There is no sign here which of the two groups intended the action.

Or consider an example where the public engagement committee invites everyone to the pub by sending around an email from an account which is not clearly linked to one or the other group. If the email lacks a signature of the committee and any other identifying marks, then one cannot tell from the behaviour which group intended to send out the pub invitation.

Generally, any example will do in which the outer-behaviour has not features justifying a difference in attribution of propositional attitudes. Sending unsigned

[22] I use the notion of a constituent rather than a member to allow that more than individuals might constitute a group and contribute to the behaviour of the group.

[23] One option I am not considering here and in the following is that groups might exist intermittently, so that the teaching committee only exists whenever the public engagement committee doesn't. While this would work against my examples here, they could easily be adapted to avoid this issue. The problem does not arise for my final counterexample.

emails from generic accounts, playing hidden pranks, or any other action either group might engage in suffices. These cases are so easy to come by for two reasons:

1. The behaviour of the coinciding groups is shared, that is the outer-behaviour in which interpretivism is interested does not differ.
2. Often the only identifying marks are found in the deliberation within the group, but the deliberative behaviour between group members does not count as group behaviour.

As discussed above, a behaviour of a constituent of a group is an instance of an outer-behaviour of the group if and only if the behaviour engages with an entity external to the group. In both the case of the prank and pub-invitation, knowledge of the internal deliberations i.e. knowing at which group meeting the members decided to pull the prank and at which they decided to invite everyone to the pub, would allow a difference in attribution of propositional attitudes. The interpretivist, however, cannot take these inner-mechanisms as difference-making on pain of becoming indistinguishable from the functionalist.

These examples put pressure on the interpretivists, but they have still one trick up their sleeves. In addition to actual behaviour, behavioural *dispositions* might count in the attribution of propositional attitudes.

7 The dispositionalist complication

Interpretivists do not just look to realised behaviour but consider also behavioural dispositions. For example, Tollefson writes that "[p]ropositional attitudes are not internal states of a system but dispositional states of whole systems" (Tollefsen 2015: 110). Such dispositions might allow for fine-grained attributions of propositional attitudes when actual behaviour fails to do the job.

This dispositionalist move solves the problem with the prank and pub cases. The two committees have dispositions to show behaviour clarifying which committee played the prank or sent out the pub invitation under appropriate circumstances. Prompted by an inquiry, perhaps enforced by threats of being fired, the groups would clarify the intentions of the committees. However, the dispositionalist move risks making interpretivism uninteresting as a position.

As emphasised, interpretivism is substantially different from functionalism because it avoids giving the inner-mechanisms a role in making ascriptions of propositional attitudes true. Whatever the behavioural dispositions are, they must be such that they are not merely read off the inner-mechanisms. Since interpretivism takes the attribution of propositional attitudes to be all about prediction of behavioural patterns, pointing to an unrealisable disposition, for example a finkish disposition, is an illegitimate move within the framework. Only by limiting themselves in such a way, can interpretivists turn dispositionalist without becoming uninteresting.

This limitation of the dispositional move is corroborated by Mölder's discussion of Dennett in which the former argues for the following construal of behavioural

dispositions: "On this reading, the dispositional patterns are just any patterns of the system that are apt to appear or become manifest to interpreters." (Mölder 2010: 112) While this quote offers a broad notion of dispositions, it acknowledges the importance of the pattern *being apt to manifest itself* to an interpreter.

For the previous examples, this limitation does not matter. The teaching committee has a plausibly realisable disposition to reveal its authorship of the prank. The public engagement committee has a disposition to clarify its intention to invite the others to the pub and this disposition is apt to manifest itself to an interpreter. The nature of group agency, however, allows us to construct a revised and final counterexample.[24]

8 The counterexample

Assume that for some arcane reason of university bureaucracy, both the teaching committee and the public engagement committee have the capacity to dissolve themselves and the other by simply intending to do so. At any point in time, each group can end its own existence or that of the other group by forming the appropriate intentions. Let us, furthermore, assume that the two groups have a joint session at which both of them end their tenure by forming appropriate intentions. The two intentions are instituted by a committee member stating: "It is hereby decided that the teaching committee intends to end its tenure and the public engagement committee intends to end its tenure."

Both groups came to an end, but how could interpretivism settle which group intended to dissolve which? Did each group intend to dissolve itself or the other? Or perhaps one group intended to dissolve both? The interpretivist cannot tell. The actual behaviour of the groups doesn't justify a difference in attributions. There are also no dispositions that could be realised after the act of forming the intentions since the group life ceases immediately. To use Mölder's phrase, the dispositions have to be the "apt to appear or become manifest to interpreters", but the groups cannot manifest any dispositions after they stop existing.

