Synthese (2021) 199:3367–3403
https://doi.org/10.1007/s11229-020-02938-z

Hybrid collective intentionality

Thomas Brouwer[1] · **Roberta Ferrario**[2] · **Daniele Porello**[2]

Received: 21 June 2019 / Accepted: 26 October 2020 / Published online: 22 November 2020

Abstract
The theory of collective agency and intentionality is a flourishing field of research, and our understanding of these phenomena has arguably increased greatly in recent years. Extant theories, however, are still ill-equipped to explain certain aspects of collective intentionality. In this article we draw attention to two such underappreciated (and intertwined) aspects: the failure of the intentional states of collectives to supervene on the intentional states of their members, and the role of non-human factors in collective agency and intentionality ('hybrid' collective intentionality). We propose a theory of collective intentionality which builds on the 'interpretationist' tradition in metasemantics and the philosophy of mind as initiated by David Lewis and recently developed further by Robbie Williams. The collective-level analogue of interpretationism turns out to look different in some ways from the individual-level theory, but is well-suited to accommodating phenomena such as hybrid collective intentionality. Complemented with Kit Fine's theory of variable embodiment, such a theory also provides a diachronic account of intentional collectives.

Keywords Collective intentionality · Interpretationism · Variable embodiment · Structure

1 Introduction

It is not uncommon in everyday thought and talk to ascribe intentionality, and more specifically propositional attitudes, to collectives. While some have downplayed the significance of such attributions, regarding it as loose talk (e.g. Quinton 1975) it has become gradually more common among philosophers to take such ascriptions seriously, treating collective agents as genuine entities which genuinely have propositional attitudes. That is not to say that they are inclined to put the intentionality that collectives

✉ Thomas Brouwer
tnpabrouwer@gmail.com

1 School of Philosophy, Religion and History of Science, University of Leeds, Leeds, UK

2 Laboratory for Applied Ontology, ISTC-CNR, Trento, Italy

exhibit on a par with that exhibited by individuals. The default approach is to treat the intentional states of collectives as deriving in some manner from the intentional states that their members enjoy: collective intentionality is a secondary kind of intentionality built upon the primary intentionality of individuals. This is not an implausible thought–if the intentionality of collectives doesn't derive from the intentionality of individual members, then where does it come from? However, we will argue that this thought is often developed into an approach to collective intentionality which is too narrow to accommodate the full variety of collective intentionality phenomena. We must take a somewhat broader and somewhat more subtle view of how collectives come to exhibit intentionality. In this paper we set out such a view.

We will start by explaining the type of model of collective intentionality that we are hoping to improve upon (Sect. 2.1). We then go through some cases that illustrate how the attitudes of collectives do not depend in a straightforward way on the attitudes of members of those collectives (Sect. 2.2), and draw attention to some of the additional factors that are at work in determining these attitudes, factors which an adequate theory should accommodate (Sect. 2.3). In particular, we draw attention to the phenomenon of *hybrid* collective intentionality, that is, collective intentionality that comes about as the result of interaction between human agents and artificial entities which may or may not be agents (Sect. 2.4). It has so far drawn little attention in the collective intentionality literature, despite its considerable real-world significance. In Sect. 3 of the paper we set out our positive view. After a clarification of the terms used to present the approach (Sect. 3.1), we begin by articulating an *interpretationist* approach to collective intentionality. This type of theory of content has been developed at a fairly sophisticated level for the case of individual intentionality (Sect. 3.2), and (we argue) gives us the tools we need for modelling collective intentionality with the requisite degree of flexibility and generality (Sect. 3.3). An adequate model of collective intentionality needs to pay particular attention to the *structure* of collectives (or so we argue), so we spell out the notion of structure that is involved and articulate it using the theory of *variable embodiment* (Sect. 3.4).

2 Some limitations of current approaches

We can ascribe intentional states to collectives in a number of ways. Christian List (2014) helpfully classifies these into ascriptions of aggregate, common and corporate attitudes:

- When we ascribe *aggregate attitudes* to a collective, we use some kind of aggregation rule (e.g. majority opinion, largest plurality opinion, average estimate) to summarise the attitudes of members of the collective. The aggregate attitudes of a collective are thus relative to what rule we choose, and what rule is appropriate in a given case depends on our theoretical or practical interests.
- The notion of a *common attitude* is a generalisation of the familiar notion of *common knowledge*. A collective can be said to have such an attitude if (a) all the members of the collective have it and (b) they all believe of each other that they have it, they all believe of each other that they believe of each other that

they have it, and so on and so forth.[1] Common beliefs play an important causal-explanatory role in explaining how groups function and how individuals function within groups.[2]

– *Corporate attitudes* are the kind of attitudes that we ascribe to collectives when we treat those collectives as agents in their own right, rather than just as collections of individual agents. If a collective displays behaviour that is unified enough to be aptly described as the behaviour of a single agent, then it may make sense to enquire what that agent believes, desires, etc. Those attitudes would be the corporate attitudes of the collective.

In this paper we focus on corporate attitudes. Compared to aggregate attitudes and common attitudes, corporate attitudes give rise to some quite substantial metaphysical questions. What makes it correct to ascribe attitudes of the former kinds to some collective is just that the individuals in that collective instantiate some distribution of attitudes: they arguably *just are* certain kinds of distributions of attitudes across individuals. But there is an entirely non-trivial question about what facts underlie and make true the ascription of corporate attitudes to collectives. It seems very plausible that there is *some* important relation between the attitudes of a corporate agent and the attitudes of its members, but it is not transparent what that relation is, and it is not a priori that the corporate attitudes of a collective are determined only by the attitudes of its members. In other words: corporate attitudes require a substantive metaphysical explanation of a sort that common and aggregate attitudes do not seem to call for.[3]

There has been a fair bit of attention paid to corporate attitudes in recent years, and it is clear now that they are too complicated to be simply identified with the result of some aggregation rule or other (as amply documented by List and Pettit 2011). But we contend that there are aspects to corporate attitudes that have yet gone unappreciated, and that overly restrictive assumptions have been made about their nature. To make clear why these assumptions should be contentious, we will start off with a general look at how collective attitudes fit into the world at large. From here on in, by 'collective attitudes' we mean corporate attitudes, unless otherwise indicated.

[1] This, at least, is one way of explaining what it is for a group to have a common attitude. List also discusses what he considers to be rival understandings of the same general phenomenon, e.g. Searle's notion of we-attitudes (Searle 1995) and Gilbert's notion of plural subjects (Gilbert 1989).

[2] Common attitudes are arguably just one instance of a wider class of collective attitudes, that of attitudes which play robust causal-explanatory roles in group behaviour but which do not necessarily amount to corporate attitudes. In this class we might place for instance pooled knowledge, and also perhaps joint intentions, as understood by e.g. Bratman.

[3] That's not to say that they don't give rise to theoretical challenges. With regard to attitudes like joint acceptance and joint intention, there is long-running debate concerning the kind of individual-level attitudes that are involved, and how they must be interrelated, in which Bratman (1999), Gilbert (1989), Tuomela (1995) and Searle (1995) are central figures. Do joint attitudes involve special psychological capacities (special attitude-types) unlike those used in individual thought? Do they involve familiar attitude-types directed towards special contents? Do joint attitudes intrinsically give rise to normative constraints? Aggregate attitudes give rise to various technical challenges, since many ways of combining individual attitudes may result in inconsistencies, Condorcet cycles and the like. These difficulties are closely related to impossibility results such as Arrow's Theorem (Arrow 1951), and are studied using the tools of social choice and judgement aggregation theory. List and Pettit (2011) discusses these issues in detail; see also List and Puppe (2009), Dietrich and List (2010) and Porello (2018).

2.1 How collective intentionality gets into the world

Pretty much everyone agrees–and we do too–that collective intentionality is not a metaphysically *basic* phenomenon; wherever it rears its head, we ought to be able to point to some underlying facts which constitute that phenomenon. Even theorists that class themselves as (in some sense) anti-reductionists tend to subscribe to this plain, ontological kind of reducibility, which does not of itself imply any kind of theoretical, epistemic or definitional reducibility for the phenomena in question. To deny that collective intentionality is metaphysically non-basic in this way would be to assert that collective attitudes are among the basic constituents of the universe, a view few would incline towards.

Though this is not a controversial claim it does constrain how we theorise about the phenomenon, so it is worth articulating this assumption a bit, using the tools of contemporary metaphysics. Assume there is some collective c which has a set of propositional attitudes. Take those propositional attitudes to be captured by a set of facts A–to know those facts is to know everything about the attitudes.

Because these A-facts are non-basic, there is some further set of facts, which for now we'll blandly call "B", which consists of all and only those facts which together metaphysically determine the A-facts. That means that (i) the A-facts are a function of the B-facts (or *supervene* on them, to use another term), but not only that: (ii) the A-facts are what they are *because of* or *in virtue of* the B-facts.[4] Clauses like (i) and (ii) can be expressed in metaphysical lingo by saying that the B-facts *ground* the A-facts, and the A-facts *are grounded in* the B-facts.[5]

So facts about the attitudes of collectives are grounded in other facts. This is a very unspecific assumption, and while there is debate about how to construe the notion of grounding, the claim itself shouldn't come across as contentious.[6] It does immediately generate a set of questions the answers to which will be contentious. First, with regard to the nature of the A-facts: what attitudes should we be willing to ascribe to collectives? For instance, is a collective the sort of entity that can have credences, or could we only ascribe outright beliefs (cf. List and Pettit 2011, p. 37) or maybe just 'acceptances' Tuomela (2000), Wray (2001)? What we say to this question has knock-on effects for the metaphysics, for the 'richer' the attitudes we ascribe (in terms of type and content) the richer the base of grounding facts has to be. Second, with regard to the determining relation which ties those A-facts to lower-level facts: what is this relationship? Something more needs to be said than just that the latter determine/ground the former, for they could in principle determine them in many ways (functions are cheap) and we need some substantive story that explains why we get *these* A-facts out of *those*

[4] The qualification is necessary because the supervenience relation (or the relation of being-a-function-of) can be, and often is, symmetric, whereas the relation of metaphysical explanation we're trying to get at here is clearly directional and asymmetric. The extra clause expresses this.

[5] Which is not yet to buy into any particular theory of what grounding amounts to and how it behaves. For our purposes we don't need to venture any opinions on that issue.

[6] Which doesn't mean it can't be doubted at all. A theorist could conceivably hold that, while the A-facts are in some sense metaphysically posterior to some set of B-facts, the B-facts nevertheless do not supervene on the A-facts. This would amount to an 'emergentist' view about collective intentionality. Like many, we do not find it attractive: it raises more questions than it answers.

B-facts. Third, with regard to the B-facts: what sorts of facts do these include? Will they be facts about the attitudes of the members of the collective in question, or facts about their behaviour, or both, or could they include facts about other things entirely? By being liberal about what we include here, we help ourselves to more resources to explain the A-facts we encounter. But a more heterogeneous grounding base also makes it more difficult to give an elegant story about the determining relation between B- and A-facts.

By framing the enquiry into collective intentionality as a metaphysical question about what grounds what and how, we get a sense of the questions that need answering and a conception of the logical space that is there for us to explore. That logical space is currently underexplored. Let's take the question about B-facts: what sorts of facts should we expect to ground facts about collective intentionality? One straightforward and *prima facie* tempting picture of the metaphysics of collective attitudes has it that B-facts are all of a specific sort: for a collective c, the B-facts are all facts about the attitudes of the members of c. For example, if c is a bowling team (we'll call them the Sunday Afternoon Champions) that believes, collectively, that it can win the trophy, then the fact that c has this belief is determined by the distributions of beliefs (and, possibly, desires and other attitudes) that Alice, Bob, Charlie and Delilah, the members of c, have. This is an elegant picture, and for many cases entirely plausible. It comes as no surprise, then, that it often plays the role of a background assumption in theories of collective intentionality, and in some cases, an explicit thesis. Here, for instance, we have List and Pettit's supervenience thesis: 'The attitudes and actions of a group agent supervene on the contributions of its members.' (List and Pettit 2011, p. 66)

List and Pettit are willing to countenance in principle that the 'contributions' of group members might embrace somewhat more than just their attitudes (e.g. their actions), but they downplay the potential significance of such non-attitudinal contributions, and try to make do without them in the theory they give. They also allow that the organisational structure of a collective can make a difference to how the group's attitudes are determined, but they consider this structure to itself supervene on member attitudes, so it does not constitute a failure of supervenience.

List and Pettit are not outliers in treating the attitudes of group members as the grounding base for collective intentionality. We'll argue that the facts that ground facts about collective intentionality plausibly embrace more than this. We also think that those facts can be rather heterogeneous in type. In the following two subsections, we will discuss some cases to support this view.

2.2 Collective intentional states are not grounded only in intentional states of members

Brian Epstein (2015) discusses cases which, he argues, show that facts about group intentionality can depend on facts other than those about the intentional states of group members.

– *MBTA* The Massachusetts Bay Transport Authority (MBTA) board is empowered to make decisions about subway fares, but only when gavelled into session by the board's 'parliamentarian', who is not themselves a member of the board. Depend-

ing on whether the parliamentarian has in fact gavelled the board into session–a fact the board members might be unaware of–the MBTA board members' consensus on whether to raise subway fares does or does not result in the MBTA board making a decision.

– *Shareholders* A group of shareholders makes decisions by weighted majority voting, the weights being determined by the number of shares that a shareholder possesses. Here, Epstein contends, the group's decisions are not a function solely of the shareholders' decisions, but of the shareholders' opinions combined with facts about how much shares they own.

These cases make trouble for supervenience in two ways. First, they aim to show that there are factors aside from the collective's members' attitudes that are making a difference to the collective's attitudes. In the Shareholders case, these are facts about the share distribution, in the MBTA case, these are facts about the parliamentarian's doings. Second, in the MBTA case (but not the Shareholders case) the additional factor is external to the collective: it seems to be an outside influence that makes a constitutive difference to the collective's decisions. When it comes to the first kind of supervenience failure, List and Pettit have a response in mind:

> The supervenience thesis does not sideline the role of the group agent's organizational structure. It is perfectly possible that the members of two different group agents individually have exactly the same intentional attitudes on some propositions, while the two group agents, due to their different organizational structures, hold different attitudes and act differently. But the difference in the two organizational structures will show up in some individual-level differences between the two groups; their different forms of organization mean that their members will act and be disposed to act in different ways. If one group is democratically organized while the other is dictatorial, for example, this difference will show up in different dispositions among members with regard to counting votes. (List and Pettit 2011, p. 66)

One could apply this idea to explain Epstein's Shareholders case. One could say that distribution of shares over the group members gives rise to a certain vote-counting behavioural disposition among them. This fixes a group structure whereby the group's opinions are aggregated in a way which indirectly depends on the distribution of shares. The decisions of the shareholder group can then still be seen to supervene on the opinions and behavioural dispositions of the individual shareholders, as long as we include in the supervenience base not only the members' opinions on the matter under vote, but also their behavioural dispositions relating to the share distribution.

But there are limitations to this response. For this to be a general form of response to Epstein-like cases, it would always have to be plausible that the factor which determines how members' opinions are aggregated itself supervenes on features of individual group members. Already in the MBTA case it is not clear that one can say this. The parliamentarian's behaviour is clearly making a difference to how the board's opinions determine a decision, so their activities would have to show up somehow in a description of the group's structure. But is that because of individual features of board members? It doesn't seem so. And this is not uncommon: the members of groups

frequently don't have the authority to decide facts about the group structure: a football team's captain, for example, is appointed by the coach, not by the members.

So it is doubtful that there is sufficient mileage in this suggestion alone. One could bolster it by adjusting one's account of what it takes to be a member of a group agent: one could propose to count any individual whose attitudes or behaviour form part of the supervenience basis for that group's attitudes as *ipso facto* a group member. Then the MBTA board's parliamentarian turns out, contrary to initial appearances, to be a board member, and supervenience appears to be saved. This suggestion bears a cost, as it requires us to revise things we thought we knew about group membership. It seemed antecedently very plausible, for instance, that whether or not the parliamentarian is a member of the MBTA board is something the MBTA, as an organisation, can simply stipulate. But apparently it cannot![7]

We think that the Epstein cases are best taken at face value, as counterexamples to the supervenience thesis. Cases like these are not isolated. Individuals outside of a collective can contribute something to the attitudes of that collective in various ways, and such contributions can go beyond affecting the process by which the attitudes are determined from the attitudes of the collective's members (as in the MBTA case): entire attitudes can be acquired from the outside. Consider a company that retains a legal firm to react on its behalf to any legal troubles it may encounter. Say some trouble arises and the legal firm takes a stance on the company's behalf; we would want to say that this stance is the company's stance. It's quite possible in such a case that no individual or sub-collective within the company has the capacity to understand the stance in question, if the legalities involved are complicated enough. The company might not even make an effort to register or understand the stance taken. So there is plausibly no internal state of the collective that could be subvening this attitude; but it is an attitude of the company.[8]

Cases like Epstein's are one good reason to look for a more liberal metaphysics of collective intentionality. Another good reason is the phenomenon of hybrid collective intentionality.

2.3 Collective intentionality does not only involve human beings

There is another distinct (but not unrelated) way in which collectives are often considered too narrowly: collectives don't always just consist of people. This may seem like an odd thing to say at first–aren't they groups of people by definition?–but examples will hopefully make apparent that it's not that odd. To again borrow an example from Epstein (2015), take Starbucks, the coffee chain. We're generally willing to ascribe attitudes to an entity like Starbucks, e.g. we say stuff like 'Starbucks wants to move in on the third-wave coffee market' or 'Starbucks misunderstands Australia's

[7] By way of damage limitation one could perhaps distinguish between the collective agent called the MBTA board and the institution called the MBTA board, and say that the latter, but not the former, can have its membership determined by stipulation. But that is unlovely: part of the aim of the theory of collective intentionality is to make sense of the idea that institutions can be agents, so we shouldn't let them come apart so easily.

[8] These are usually referred to as cases of *proxy action*. See Ludwig (2014) for a general discussion of the phenomenon.

coffee culture'. But what does, Starbucks, the entity, consist of? A common-sense view would be that Starbucks is made up of cafés, coffee machines, baristas, refrigerators, corporate offices, managers, company cars and all manner of other entities. It contains people but not just people.

This is not conclusive, for someone might agree that there is some entity which has both baristas and coffee machines as parts, but resist the idea that it's identical to the entity that we're attributing attitudes to. One might think there are two entities, both confusingly named Starbucks in common discourse. One of them (call it Starbucks$^-$) has collective attitudes and only people as components and the other (call it Starbucks$^+$) has other types of components as well, but no collective attitudes. Only the former is a collective, strictly speaking.

It isn't simply Ockham's razor that pushes us to identify Starbucks$^-$ and Starbucks$^+$ as a single entity that has both collective attitudes and non-human parts. It's explanatory power. Starbucks$^+$ is an integrated, functional whole, which clearly operates as a functional unit within a wider system (the coffee retail market). In explaining why Starbucks does better than one of its competitors, one may cite in one breath the fact that its employees make coffee more quickly (let's just say) and that its coffee machines are better (let's just say). That the entity Starbucks$^-$ can play a similarly useful explanatory role is a much more doubtful proposition. But we can go further: even when it comes to the specific case of explaining Starbucks' collective attitudes, it's more natural to let the non-human parts be involved as well as the human parts.

Consider this slightly science-fictional scenario, whereby Starbucks gradually replaces more and more parts of itself with artificial components. Assume, at the start, that all of Starbucks' collective attitudes have explanations which appeal only to the human parts of Starbucks. For instance, let's say Starbucks believes that the time is ripe to expand into the Namibian coffee market, and that it has this belief simply in virtue of the fact that the head of Starbucks' Europe, Middle East and Africa Division has this belief. Now imagine that Starbucks develops a new bit of planning software which announces the best places to open new locations, and that Starbucks adopts the policy of doing whatever the software says. Let's say that the new software *de facto* gives all the same verdicts as the relevant humans did before. So Starbucks still believes that the time is ripe to expand into the Namibian coffee market, but now it does so in virtue of the software's say-so. Imagine that Starbucks keeps replacing bits of itself with software programmes, robotic arms, self-driving vehicles, hard-drives for storage etc., at every point preserving the overall functional profile of the company. At the end we have an entity which has no human parts, but given that it behaves the same, there is no reason to stop attributing to it the attitudes that we were formerly attributing to Starbucks.