One might try to solve the problem by pointing to behavioural dispositions prior to forming the intention.[25] For example, the groups might have dispositions to clarify which intentions they are about to form. Just interrupt the speaker after "It is hereby decided…" and before the end of their sentence and ask about what will be decided. Under such circumstances the groups would reveal the authorship of the intentions. However, such cases can also be ruled out with a simple tweak, namely the introduction of randomness.[26] Let the group member say: "It is hereby decided that if the coin comes up head the teaching committee intends to end its tenure and the public engagement committee intends to end its tenure, and if the it comes up tail the

[24] I thank Luca Barlassina in helping me to develop this counterexample.

[25] I thank Yonatan Shemmer for pressing me on this point.

[26] A random event can here just be any event which is not the realisation of a prior disposition. Thus, the use of randomness does not rely on indeterminism.

intentions will be the other way round." They then proceed to flip a coin and the intentions are formed accordingly. Now neither dispositions prior to the formation of the intentions, nor any later dispositions, can reveal the content of the intentions.[27] Consequently, the groups' intentions remain inaccessible to the interpretivist onlooker.

To summarise my counterexample, we have here a difference in the mental lives of the two committees, each intended to end itself rather than the other, although it could have intended to dissolve the other. The interpretivist faces a puzzle: Which group formed what intention? The group behaviour and realisable dispositions do not allow us to answer the question.

I do not deny that the *group members* exhibit a different behaviour because of the groups' intentions. Asked which group formed which intention, they can clarify that each group intended to end itself and not the other. But our neurons also exhibit behaviour and dispositions when we form an intention, and interpretivism is committed to not giving them a constitutive role. Only the realisable outer-behaviour of the entity in question counts, and the groups cannot show such behaviour after the end of their existence. Interpretivism does not have the resources to justify the correct attribution of different intentions to the two group agents.

The two special features of group agency pose a particularly difficult challenge to the interpretivist: Since the groups coincide and are exhaustively constituted by human individuals, they show no behavioural difference. Nonetheless, the interpretivist will find it hard to deny the mental difference between the two groups, because we have access to the mental lives of the groups. We can simply ask the members about the meeting.

9 Four interpretivist responses

An interpretivist might respond in four ways to my challenge. First, they might be tempted to deny that we have two groups and try to collapse them into one. They would attribute the intention to end the committee tenures to the one overarching group.

But collapsing the groups goes against the growing consensus in the group ontology debate (see Gilbert 1992: 220–221; Uzquiano 2004; Sheehy 2006; Ritchie 2013, 2015; Epstein 2015, 2017; Thomasson 2016): groups which share all their members can remain numerically distinct. Generally, the argument relies on Leibniz's law that a group cannot differ from itself in its properties. Distinct coinciding groups, so the literature has argued, exhibit different properties. For example, the teaching committee has duties and a structure which the public engagement committee does not have. The groups differ in their properties. Leibniz's law dictates that they cannot be one group. We must attribute the intentions to separate groups.

Second, interpretivists might accept that the groups are numerically distinct but suggest that it is indeterminate which group had which intention.[28] There would just

[27] I used the coin example for simplicity. If we assume that some randomness occurs in the decisions of individuals, it would suffice to let them decide. In this case, there would also be no doubt that we are dealing with an inner-mechanism.

[28] Dennett (1991b: 48) allows for such indeterminacy.

not be a matter of the fact whether the teaching and public engagement committee each ended themselves or the other.

My response relies on the access the individual members have to inner workings of the groups. If we interrogate them, they will all tell the same story about the group decisions. Assuming they all properly listened during the meeting, they will insist that they *know* that each group ended itself rather than each other. They will say that they were there when it happened, that they were witnesses to the decision process. There is a matter of the fact that the coin came up one way or the other at the group meeting. To those directly involved in the group's mental life it doesn't appear indeterminate.

Given these details about the meeting, that is the inner-mechanisms of the groups sustaining the outer-behaviour, it appears wrong to endorse indeterminism regarding the propositional attitudes to the two distinct groups. From the inside everything appears wholly determined. For the individual members there is no reasonable doubt about what happened. At least for our example with all its details, endorsing indeterminism amounts to little more than an admission of inadequacy.

Third, interpretivists could attempt to deny that the groups formed any intentions at all, because they did not fulfil the requirement for propositional attitudes. This denial of a mental life would result if we had run afoul to any of the restrictions Mölder, List and Pettit, and Tollefsen imposed in response to the overgeneration problem.

As discussed, Mölder objects to the far-fetched sci-fi character of many counterexamples the literature offers regarding interpretivism. Martian marionettes lie outside the challenges we usually face in attributing agency. But in the case of the teaching and the public engagement committee we do not have an empty head and radio controls. We have two groups, which look like agents with different mental lives. Folk psychology rules out neither group as an odd fringe case. Compared with other candidates for group agency, such as the highly complex nation states or multinational cooperations, the committees of a philosophy department are innocuous. Excluding these two committees as weird from the scope of interpretivism would be tantamount to admitting that the interpretivist theories of group agency offered by Tollefsen and List and Pettit are bust.

List and Pettit restrictions do not apply since there is nothing irrational about the propositional attitudes of the committees as described in our counterexample. They behave perfectly rationally as far as our example goes. Tollefsen's requirement of linguistic interpretability poses no problem either, since we can stipulate that the committees engaged in interpretable linguistic behaviour prior to forming their self-dissolving intentions.[29]

However, even if interpretivists could find non-ad-hoc restrictions ruling out the attribution of intentions in our counterexample, this would only suggest that they are overly strict since the example provides such a clear example of groups with

[29] In my counterexample the groups can never comment on these propositional attitudes themselves. It would be overly demanding, however, if Tollefsen insisted that all propositional attitudes have to be associated with linguistic expressions. Surely, we don't need to comment or be disposed to comment on *all* our propositional attitudes. So why insist that committees have to?

propositional attitudes if there are any. Interpretivists would be left with too few group agents by ruling out groups resembling these committees.

As a fourth response and last refuge, interpretivists could argue that I have drawn the line between inner-mechanism and outer-behaviour wrongly. If the deliberation between the teaching committee members, their discussion which group intended what, were to count as outer-behaviour of the committees, then we could account for the difference in the mental lives. The behaviour of the members deliberating for the committees would allow the interpreter to make the correct attribution.

But this retreat renders interpretivism about group agency uninteresting. Interpretivism about individuals is a substantial thesis, because it stops, roughly, at the skull. As discussed, interpretivism about group agents is only interesting if it endorses an analogous limit. Tollefsen (2015: 106–107) herself recognises this in her discussion of the Ford Motor Company: The interpretivist position is interesting because it allows us to ignore the behaviour of individuals, the discussion between the president of the company and other members, and exclusively considers the outer-behaviour when we attribute to the company the intention to raise prices. If the deliberations between the members also go into the outer-behaviour box, then too little difference to functionalist accounts of group agency remains. Similarly, an interpretivist who responds to worries about misattributing intentions to an individual cannot just solve the issue by categorising neuronal processes as behaviour of the individual.

In conclusion, my counterexample of the coinciding groups establishes that interpretivists cannot offer an interesting account of group agency and correctly attribute different mental lives to the coinciding group agents in our example.

10 Towards a functionalist solution

If interpretivism cannot distinguish between the mental lives of the two committees, how do we pull it off? We know that the teaching committee intended to dissolve itself, because we know that its members discussed this at the meeting and that the coin came up a certain way. By looking more closely at the interaction of the members, we can pry apart the mental lives of the two groups. A functionalism which considers *how* the members realise the group's propositional attitudes has the capacity to provide the correct answer.

Functionalist proponents of group agency have to explain why details about the group members matters for the realisation of the teaching committee's intention. They must specify the Ramsey sentence for propositional attitudes in such a way that the intentions are attributed to the appropriate committees. It would be no good, however, to mention the awareness of group members as a condition for a group to have an intention. At least, if we want to use the exact same analysis for propositional attitudes for all kinds of agents, then the condition has to work for individual as well as group agents. Thus, an explicit mention of group members in the Ramsey sentence is ruled out.

Nonetheless, the functional profile can be such to allow a difference in the attribution of the intentions. For example, the Ramsey sentence might demand that the realiser of an intention stands in a certain causal connection to a realiser of the

agent's self-representation. This link to a self-representation of the group would allow the functionalist to argue why the teaching committee forms one intention and the public engagement committee the other: There are two self-representations, that is one for each group and each of them only stands in an appropriate relation to one of the intention-realisers. In the case of the coinciding group agents, the connection might be realised through the individual members and how they conceive of their deliberation. In the case of individual agents, the realiser of such a self-representation would not involve any members, but some neural state. Demanding such a causal connection would be on the right level of description: It concerns internal states but allows for group as well as individual agency.

Of course, the functionalist would have to justify such a requirement independently of the presented problem case so as not to appear terribly ad-hoc. My proposal involving self-representation is just one amongst many options which are opened up by allowing ourselves to draw on the internal states of the agents. To provide all the details of the functionalist solution, we would need to defend a general functionalist theory covering all kinds of agents. Such an effort goes beyond the scope of the present paper and is not needed.