Such a case would show that we could replace the human parts of a collective agent with artificial parts and still retain an agent. It's perhaps not so obvious that what remains at the end of the transformation is still a *collective* agent. Perhaps we've just replaced the whole of Starbucks with a complicated robot![9] Settling this requires knowing what makes an agent a collective one. This is not a question often discussed: in the literature on collective agency the salient question is usually whether a given

[9] We thank an anonymous referee for discussion of this point.

collective counts as an agent, rather than whether a given agent counts as a collective. But here is a suggested answer: an agent is a collective one if its agency arises from a system of interacting elements, where at least some of these elements are themselves agents. We use 'element' here, rather than just 'part', to signal that the agents and their (inter)actions would need to actually show up in the explanation of the agency of the whole. There might in principle be composite agents which have parts which are in fact agents, but where the fact that they are agents does no explanatory work. Those would not be collective agents, properly speaking.

With this characterisation in mind, we can see that the Starbucks replacement case is as yet underspecified. There are ways of conducting the wholesale replacement operation which would preserve its collectivity, and ways that would not. It would depend on whether at least some of the artificial parts we swapped into Starbucks would still contribute to the overall agency of the whole *through being agents*. There are certainly ways of telling the Starbucks replacement story on which this wouldn't be the case. But there are also ways of telling it on which it would. And that is all we require to illustrate our point: that human involvement, though typical, is not essential to the phenomenon of collective intentionality.[10]

The Starbucks story is a bit far-fetched, perhaps, but real-world examples of essentially the same process abound, especially these days. In the wake of Uber, taxi companies everywhere are replacing traditional dispatchers with software that does essentially the same job but more quickly and for less money. Decisions that are made within a company and on behalf of a company can be made by artificial means, and the attitudes this gives rise to are genuinely those of the company. Collective intentionality is not an exclusively human phenomenon.

We refer to cases in which the non-human parts of collectives play a role in generating their attitudes and actions as *hybrid collective intentionality*.

2.4 Hybrid collective intentionality

Imagine that we are in an airport. We have passengers walking along the hallways; airline employees sitting at check-in desks; clerks in the shops; security company employees checking the luggage of passengers by means of x-ray machines and metal detectors; policemen, some patrolling the premises of the airport, others checking passenger's passports, yet others sitting in a room and watching the security video camera feeds. This whole complicated system is not plausibly regarded as one collective agent, but parts of it can be. For instance airport security. This is an organisation which employs a large number of people–security guards walking the hallways, watching the video feeds, operating the luggage scanners, patting down the passengers–and also involves a large amount of relatively advanced machinery, e.g. the aforementioned cameras and luggage scanners, metal detectors, and so on. Other, less advanced things are also involved: the security guards' uniforms, weaponry, coffee machines, what have you. Some, but not all of these non-human things are plausibly contributing to

[10] To retain the collective nature of the resultant agent, it would not be necessary that we replace human agents for artificial agents one for one. We could replace multiple human agents with single artificial agents, or vice versa.

the collective's attitudes in fairly direct ways. The video cameras, luggage scanners and metal detectors are, first off, acquirers and bearers of information, which the security guards are in a position to extract and act upon. But they can also be processors of information and decision-makers. For instance, a camera's software may enable it to recognise shapes which it classifies as signs of danger, and raise the alarm as a result. Airport security, as a collective, might spring into action on the say-so of a camera.

We're inclined to view a system like this as a collective which has components which are human beings and components which are not. Both the human beings and the other components do their bit in gathering information, processing it, making decisions on the basis of gathered and processed information, and implementing those decisions. Different components play different roles, and some are more important than others or more authoritative than others, in that they can overrule, reinterpret or aggregate the contributions made by others. We assume no in-principle differences in the kinds of roles that human and non-human components can be accorded in modelling a collective agent. Not all of the collective's components have to be regarded as full-fledged agents (or as agents in any sense) to figure in the model; even a non-agent like a simple motion sensor can bring information into the collective's cognitive processing in much the same way that a human observer could, while lacking the full range of capacities that make the human observer an agent.[11]

That is how we are inclined to view things. Another possible diagnosis would take inspiration from the 'extended mind' tradition: the proposal would be that that the cameras, metal detectors etc. are serving to 'extend' the cognitive activity of the human agents involved. The extended mind hypothesis is not a view we're in principle opposed to, and the view that we'll go on to propose can perhaps be combined with it. But it depends on how we apply the extended mind conceit; in particular, whether we see the artificial components as extending the collective's mind as a whole, or just as extending the minds of the human beings in the collective.[12]

On the latter view, all of the artificial components in the example above that we suggested help generate hybrid collective intentionality get incorporated in the model of the collective agent as sub-personal components of the human agents. For example, the metal detector might be regarded as part of the security guard who is on metal detector duty. This view would not require that every artificial component of the collective be modelled as an extension of some one particular individual in the collective; in principle the same component could be extending the minds of multiple individuals, or even all individuals in the collective. It does, however, require that we be able to understand the contribution these artificial components are making as wholly 'routed through' the human individuals in the collective: they do not contribute on the same level as these individuals, but rather just contribute indirectly, in the way that sub-personal parts of those individuals might. This places a significant constraint on the way that we can model hybrid intentionality, and there are going to be cases where this

[11] Of course, if one's notion of agency happens to be undemanding enough, then even something like a motion sensor might be an agent. We happen to favour a relatively undemanding, functionalist characterisation of agency, but the points we want to make go through even on a more restrictive account, because being an agent is not on our view a precondition for contributing to collective attitudes or collective action.

[12] For a discussion of the connections between collective intentionality and the extended mind thesis, see Tollefsen (2006).

does not look like an apt treatment: notably, the cases where artificial parts of the collective are full-fledged agents (e.g. robots, or autonomous artificial intelligences), but more broadly, cases where artificial components that may or may not be full-fledged agents are playing roles in a collective that we could just as easily see played by a human individual. It seems somewhat chauvinistic to model these as mere extensions of nearby human agents when we presumably would not have done so if their role had been played by an actual human–in the absence, at least, of some further independent reason for treating them differently.

Even if we could find a way to model each and every hybrid intentional collective so that it obeys this constraint, it is a constraint that we would rather not have to put ourselves under, in theorising about hybrid collective intentionality, without some strong motivation. More congenial to us seems a second option, that of seeing the artificial components as extending the mind of the collective as a whole.[13] Here, there would be no built-in difference between the kind of contribution made by the human components of the collective and its artificial components; they would be alike in all being part of the supervenience base of the same 'mind'.[14] Of course, this by itself does not tell us much about how the collective's intentional states arise from this supervenience base, and it does not tell us whether there might be any further, less fundamental differences between the contribution made by the human and non-human components of the collective. For that, we need a different kind of theory, which, fortunately, we intend to provide.

3 An interpretationist proposal

3.1 Clearing the ground for a positive account

An account of collective intentionality that is adequate to the variety and complexity of the phenomenon cannot build in, as an assumption, that collective intentional states are grounded in the intentional states of individual members of collectives, nor should it assume that it is only ever human beings that we have to pay attention to. In what follows, we are going to set out an account that we hope is adequately powerful and flexible. It is an account that is *functionalist* and *interpretationist*, and fits into a tradition of theorising about intentionality that derives from David Lewis (1974), and has recently been further developed by Robbie Williams (2016, 2020).[15]

First, we will explain how we use some important terms. We use the term 'collective' as a an umbrella term for talking about collections of entities, some of which entities

[13] To be clear, we have no objection to treating, where apt, artificial components as extending the minds of human components. We're simply reluctant to see this as the general mechanism by which artificial components figure in collectives.

[14] It is interesting that in the literature on collective intentionality, authors who are quite happy to ascribe intentional states to collectives often (if not always) refrain from speaking of collectives' 'minds'. We feel this reluctance too, but nevertheless it is not obvious to us that there is anything substantially more to ascribing a mind to a thing than ascribing intentional states to it.

[15] There is sometimes disagreement about whether Lewis' proposal is aptly described as functionalist. We say more below about how we understand the term.

are agents. We reserve the term 'group' for collections of only human agents, as seems standard in the literature. Collectives can be just groups, or they can be a mix of human and artificial components–we'll call those 'hybrid collectives'–or they can be entirely artificial. We will call those collectives that qualify as agents 'collective agents'. We will call components of collectives that are agents 'agentive components' (or sometimes 'agentive parts', for variation) and others 'non-agentive components'. In what follows we seek to develop a theory that is adequate for understanding hybrid collective agents which have both agentive and non-agentive components. Since these are about as complicated as collective agents can get, they constitute a good explanatory benchmark for a theory of collective intentionality.

The concept of agency also requires characterisation. It is not only a contested notion within philosophy, but also deployed in rather different ways in different disciplines outside of philosophy. In sociology, for instance, agency is often strongly connected with the independence of the subjects in making their choices, while in computer science the notion is often used to characterise artificially intelligent entities, in conjunction with notions such as autonomy, proactiveness and self-direction. For the sake of this paper, we will assume a quite general, relatively undemanding notion, with the aim of preserving the broadest possible common ground. By some readers' lights, not everything we refer to as agency in what follows will count as full-fledged agency, and those readers are encouraged to relabel it however they see fit.

We will use a functionalist characterisation of agency–anything that walks like an agent and quacks like an agent will be an agent by our lights. Specifically, for us an agent is an entity that responds to evidence in ways that are (more or less) rational in the light of its desires/aims/preferences. Thus it has to be possible to ascribe belief- and desire-like representational states to it, even if only of a very coarse-grained, inflexible sort, to view it as making choices out of a space of options–for that is how we model action–and there has to be some coherence (by the lights of one's favoured theory of rationality) between these elements. That's it–we do not assume that it has to have anything like self-awareness, consciousness, or qualia. It need not be free, or enjoy any substantive kind of autonomy.

Finally, we use 'collective action' to refer to actions taken by collective agents. We use 'joint action' to refer to any case in which pluralities of agents act together, without necessarily thereby constituting a collective agent. We assume that collective actions can consist in joint actions so understood, but we don't assume that they have to. Conversely, we do not assume that any joint action is a collective action.

3.2 Introducing interpretationism

Interpretationist accounts of the content of language and thought try to explain how such contents are determined by employing the philosophical device of *radical interpretation*. In radical interpretation, we imagine an idealised interpreter who has access to all the basic facts about the interpretee and their situation–all the facts that don't presuppose the target facts. They know the physical facts about them, they know about their behaviour (including the sounds and signs they make) and they know about the

environment they're in. But when it comes to the contents of their attitudes, the interpreter starts from zero (hence the term 'radical', in the older sense of 'from the roots').

The interpreter is idealised in certain respects: they are omniscient in the above-mentioned spheres, and their ability to process information is unlimited. Their task is to ascertain the contents of the interpretee's attitudes. We then ask some questions. What information would they need in order to fulfil their tasks? What methodological principles would they need to presuppose (e.g. principles of *charity* or *humanity*)? What steps would they need to go through to arrive at their conclusions? We then use the resulting description of the ideal interpreter's activities to understand how facts about content are determined by facts that underlie them. Thus interpretationism can be viewed as a pseudo-epistemological account of the metaphysics of meaning.

The interpretationism we're describing is a style of theory that derives from the work of David Lewis (1974), and has recently been further developed by Robbie Williams (2016, 2020), upon whose version we shall build. It differs in important respects from other interpretationist theories, like those of Donald Davidson and Daniel Dennett. Dennett's interpretationism in particular has been used by Deborah Tollefsen (2002) as the basis for a theory of collective intentionality. So let's note the differences between the styles of theory which explain why our theory ends up looking quite different from Tollefsen's. Here Tollefsen characterises the thinking behind Dennett-style interpretationism:

> Interpretationism is the view that if an agent is interpretable, the agent is an intentional agent. [...] It is an approach to intentionality that starts not with metaphysical speculations about the nature of the mental, but with our practice of attributing intentional states. It asks, what are the constitutive features of our practice that guarantee its explanatory power? That is, what assumptions do we need to make about an agent in order to interpret her behavior successfully? Approaching intentionality from this third-person perspective allows us to avoid having to speculate about what beliefs are or the conditions under which one can be said to believe a certain proposition. If interpretation is successful, then the assumptions we make about the agent are justified. (Tollefsen 2002, p. 397)

Lewis/Williams interpretationism differs in two ways. First, for Tollefsen and Dennett it matters that there actually be interpreters, with some established practice of interpreting. The Lewisian interpreter merely stands in for an abstract interpretation *function* which in turn represents a determining relation holding between a lower-level set of facts about the world and a higher-level set of facts about the world. Second, it does not aim at avoiding metaphysics. It happily makes claims about how facts about intentional states obtain in virtue of lower-level facts about e.g. brain states, and appeals to an idealised procedure of interpretation to explicate how those lower-level physical states come to play the role of representational states. This is meant to yield a robust metaphysics of intentionality, the sort of metaphysics that Dennett and Tollefsen seek to side-step.

The second point is crucial. With regard to the goals we've set ourselves–accounting for ways in which collective intentionality fails to be grounded in individual intentionality, and accounting for hybrid intentionality–there is a certain sense in which

Tollefsen's form of collective interpretationism scores very well, in that nothing in her theory excludes either possibility. That is for a specific reason: unlike the theories we have discussed above (e.g. List and Pettit's), her metaphysically quietist form of interpretationism makes no claims at all about what grounds collective attitudes or about the way that non-human elements might figure. But the metaphysical nitty-gritty that Tollefsen seeks to avoid is precisely what we are interested in finding out about–to side-step the metaphysics would be to give up the game. Hence we prefer the more explanatorily ambitious kind of interpretationism that Lewis and Williams offer.[16]

To explain our proposal for a collective interpretationism it will help to briefly set out how the view is meant to work in the individual case. Let's say (borrowing from Lewis) that our interpreter is interpreting an individual named Karl. Let their task be understood as giving an entire interpretation of Karl, that is, to assign him a full suite of beliefs and desires (and, perhaps, other contentful attitudes, if those cannot be defined up in terms of his beliefs and desires).[17] Such interpretations can be understood as *theories of Karl's mind*–psychologies of Karl, if you will, which serve the purpose of explaining how Karl's behaviour makes sense in the light of the situation he's in.[18]

We make the following methodological assumption to get us going: the *correct* interpretation of Karl–the one that assigns him the attitudes he really has–is to be identified with the *best* interpretation a fully informed ideal interpreter could devise. Thus we need to know how to rank interpretations as better or worse. Since we're understanding interpretations as a kind of *theory*, we know roughly what kinds of criteria to apply: criteria of *theory selection*, as familiar from the philosophy of science. Some of these criteria are generic ones, e.g. things in the order of simplicity, conservativeness, or unification.[19] Some of them will be specific to the particular type of theory we are dealing with here. A criterion specifically relevant to this type of theory is that it ought to make an agent come out as rational: our theory of Karl ought to describe him as acting in such a way that he is responding to his evidence in reasonable ways in the light of his desires and background beliefs.

One may wonder why rationality comes into it; making things come out as rational is not ordinarily found on lists of theoretical virtues. Well, just as with any theory, we want the theory of Karl to have explanatory power with respect to the phenomenon it targets (Karl's behaviour). And we are after a specific type of explanation: one that sheds light on Karl's *reasons* for acting as he does–as opposed to, say, the physical causes of his behaviour, for which we look to e.g. a biological, neurological or chemical

[16] Another approach in the interpretationist ballpark is Huebner (2014). Though it too takes inspiration from Dennett, it ends up putting more emphasis than Dennett or Tollefsen do on the ways that cognitive systems are concretely realised by causal mechanisms, in demarcating the phenomenon of intentionality. In that respect, it is a little closer to the more metaphysically heavy-duty Lewisian approach we develop. Huebner's account is, however, different from ours in a number of respects which we do not have the space to discuss here.

[17] Or, depending on your version of interpretationism, the beliefs and desires could be assigned to temporal stages of Karl, or to states of Karl's brain. There are delicate theoretical choices here, but we won't need to go into them.

[18] This style of interpretationism typically aims for a combined account of both mental content and linguistic meaning. We will concentrate only on the case of mental content.

[19] Or whatever we take the list of genuine theoretical virtues to include. It can be left open for present purposes.

theory of Karl. Thus, a theory of Karl's attitudes which makes him come out as more rational does a better job of *explaining* him, in the specific way that we want him explained, and to that extent counts as a better theory.[20]

What is rationality, for these purposes? Williams (2016) suggests that we start by understanding rationality in terms of Bayesian decision theory: perfectly rational agents make choices that maximise expected utility given their beliefs (which we treat as a probability function) and desires (which we treat as a utility function), and they update on evidence in the ways prescribed by the Bayesian framework.[21]

Given certain preferences on Karl's part, and certain beliefs about the world, the decision theory allows us to predict that Karl, if rational, will take certain actions if he receives certain evidence. Knowing his actual evidence and actions, and on the assumption of rationality, we can reverse-engineer his beliefs and desires.[22] This, then, allows us to determine the interpretation of Karl that makes him most rational, which is *ceteris paribus* the best interpretation of Karl.[23]

This proposal invokes Karl's *evidence* and *actions*. We attribute beliefs and desires to Karl by rationalising his actions given his evidence, using our Bayesian apparatus. In order to do this, the evidence and actions must be characterised in a certain way: his actions need to be framed as *choices* from sets of options, and his evidence needs to be framed propositionally, as items Karl can update on. In other words, we need an intentional characterisation of Karl's actions and evidence.

On the one hand, this shows a limitation of interpretationism. It does not give us a complete characterisation of intentional states of Karl in terms of a purely physicalistic description of Karl's behaviour and environment. Rather, it takes as given certain 'low-level' intentional states of his (his evidence and choices, or at least a certain range of them) and, using the device of interpretation, fills out the rest of Karl's contentful states. If we want a theory that fully characterises intentionality in terms of the non-intentional, we'd need to add an independent, prior theory that tells us how to attribute a basic set of evidence propositions and choices to an agent. We'll not go into this, as

[20] For all that's been said, we may not always get the result that there is a single best candidate interpretation of Karl. If multiple interpretations end up tied for first place, Karl's attitudes would be indeterminate where those interpretations disagree. Sometimes that's a plausible result: most likely not all of Karl's attitudes have determinate content, and we would want to capture that. If the theory ascribes more indeterminacy than is plausible, that would suggest a deficiency in the theory. We may need to rethink what evidence we consider available to the interpreter, or go looking for additional or more fine-grained criteria.

[21] Williams ultimately opts for a notion of rationality that's richer than this. But the considerations that motivate this are not particularly relevant for present purposes, so we will stick with the simpler view.

[22] In so doing we would need to invoke decision-theoretic representation theorems, which tell us (for a given decision theory) that we can derive a unique set of preferences and credences, given a specification of an agent's choices and evidence.

[23] Though the winning interpretation is the one that all else equal does best at making Karl rational, this does not mean that we should expect it to make him a perfectly rational agent (which would also seem like the wrong theoretical result). Rather, it's one that doesn't ascribe to Karl any gratuitous failures of rationality. Even when Karl's behaviour is strange and seemingly incoherent, *some* assignment of beliefs and desires would make those actions into rational responses to his evidence. But it won't necessarily be the best interpretation all things considered. It might for instance involve attributing to Karl a set of rather disjunctive, gerrymandered concepts, or it might involve attributing some psychologically implausible final desires. The *ceteris paribus* clause is there for a reason. (Additional work may be done via an enriched notion of rationality–see Williams (2020).)

it goes beyond our purposes, but we direct the interested reader to Williams (2020), which tries to execute this additional reduction by means of a teleosemantic theory that builds on Neander (2017).

On the other hand, it makes sense that we'd want to characterise Karl's evidence and actions in a fairly fine-grained way, rather than just give a physical description of his environment and his movements. Karl, given the limited sort of creature he is, can only take up a limited amount of information from his environment, and it only makes sense to rationalise his behaviour in terms of the information he is taking up, rather than the information that is in some general sense there to be had. There may be a certain precise number of blades of grass in the field that Karl is standing in, and that information may be in some broad sense available to him, but it's nevertheless not reasonable to treat this as evidence that's informing his behaviour. And only a subset of his behaviours are the sorts of things that call out for any kind of rationalisation—the things we'd properly call his actions, rather than just all the activity that his body engages in.