I neither claim to provide nor have to provide the fully specified Ramsey sentence for propositional attitudes. The inclusion of internal states allows to account for the attribution of intention in our counterexample in one way or another. There is plenty of room to find responses to the case of coinciding groups as soon as we look into the internal lives of group agents. The availability of such a solution renders functionalism the more attractive theory of group agency. In other words, while functionalist proponents of group agency face the challenge of providing the correct Ramsey sentence, interpretivists close the door to a solution.

11 Conclusion

We considered an example of group agency where the internal states of how the members realise the groups' mental lives make all the difference, while the outer-behaviour, realised and dispositional, remains indistinguishable. Interpretivist accounts cannot deal with these cases since they hold that "the performance [of the system] itself should dictate the representations and motivations we ascribe to the agent" (List and Pettit 2011: 28). For the case of coinciding group agents, these interpretivist approaches fail since the behavioural performances of the groups do not allow to explain a difference in attributing propositional attitudes. A functionalism which looks beyond the performance of the system and considers how internal mechanisms realise it can solve the problem.

Funding Funding was provided by Arts and Humanities Research Council (Grant No. AH/L503848/1).

References

Block, N. (1978). Troubles with functionalism. *Minnesota Studies in the Philosophy of Science, 9,* 261–325.

Block, N. (1981). Psychologism and behaviorism. *Philosophical Review, 90*(1), 5–43.

Block, N., & Fodor, J. A. (1972). What psychological states are not. *Philosophical Review, 81*(April), 159–181.

Brooks, D. H. M. (1986). Group minds. *Australasian Journal of Philosophy, 64*(4), 456–470.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*(1), 7–19.

Dennett, D. C. (1987). *The intentional stance*. Cambridge, Massachusetts: MIT Press.

Dennett, D. C. (1991a). *Consciousness explained*. London: Penguin Books.

Dennett, D. C. (1991b). Real patterns. *Journal of Philosophy, 88*(1), 27–51.

Epstein, B. (2015). *The ant trap: Rebuilding the foundations of the social sciences*. New York: Oxford University Press.

Epstein, B. (2017). What are social groups? Their metaphysics and how to classify them. *Synthese*. https://doi.org/10.1007/s11229-017-1387-y.

Fodor, J. A. (1968). *Psychological explanation: An introduction to the philosophy of psychology*. New York: Random House.

Funkhouser, E. (2007). Multiple realizability. *Philosophy Compass, 2*(2), 303–315.

Gilbert, M. (1992). *On social facts*. Princeton, NJ: Princeton University Press.

Huebner, B. (2014). *Macrocognition: A theory of distributed minds and collective intentionality*. New York: Oxford University Press.

Lewis, D. (1966). An argument for the identity theory. *Journal of Philosophy, 63*(1), 17–25.

Lewis, D. (1970). How to define theoretical terms. *The Journal of Philosophy, 67*(13), 427–446.

List, C., & Pettit, P. (2011). *Group agency: The possibility, design and status of corporate agents*. Oxford: Oxford University Press.

Lycan, W. G. (1987). *Consciousness*. Cambridge, MA: MIT Press.

Millikan, R. G. (2004). *Varieties of meaning: The 2002 Jean Nicod lectures*. Cambridge, MA: MIT Press.

Mölder, B. (2010). *Mind ascribed: An elaboration and defence of interpretivism*. Amsterdam: John Benjamins.

Peacocke, C. (1983). *Sense and content: Experience, thought, and their relations*. Oxford: Oxford University Press.

Polger, T. W., & Shapiro, L. A. (2016). *The multiple realization book*. Oxford: Oxford University Press.

Putnam, H. (1975). *Mind, language, and reality*. Cambridge: Cambridge University Press.

Ritchie, K. (2013). What are groups? *Philosophical Studies, 166*(2), 257–272.

Ritchie, K. (2015). The metaphysics of social groups. *Philosophy Compass, 10*(5), 310–321.

Sheehy, P. (2006). Sharing space: The synchronic identity of social groups. *Philosophy of the Social Sciences, 36*(2), 131–148.

Shoemaker, S. (2003). *Identity, cause, and mind: Philosophical essays*. Oxford: Oxford University Press.

Thomasson, A. L. (2016). The ontology of social groups. *Synthese*. https://doi.org/10.1007/s11229-016-1185-y.

Tollefsen, D. (2006). From extended mind to collective mind. *Cognitive Systems Research, 7*(2–3), 140–150.

Tollefsen, D. (2015). *Groups as agents*. Cambridge: Polity.

Uzquiano, G. (2004). The supreme court and the supreme court justices: A metaphysical puzzle. *Noûs, 38*(1), 135–153.