By saying that the ideal interpreter determines Karl's beliefs and desires on the basis of Karl's evidence and action, the interpretationist should not be taken to imply that the former states are ontologically grounded in the latter states. All of these states of Karl—beliefs, desires, perceptions, action-intentions—are grounded in physical states of Karl (many would say brain states of his, but we don't have to commit to that), and are thus on the same level ontologically. The relations between them are not constitutive but causal (what Karl perceives causally affects what he believes, which in conjunction with his desires causally affects what he does) and informational (knowing what he perceives and does tells us about what he believes and wants).

As we mentioned above, the Lewis/Williams ideal interpreter (as opposed to the Dennett/Tollefsen interpreter) is meant to be in possession of the full physical facts about Karl, including things like his brain states. Thus the interpreter doesn't just know Karl's perceptions and action-intentions, but also what physical states subvene these states, and what other physical states of Karl causally mediate between these. Having interpreted Karl as having certain beliefs and desires, they will then also be in a position to make a theoretical identification between certain belief- and desire-states of Karl and further physical states of his.

The import of this is that we should take care to track the difference between two sorts of determination relations involved in the interpretationist approach. There is an informational relation between facts about evidence and actions, and facts about beliefs and desires. This has no generally accepted name, but let's call it the 'fixing' relation. And there is an ontological relation between facts about beliefs and desires and some type of physical facts (e.g. facts about brain states) that subvene and constitute these. Along both directions the interpretationist (if successful) might be said to be effecting some kind of reduction. But they are distinct explanatory connections. This is worth clarifying, as we'll want to be in a position later to say how the collective interpretationist performs with respect to the reductive ambitions that are typical of the research programme on collective intentionality.

3.3 Collective-level interpretationism

Our aim is to extend the interpretationist approach to the case of collective entities and their attitudes, including hybrid ones. The hope is that this yields an account that deals smoothly with the kinds of complications identified above, involving components of collectives that aren't people or even agents, and involving factors outside collectives that contribute to fixing their attitude contents.

To develop such an account in full detail would be a book-length undertaking, so we're setting ourselves a more modest goal: to set out the general features of a collective-level interpretationism, to identify and assess the theoretical choices that arise along the way, and to apply the approach to the aforementioned complications. The strategy will be to copy over the structure of individual-level interpretationism, and then see how the elements of that structure could be plausibly filled in for a collective-level entity. In the best-case scenario, this process will be instructive not only by shedding light on the phenomenon of collective intentionality, but also by telling us something about intentionality in general, as a phenomenon that can manifest at various levels and in various forms.

Just as in the individual case, we imagine an ideal interpreter, unlimited in their ability to process information, who has access to all the information about the interpretee that doesn't involve the target facts–they know all that goes on in and around the collective. Using those facts, and in particular facts about the collective's evidence and actions, they come up with the set of beliefs and desires that best explains, in a rationalising way, what actions the collective takes in light of its evidence.

The first challenge we run into is to characterise the materials that the interpreter works with in order to do their interpreting. What is evidence for a collective, and what counts as an action of a collective? These are fairly thorny questions–arguably thornier than they are in the case of individuals–and the matter is complicated in particular by the fact that one collective agent can differ more profoundly from another in its cognitive architecture than one human agent does from another. We discuss collective evidence and collective action in Sects. 3.3.1 and 3.3.2. The general lesson from these discussions, discussed in Sect. 3.3.3, is that an ideal interpreter will have to pay attention to the *structure* of a given collective in order to interpret it successfully. As a result, the procedure that an ideal interpreter undertakes in the collective case looks somewhat more complicated than it does in the individual case–or rather, complications are revealed in the collective case which in the individual case are obscured. In Sect. 3.3.4 we draw conclusions from this discussion for the metaphysics of collective intentionality, and hybrid collective intentionality in particular.

3.3.1 Collective evidence

For the same reasons that we gave in the individual case, it is necessary for our theoretical purposes to construe the notion of evidence as something already intentional in nature–as propositional information gathered by the entity from its environment, the sort of thing it could be modelled as updating its credences on. What kinds of states of a collective agent, or of parts of a collective agent, look like they could be called upon to play this role? Some options would include:

- High-level intentional states–in particular beliefs–of the individual agents involved in the collective.
- Low-level intentional states–e.g. perceptions–of individual agents involved in the collective.
- Some essentially collective state of possessing evidence, defined in some manner out of high- and/or low-level intentional states of agents involved in the collective.

A tempting way to proceed would be to pick one such option and make a straightforward theoretical identification between those states and a collective's states of having evidence, i.e. say that across the board, the former kind of state realises the latter kind of functional state.[24] However, this is not going to work. Suppose for illustration that we said that a collective c has evidence that p iff c has a member that believes that p, and left it at that. We would run into a problem of *overinclusion* and a problem of *underinclusion*.

Let's start with overinclusion. It is implausible to identify the collective's evidence with *all* of the beliefs of individual agents involved in the collective. For only some of these agents' attitudes are relevant to the part they play in the collective, and many of them would get systematically ignored by the information processing that happens within a given collective. Suppose that Karl works within a collective: an airline. Suppose also that he has extensive opinions on brands of breakfast cereal. His role within the collective–let us say he purchases airplanes for the airline–does not involve breakfast cereal, and there is no mechanism by which those opinions of his would affect the collective's behaviour aside from by peculiar happenstance. It is not plausible, then, that those opinions of Karl play a role in determining the collective attitudes of the airline. So if we say that the beliefs of members constitute a collective's evidence, we should at best take ourselves to refer to a refined subset of those beliefs.[25]

Then there is underinclusion. As we've argued in Sect. 2.3, collectives can have components which are not agents and may not have states which qualify as beliefs or perceptions in a functional sense. Nevertheless, such entities can contribute to the information that the collective acts on. Suppose that in some city Uber installs sensors which go 'ping' when it rains, so that the collective can anticipate a surge in demand and adjust prices to take advantage. Is the rain sensor an agent which believes that it is raining? Perhaps on a very undemanding construal of agency, but more plausibly it plays the role of an agent at best partially: it does do something like perceiving, but we would not say that it responds to those perceptions rationally in the light of its desires. In functionalist theories of mind, functional states are defined with reference to each other: for some state to function as a belief, it has to interact in a certain way with desire states, and for some state to function as a desire, it has to interact in a certain way

[24] A note on terminology: for brevity we speak of 'identifications' between a functional state and a realiser state. There is, of course, a debate on whether or not the relation between functional roles and realisers is correctly understood as identity and, if so, whether the identities posited are between type states and token states. We don't mean to take a position on this debate: for 'identifying', the reader should feel free to substitute the positing of whatever relationship one takes to hold between realiser states and functional states.

[25] This reflects a general theoretical point which already applies to collective attitudes of the aggregate variety (see Sect. 2). Aggregation functions need not take into account all of the opinions that are available, and for some theoretical purposes it is best to ignore some as irrelevant.

with belief states. Functions like perceiving or intending are likewise defined within a broader system of functions. Thus it is difficult to count as an agent something which does only one of these things.

But even if the rain sensor is not an agent, we still want to consider the information that it holds as evidence that the collective Uber has at its disposal.[26] Thus, however we conceive of a collective's evidence within our account, we will not want that conception to be so restrictive as to allow only the beliefs of full-fledged agents to contribute to it. Whether the rain sensor perceives or has beliefs should not matter, as long as its information can be treated, by the collective that it is part of, as functionally equivalent to a belief on the part of one of its human agents, which is to say: as alike in playing the functional role of a perception on the part of the collective.[27]

This is a general point. If we start with some other general identification between the state of collective perception and some other type of lower-level state (say, the perceptions rather than the beliefs of members) these problems manifest differently but they still arise. The conclusion to be drawn is we shouldn't hold out for a general correlation between collective perceptions and some antecedently unified type of state of its components. Things of various kinds can turn out to play the functional role of perceptions of a collective agent, and not all things of those kinds need to play that role. Nor should we presume that because such-and-such states serve as perceptions for a given collective, those same kinds of states will also play that role in a different collective. That being said, limited generalisations will nevertheless be available, since many collectives are instances of general types of collective (e.g. courts of law, research teams) instances of which are alike in how they process information.

So how *does* the ideal interpreter latch on to the evidence of a collective that they're trying to interpret? We can get a better idea by first looking at the other kind of input: the actions of collective agents.

3.3.2 Collective actions

As indicated earlier, we need to construe the notion of an action as something already bearing an intentional characterisation. Specifically, Williams-style interpretationism models actions as choices among sets of options, so as to render the theory's decision-theoretic apparatus applicable. It is such choices that the interpreter seeks to rationalise relative to the agent's evidence.[28]

[26] An alternative construal would be that the information in question only becomes part of the collective's evidence if and when some fully-fledged agent within the collective takes note of the 'ping'. This seems a less plausible construal to us: if the 'ping' went unnoticed by anyone in the collective, it would seem more natural to say that Uber failed to act on evidence that it had, rather than that it failed to get the evidence. But even if this example proves unconvincing, the general structural point should be clear enough.

[27] Note also, for what it's worth, that there is typically a kind of transformability here: the information in the computer can be taken up by an employee of the company, in the form of a belief, and the content of an employee's belief can, in some fashion, be put into the computer.

[28] There is always a risk of running together two senses of 'intentional', one a broader sense that applies to all states or things that have a content or are about something, and a narrower sense that applies to actions when they've been 'intended'. In this case, the two just happen to coincide: we give actions an intentional characterisation in the broader sense precisely by associating them with an intention in the narrower sense.

In the individual case, this creates a need to identify what, from among all the stuff that goes on in and around an agent, should count as their actions, and what the correct intentional characterisation of those actions is. This is also the case at the collective level, but there are additional complicating factors: collective agents can count as performing an action in quite a few ways, and in addition one collective can differ quite a bit from another in what actions they can perform and how they can perform them. Here are some examples of ways a collective agent might act:

- A collective agent might perform an action by means of some individual agential component (i.e. a human agent, a non-human agent, or a partial agent) of the collective performing an action on behalf of the collective;
- A collective agent might perform an action by means of some subset of the collective's components, which itself constitutes a smaller collective agent within the larger collective agent, performing an action on the larger collective's behalf;
- A collective agent might perform an action by means of some subset of the collective's agential components acting jointly on the collective's behalf, but without constituting a full-fledged collective agent within the larger collective;[29]
- A collective might perform an action by means of all of the components (agential or non-agential) acting jointly on behalf of the collective.[30]

For example, Starbucks might decide to invest in some new type of coffee machine simply in virtue of the fact that some one person in the company has, and uses, the authority to make that decision on behalf of the collective. Or Starbucks' Beverage Development Team might, through close cooperation, develop a new kind of frappuccino, thereby making it the case that Starbucks has developed a new kind of frappuccino. Or the entirety of Starbucks might decide, through some voting procedure, to buy out its stockholders and turn itself into a worker's cooperative. Despite the differences in realisation, all of these seem to count equally well as actions of Starbucks.

Note that the 'on behalf of the collective' condition appears in all these cases. It is crucial, for any of these things might occur without counting as actions of the collective in question. Individuals in the collective may act on their own behalf rather than on the collective's behalf, and the same goes for pluralities of individuals acting jointly. Any sub-collective of the collective might simultaneously constitute a different collective and some of its actions might count only as actions of this different collective. At the limit, all of the collective's components might simultaneously constitute a different collective. It is neither necessary nor sufficient for a collective agent to act that the totality or a plurality of its individual components are involved in the act. Thus a theory of joint action, of the kind that e.g. Bratman has developed, is importantly different from a theory of collective agency. In order to account for all the ways in which collective agents can act, we will want to have a theory of joint action to build on. But equally crucial is to be able to say when some action, be it one executed jointly

[29] As noted in Sect. 3.1, we assume that a plurality of individuals can act jointly without constituting a collective agent. For something to be recognisably a joint action, the individual contributions involved would have to exhibit a certain amount of coordination and 'meshing', as Bratman (1999) calls it. But this does not seem to require the existence of a collective agent with its own beliefs and desires.

[30] This is really just the limit case of the former action-type where the subset is improper.

or individually, counts as the action of a collective agent. A theory of joint action as such doesn't tell us this.

The actions that do belong to the collective agent must then also be associated with the right intentional characterisation, i.e. as cases of the collective intending to ϕ, for some ϕ. Functionally, intentions are states which are the outcome of an agent's deciding (on the basis of their desires, beliefs and evidence) among some options ϕ, χ, ψ available to them, and which give rise to activity that is such as to causally bring the option about. By tying actions to intentions, we can gloss an action done 'on behalf of' the collective as an action that results from an intention of the collective. The challenge then becomes that of associating the collective agent with a set of intentions.

As with evidence, with intentions we should also expect heterogeneity in the realisers. While we may be able to identify some central, default mechanisms by which collectives intend things, collectives can be built out of varied components (some of which may not be agents) and they can differ notably in structure from case to case. Among the states which realise collective intentions, we may find the following:

- Intentions formed by either individual agents within a collective or by collective agents that are part of the collective, when those agents are acting in an official role associated with that collective.
- Joint intentions (as described by theories of joint action such as e.g. Bratman's) formed by all or some of the collective's members.
- Decisions reached by authoritative members of a collective (individually or jointly) the execution of which is delegated to other members of the collective.

In all of these cases, the collective intentions would be realised by what are arguably already intentional states. But we can and should cast our net beyond this. Imagine that the airport security system includes automatic safety doors which close when certain conditions are met. The closing of such a door (when it happens in the normal way) could count as an intentional action of the collective, even though it is automatic and the door has no mental states. It counts as such because, assuming everything is working as designed, airport security is hereby choosing among the courses of action open to it in response to what evidence it has and in the light of what it believes and desires. The door's closing plays the functional role of an intentional action within the overall system, and that suffices.[31]

These considerations imply that, in figuring out what behaviour of a collective the interpreter needs to treat as actions to be rationalised, we again need to guard against overinclusion (not every action that takes place within the collective is 'inherited up' as an intentional action of the collective agent) and underinclusion (the functional state of a collective's intending to ϕ could be realised in various ways, including states which are not, when considered by themselves, already intentional ones).

We have looked at collective evidence and collective action, and in both cases we've identified some interesting, and broadly analogous, theoretical challenges. Let's see how collective interpretationism can address these challenges.

[31] In this case, it may be impossible to identify separate states of the collective which underlie its *intending* to close the door and its *closing* the door. That is okay: the same states of the system can in some cases realise both functional states, and arguably this is what we typically find when an agent is capable of actions that are both intentional and automatic.

3.3.3 Source intentionality and structure

In the case of individual interpretationism, we already identified the theoretical challenge of characterising an agent's evidence and intentions, in such a way that the decision-theoretic apparatus involved in rationalisation can get a grip on it. Interpretationists like Williams see this as an additional leg of the research programme; the interpretationist must first 'earn the right' to speak of agents having evidence and intentions (what Williams calls 'source intentionality') by giving some theory of it (in the case of Williams (2020), a teleosemantic one) so that they can consider this as information that the ideal interpreter could have at their disposal, in interpreting an agent.

Note that though this is something that the interpretationist theorist must deliver on, characterising this source intentionality is not itself seen as part of the procedure of ideal, radical interpretation. This source intentionality is what the ideal interpreter starts from, not something they deliver as a result of interpretation. However, to address the challenges we've identified in the foregoing two sections, we may want to revise this element of the approach. We may want to consider the task of characterising source intentionality as *folded into* the task of interpretation itself. Doubtless this will sound a little puzzling, so we'll explain what we mean by that.

First note that while we have identified the potential for quite a bit of heterogeneity in what might realise the evidence and intention states of collective agents, *for a given agent* we ought to be able, after sufficient study of behaviour and its workings, to give an account of what it is, for this agent, to have evidence that p or intend to ϕ. If the agent is an instance of a general type of collective, we may also be able to generalise this account to other collectives of that type. Evidence and intention states work a certain way for a given collective because that collective is organised so that information flows into it and through it in certain systematic ways, and affects its behaviour in certain systematic ways. It is this which determines that the breakfast cereal opinions of Karl (airplane purchaser) do not count as evidence that the airline has, whereas those of Janine (head of in-flight catering) do. To put it another way, a collective agent has a certain cognitive *structure*, and insofar as this structure is stable, what makes for evidence and intention for that agent is also stable. The requirement that the ideal interpreter have access to (the correct characterisation of) the collective agent's evidence and intentions could also be expressed as the requirement that they know the agent's structure, in this sense.[32]

Something analogous is true of human individuals, of course. Human beings have certain perceptual capacities, which determine what information they can extract from their environment, and they have a certain neural and motor system which determines what basic range of actions they can intend. The difference is that for humans, this structure is by and large stable across individuals. There is cognitive diversity, but it is much less profound than in the case of collective agents, whose diversity is bounded only by the limits of humans' ingenuity in organising themselves so as to act as one. Because of this, it's not very natural in the individual-level case to picture the ideal

[32] The importance of the structure of the collective for ascribing corporate attitudes is also acknowledged by the formal ontology of group agency proposed in Porello et al. (2014).

interpreter as treating it as an open question, when they interpret some individual, what sorts of things constitute evidence and intention for this creature. It's far more natural to picture this interpreter as 'pre-loaded' with a set of (objectively correct) assumptions about what constitutes evidence and intention in humans, and thereby as having available to them the evidence and intentions of this individual, given that they already know every physical detail of what goes on in and around the individual. There is no such general set of assumptions about cognitive structure for a collective-level interpreter to fall back on, so by contrast it does make sense to picture them as treating it as an open question what constitutes evidence and intention for the particular collective agent they're interpreting.

Our proposal is that we think of the procedure of interpreting collective agents as having an extra preliminary stage which doesn't feature in the procedure of interpreting individuals, or features there in a way that's much more minimal. This is a stage in which the interpreter determines the structure of the collective under examination, and thereby its evidence and actions. How does the interpreter do this? Our answer: by an expanded procedure of rationalisation.

Imagine our collective-level interpreter as starting with an unrefined, 'blurry' picture of the collective it seeks to interpret. This gathers together all of the activity performed by components of the collective, and all the 'intentional' states of components of the collective, where this is broadly construed so as to include the quasi-intentional information-bearing states of non-agent entities such as metal detectors. But not only that: they also take into account states of the world well beyond what we might pre-theoretically consider part of the agent itself. Because the ideal interpreter is cognitively unlimited, this is no undue burden on them. Anything causally connected to the agent is in principle in the frame, when it comes to characterising the cognitive structure of the agent.

The interpreter then comparatively evaluates *candidate structures* of this collective, each of which is such as to generate a certain assignment of evidence and intentions to this collective on the basis of the unrefined information. (Considered concretely, a structure can be viewed as a description of how information is processed and acted upon by the interacting components of the collective; more abstractly, as a function from physical states of the collective to sets of evidence- and intention-states of the collective.) These candidate structures are compared on how well they do in rendering the collective as a rational agent.

To fulfil their task of rendering the collective agent rational, the interpreter should, all other things being equal, assign that structure to the agent which allows them to maximise the rationality of the agent's behaviour. That is to say: they will favour the structure which puts out a set of evidence and actions which, when interpreted, allows for the attribution of the most rationalising set of beliefs and desires, all else being equal.[33] The stage in which the interpreter assigns a structure to the collective should not really be imagined to temporally precede the stage in which they assign beliefs

[33] There are of course devils in the details. This proposal assumes that we can take a candidate interpretation available under one structure S and measure how well it does against a candidate interpretation available under another structure S*. That there will be commensurability of interpretations across candidate structures seems plausible, but is hard to argue for decisively without pinning down the formal structure of the interpreter's theory-choice situation in greater detail than we have space to do here.

and desires to the collective; rather, the interpreter should be imagined to solve for these variables together.

3.3.4 Grounding and fixing belief- and desire-states

There are further interesting things to be said about the structure of collective agents, which we'll do in Sect. 3.4. But at this stage, we are in a position to return to the two challenges we originally set ourselves: incorporating cases where attitudes of collectives fail to supervene on states of their members, and incorporating cases of hybrid collective intentionality. Let us take these in turn. With regard to supervenience failures, there are two main points to make.

First, recall that in the discussion of individual interpretationism, we distinguished two different determination relations: the 'fixing' relation and the grounding relation. In the collective case the same distinction applies. By looking to the evidence and intentions of a collective agent, the interpreter can figure out what beliefs and desires it needs to attribute to this agent. In so doing, it is tracking the informational 'fixing' relation that holds between facts about source intentionality and facts about intentional states like beliefs and desires. Since it also knows all of the physical facts about what goes on in and around the agent, it is then in a position to infer what physical states of the agent constitute these belief- and desire-states; in so doing it would detect the metaphysical grounding relations between facts about belief- and desire-states and non-intentionally characterised physical facts. In the expanded procedure of interpretation we have envisaged for the collective-level ideal interpreter, it would similarly associate the evidence- and intention-states of the agent with physical states that ground them; this is effectively what it is for the interpreter to assign a structure to the agent.

As discussed in Sect. 2.2, an assumption is frequently made in the literature that espousing reductionism about collective intentionality involves thinking that collective intentional states supervene on individual intentional states. This now turns into two questions. First, are the facts that fix collective beliefs and desires all facts about individual attitudes? Second, are the facts that ground collective beliefs and desires all facts about individual attitudes?[34] On the basis of the cases discussed in Sect. 2.2, one may suspect that the answer will be 'no' in both cases. But since we had not yet then distinguished the two questions, it is worth discussing examples that speak specifically to each.

As an example of the first sort of failure of determination–relating to the fixing relation–consider proxy actions. A company might have lawyer on retainer who steps

[34] Brian Epstein (2016) distinguishes these two questions (or very similar ones) in terms of his own distinction between a 'grounding' enquiry and an 'anchoring' enquiry. In a grounding enquiry, one tries to find the grounds of some social fact; in the case of a functionalist theory of collective intentionality, that amounts to finding the realisers for the collective's intentional states. In an anchoring enquiry, one tries to find the facts that determine the operative grounding conditions; in the case of a functionalist theory of collective intentionality, that amounts to identifying the functional roles that are being realised. One could be individualist about either question–that is, one could propose that the facts that respectively do the grounding or do the anchoring are all facts about individual members of collectives, or (less demandingly) just facts about individuals, full stop. Epstein suggests, as do we, that individualism is likely to be true neither in the case of grounding (our question (b)) nor in the case of anchoring (our question (a)).

in to deal with certain kinds of legal troubles as they arise, using their own judgement in choosing what to do. Because they act on behalf of the company, their actions in these cases are actions of that collective, yet they do not consist in any attitudes or behaviours of the collective's members. The interpretation (that is, assignment of belief- and desire-contents to states of the collective) that best rationalises the evidence and actions of the company should rationalise these actions too; but were we to consider only the attitudes and behaviours of members, we would fail to be responsive to this demand, and we would have made unavailable to the interpreter an interpretation that would do a better job of rendering the collective agent rational.

As an example of the second sort of failure of determination–relating to the grounding relation–consider an electricity company that deals with its customers primarily through an automated system that keeps track of their electricity usage, account balance, etc. When we want see what this company believes about a certain customer's account balance at a certain time, we should be looking at the state of its automated system at that time, rather than at anything in the heads of any of the company's members. The most rationalising interpretation of this collective agent would attribute to it beliefs about e.g. the customer's account balance. But if we were to ignore the physical states of the automated system as possible grounds of attitudes of the collective, we would not have anything available to us to ground these beliefs we would like to attribute to it.

The collective interpretationism we have outlined would not run afoul of these sorts of cases. As to the first, as we have characterised the procedure of ideal interpretation, we have left it effectively at the discretion of the ideal interpreter to select, among all that goes on both in and around the collective agent, what would best serve as the collective agent's evidence and actions, for the purposes of radical interpretation. No specific metaphysical theses about what sorts of states *could* in principle serve this role are built in. So the action-intentions of the lawyer, though they are not action-intentions of a member of the collective, can be pressed into the role of action-intentions of the collective agent, on the basis that they relate appropriately to the cognitive structure of the agent. As to the second, we have similarly not built in any restriction as to what sorts of states the collective interpreter can identify as grounding states for the belief- and desire-states they attribute to the collective agent. Any physical state of the world that stands in the appropriate causal relations to other physical states of the cognitive agent can be pressed into this role. So the physical states of the electricity company's automated system can serve this role, in virtue of them appropriately causally interacting with the states that ground evidence-, intention- and desire-states of the agent.

This leads directly to the second point, which can be made more briefly. Collective interpretationism can deal smoothly with these types of cases because, in setting it out, we have decoupled the question of what figures in a collective agent's cognitive structure from the question of what is and isn't part of the collective agent. Because of this, the former can range beyond the latter. One need not necessarily take this result at face value. One response to it is to be *revisionist* about the question of what is and isn't part of a collective: one could take it to be evidence that the boundaries of collectives may turn out to be quite different from what we pre-theoretically consider them to be, and that we can only really find out how to draw these boundaries by paying attention

to cognitive structure. It may then turn out that the MBTA board's parliamentarian is in fact part of that collective, and that the lawyer-on-retainer is in fact part of the company.

This revisionary approach could be advertised as giving us a systematic approach to certain ontological questions about collectives. From a different perspective, that may equally be taken to be its drawback: one might feel that the ontology of collectives shouldn't be approached in this systematising spirit, because collectives, as human artefacts, have many of their features–among them their membership conditions–by convention and/or stipulation. Answering the question of what is and isn't part of a collective–one might then think–does not so much call for abstract theorising as for inspection of what norms and principles we have in fact contingently laid down for the collective in question. Those who feel this way should leave the matter of collectives' boundaries decoupled from that of collectives' cognitive structures.[35]

This leaves the matter of hybrid collective intentionality. We argued in Sect. 2.4 that collective agents can involve components that are not human and not even agents, and that these components can play an explanatory role with respect to the intentionality exhibited by the collective agent. We have intended to make room for this phenomenon with our collective interpretationism. It is common to approach the question of collective intentionality as a challenge to explain how one form of intentionality–individual intentionality–is transformed into another form of intentionality–collective intentionality. Though it feels natural enough to frame the question this way, it makes it difficult from the get-go to incorporate hybrid collective intentionality. One could make headway with artificial components of collectives that themselves qualify as agents, for these could be taken, in virtue of their agent-hood, to be contributing intentionality in the way that human agents do. But one would struggle to incorporate anything less sophisticated than this into the picture.

We have approached the question of collective intentionality in a 'top-down' and thoroughly functionalist way. The ideal interpreter we posit is tasked with finding an interpretation of a collective agent that makes it maximally rational, other things being equal, and it works back from there to find, among what goes on in and around that agent, states that can be interpreted so as to achieve this task. It is tempting to interpret the collective that is airport security as responding rationally to its evidence in the light of its beliefs and desires, when it closes an automatic door to stop a passenger who sets off a security alarm. But to make the interpretation available on which this is indeed counted as rational behaviour on the part of the collective, the interpreter needs to endow relevant states of these components with contents so as to make them, functionally speaking, perceptions and action-intentions of the collective agent. This it could not do if its attention were restricted to states that would already, outside of the context of the collective agent, be considered intentional ones, like the attitudes

[35] One could appreciate this latter way of thinking without concluding that who and/or what can be included in a collective is completely arbitrary. First, *given* a membership convention for a given collective, who counts as a member is of course a systematic matter. Second, it may be perfectly possible to state local generalisations governing membership for types of collectives, e.g. parliaments, or football teams. And the possible variation in membership conditions might be constrained by more general (if still conventional) higher-level rules, e.g. legal principles. Third, it might be that there are more-or-less natural default membership principles that apply for informal groups that form spontaneously, like friendship groups.

of individual agents. We have not restricted our ideal interpreter this way, allowing it instead to be led entirely by the requirements of its rationality-maximising brief. This allows our account to generalise smoothly to cases of hybrid collective intentionality.

3.4 Collectives and their structure

3.4.1 Ascriptions of structure

We have set out the architecture of our collective interpretationism and explained how it allows us to respond to the challenges we originally set ourselves. In the remaining sections of this paper we take a closer look at the notion of a collective's structure. Although we've tried to illustrate it somewhat, we have mostly characterised this notion abstractly, as a function from physical states of a collective to sets of contentful states. Here we unpack it more, and relate it to familiar ways of thinking and talking about the structure of collectives.

If we step away from our particular theoretical purposes for a moment, and ask the general question of what the structure of a collective is, it quickly becomes clear that there is more than one theoretically respectable way in which one might ascribe a certain structure to a collective.

- One could be ascribing a certain *metaphysical* structure (as one could call it) to the collective. This could be (very abstractly) thought of as a big grounding function which tells us how all the facts about the collective depend on lower-level facts. It could be cashed out as a set of conditionals expressing grounding conditions for various kinds of facts about the collective. The same people and objects, doing the same things, could give rise to a differently-behaving collective, depending on how those people and objects relate to the collective in question. The metaphysical structure is what captures how the facts about the whole depends on the facts about its parts (and possibly on things not among its parts).
- Second, one could be referring to the *cognitive* structure of the collective qua cognitive agent. This would describe the kinds of intentional states the agent can be in and how those states relate to one another, both constitutively (i.e. how the states depend on each other for the fixing of their content) and causally (how they are formed and how they affect each other under conditions of well-functioning).
- Third, one could be referring to the *organisational* structure of the collective: what roles there are to be fulfilled in the collective, how the people/entities filling those roles relate to each other and communicate with each other, who has authority over what and over whom, who is responsible for what tasks, etc. In short, it is the kind of description that an organisation theorist might give of an organisation.

These are all legitimate ways of talking about the structure of a collective, and so two questions arise. (1) How do these kinds of structures relate to one another? (2) How do these kinds of structure contribute to fulfilling the theoretical role that we carved out for the notion of the structure of a collective?

Let's start with the second question. Our radical interpreter pins down a collective's structure in order to isolate the collective's perceptions and action-intentions from all the activity and intentionality that goes on in and around the collective. What this

structure tells the interpreter is how lower-level facts (about what individual agents within the collective agent think, say and do, about the activities of various non-agential components of the collective agent and–in some cases–about states of the world beyond the collective agent) ground states of perceiving and intending on the part of the collective agent. Having acquired such information about perceptions and intentions, the interpreter can attribute to the collective a set of higher-level attitudes: beliefs and desires. The interpreter would then be in a position to correlate the attributed attitudes with other lower-level facts about the agent, to discern grounding relations between the agent's beliefs and desires and lower-level goings-on.

So structure, in this story, is in the first instance *metaphysical* structure as described above, a function from lower-level grounding facts to higher-level grounded facts. We could characterise a structure of this sort as a set of conditionals associating types of facts about collectives with grounding conditions stated in terms of lower-level facts. The radical interpreter will, however, be interested only in those conditionals which relate lower-level facts to facts specifically about intentions, actions, beliefs and desires, a subset of all the grounding conditionals that pertain to the collective.

But (turning to the first question) the notion of *cognitive* structure as described above is also in the picture. For the interpreter's activities of picking a structure and picking an interpretation under a structure are primarily constrained by the goal of making the collective agent as much rational as possible. And for this, they have to attribute a set of attitudes that hang together appropriately: the perceptions have to be states which register states of the outside world and constrain belief-formation; the beliefs have to be responsive to perceptions and interact with desires in the right way; desires, by interacting with beliefs, have to give rise to appropriate action-intentions, which in turn have to cause suitable behaviour, and so forth. In short, the states that are attributed have to stand in the right kind of structure, a cognitive structure that's rational, or at least the most rational of those available to the interpreter. This kind of structure doesn't consist in inter-level relations between intentional states and lower-level states, but in intra-level relations between collective-level intentional states.

Since we're dealing here specifically with *collective* agents, the intentional states we're interested in aren't realised by means of eyeballs, neurons, etc. but primarily by people interacting socially in certain regulated, effective, purposeful ways. So to understand how a collective could be pulling off the trick of being an agent, we look at the relations that hold between those people, the norms that guide them, the way information flows between them, how they interact with bits of technology, etc. In so doing we look at a structure of the third kind, the organisational sort. This is again a sort of structure that's intra-level, in this instance one that consists in relations between entities at the non-collective level.[36]

The theory of collective intentionality which we have set out turns out to have interesting applications to the general theory of social collectives, beyond the specific case of their actions and attitudes. To show that, we'll draw on what we regard as one of the more promising approaches to the ontology of collectives, the theory of collectives as variable embodiments.

[36] Bearing in mind of course that some of the entities that show up at this level may themselves happen to be collective agents.

3.4.2 Embodied collectives

The importance of structure for collectives and of the functional roles played by their members is well-acknowledged in the literature on the ontology of groups (see, for instance, Varzi (2006), Ritchie (2015) and, more recently, Uzquiano (2018), Ritchie (2020) and Harris (2020)).

We employ here Kit Fine's theory of collectives, for two reasons. The first reason is that, as we shall see in this section, it provides a way to single out those individual attitudes of the members that contribute to forming the collective's attitudes, enabling us to disentangle the evidence- and action-states that matter for attributing beliefs and desires to collectives. The second reason is that, as we shall see in the next sections, Kit Fine's account is well-suited for approaching the issues of persistence and diachronic identity of groups.

We start by introducing Fine's notion of a *qua-object*. In Fine (1982), *qua-objects* describe the relationship between a material thing and the matter by which it is constituted. Interestingly, he includes social roles and their players among *qua*-objects, even though his theory isn't concerned with them especially:

> Given any object x and description (property) ϕ possessed by x, we shall suppose there is a new object x *qua* ϕ or x *under the description* ϕ.
> […] Given such an object as x *qua* ϕ, we shall call x the *basis* and ϕ the *gloss*. The resulting object itself will be called a *qua-object*, and the operation by which it is formed *glossing*.
> […] the *qua*-object should be regarded as some sort of amalgam of the given object and the property, like the given object but wearing the property on its face.
> […] a *qua*-object exists just when its basis does and satisfies the gloss. (Fine 1982, p. 100)

If we apply Fine's terminology to the problem of appropriately ascribing evidence and actions to collectives, we might say that the contentful states that matter in this regard are attitudes and actions of *qua*-individuals (or, more generally, the features of *qua*-objects). In the bowling team example of Sect. 2.1, we are not interested in Alice's, Bob's, Charlie's, Delilah's and Eve's attitudes and actions *simpliciter*, but in the attitudes and actions of Alice-*qua*-player, Bob-*qua*-player, Charlie-*qua*-player, Delilah-*qua*-team-captain, Eve-*qua*-team-manager, etc., of the Sunday Afternoon Champions.

An interesting feature of *qua*-objects is that they inherit some of their properties from their 'basis' and some from their 'gloss'. This seems very apt when we are dealing with agents-in-a-role: the *qua*-individual inherits some of its properties from the basis and the powers, obligations, permissions etc. from its gloss. How Alice will play in the next bowling match will depend on her capabilities, her physical fitness, her training and so forth; these properties of Alice, the basis, will determine how Alice-*qua*-team player will execute her throws and so contribute to the team's performance. But the fact that the score obtained with her throws will sum up with those of the other players of the Sunday Afternoon Champions depends on the fact that she's a member of that team while that very match takes place, so it depends on the gloss of the *qua*-individual. A

similar mechanism applies also to artificial components of a collective. For example, the way in which the Uber's dispatcher software performs its tasks depends on how its algorithm is built–the algorithm here constitutes the basis of a *qua*-object–while the fact that the software is allowed to send messages to Uber drivers, who may commit to go where the dispatcher tells them, depends on the fact that such software plays the role of dispatcher for Uber–i.e. it depends on the properties of the gloss.

We can tell from this example that the basis and the gloss don't just separately make their contributions to the properties of the *qua*-individual, but interact in a certain way. Specifically, the gloss determines *how* and to what extent the properties of the basis contribute to the properties of the *qua*-individual.

The notion of a *qua*-object can also be applied at the collective level, by constructing a *qua*-entity out of a collection of entities, i.e. a *qua*-entity the basis of which is given by a collection of entities.[37] In this case, the gloss is a complex description that states the relations holding among the entities that compose the collective. We denote this complex relation by R. Following Fine in replacing the *qua* expression with the forward slash ('/'), we represent the *qua*-object constructed out of a collection of individuals a, b, c, ... by $(a, b, c, \ldots)/R$ and we term this complex *qua*-object an *embodied collective*.[38]

This gives us our abstract understanding of the notion of *structure* for collectives: it is represented by the R at play in determining the embodied collective. At this level of abstraction, in principle, the complex relation R can represent the views of structure that we highlighted in Sect. 3.4.1.

We illustrate now how the structure R works in specifying the roles of the entities that compose the embodied collective. When we turn from *qua*-entities whose basis is an individual to embodied collectives, it seems we could construe them in two different ways:

(i) the embodied collective is (as we defined) a single complex *qua*-object, the basis of which is a collection of agents and the gloss of which is its structure R;
(ii) the embodied collective is a collection of *qua*-objects, each composed by a basis and a gloss.

Following Fine's notation, the two situations are depicted as follows: i) corresponds to $(a, b, c, \ldots)/R$, whereas ii) corresponds to $(a/\phi_1, b/\phi_2, c/\phi_3, \ldots)$

Happily, these two alternatives are, under certain circumstances, equivalent: in order to understand whether a member of the embodied collective is a *qua*-object of a certain sort, what we have to do is look at whether it plays a certain role in the whole structure provided by R.[39]

[37] See Porello et al. (2014) for a view of group agents constructed in this way. Notice that the ontological nature of the basis requires clarification. For instance, the collection of entities that form the basis may be construed as a mereological sum, a set, or as a plurality of entities. We leave this matter to be discussed elsewhere, as the intuitive understanding of a collection of entities suffices here.

[38] As we shall see in the next section, this is indeed the notation that Fine uses to represent (rigid and variable) embodiment.

[39] Here we draw on Masolo et al. (2004), where, as for Fine, *qua*-individuals are entities 'generated' when an entity plays a role, which inherit some of the properties from the entity playing the role (the basis) and some from the role (the gloss). Masolo et al. (2005) then introduces 'relational' roles: each time a collection

The two alternatives could thus be seen as two equivalent descriptions of the same entity, the latter being a sort of distributional version of the former, which enable us to elicit the roles of the members of the collective from the structure of the collective.

We have the abstract beginnings of an ontology of collectives: they are embodied collectives, i.e. complex *qua*-objects that are the result of imposing a certain gloss (a compound of relational properties) on a basis which is a plurality of appropriate individuals. Note that nothing in the picture so far forces us to think of the individuals as human agents; they could be any object that has some properties to contribute to a collective.

The understanding of structure that we have proposed complements our discussion in Sect. 3.3.3. Summarising, the interpreter firstly ascribes a candidate structure (R) and candidate members (a, b, c, …) to the collective. Then, since the structure R provides the glosses of the members, R enables the interpreter to select the actions and evidence of the members of the collective that count as evidence for the collective's behaviour, disentangling them from the surrounding noise. The action and evidence of the members that counts for the collective are indeed those of the members *qua* members of the collective.

This mechanism works well if we imagine the task of the radical interpreter frozen at a single instant in time. However, collectives are entities that may exhibit quite complex variations in time. Everyday experience shows very clearly that collectives can radically change. Firstly, the members of a collective can change. Also other crucial features of collectives may change: the requirements, powers, duties, responsibilities, etc. that are associated with roles within that collective, the internal norms regulating the collective–including norms that regulate such changes–the tasks, the purpose of the collective, etc. It is very easy to envisage examples of groups that manifest such complex changes. Basically every feature of the collective can change over time, and such changes pertain both to the way that the collective is realised (the basis) and to its structure (the gloss). We nevertheless want to be able to talk about such changes as things that happen to a single continuous entity. It should make sense to say, e.g., that nowadays ethical values play a much more central role for company x than they have in the past.

To deal with change, we shall follow Fine further, and rely on his *theory of embodiment*. This theory can be seen as a natural extension of the theory of *qua*-individuals.[40] And indeed, recently Fine himself has applied his theory to the case of social groups,

Footnote 39 continued
of individuals is classified by a relational role (like our R), a collection of *qua*-individuals is created. Thus, every time we can create an embodied collective $(a, b, c, \ldots)/R$, the complex pattern of relations R is capable of ascribing the relevant gloss to the component individuals a, b, c, …, forming the collection of *qua*-objects a/ϕ_1, b/ϕ_2, c/ϕ_3, etc. On the other hand, the direction from a collection of *qua*-objects to the embodied collectives works as well, provided the glosses $\phi_1, \phi_2, \phi_3, \ldots$ are properties that include suitable relations to each other.

[40] In his recent unpublished draft, titled "Acts and embodiment" Fine (20XXa), Fine explicitly presents the theory of *qua*-individuals as a theory of rigid embodiment. The paper is forthcoming in a collection edited by Alec Hinshelwood, titled *Being and Doing*. The draft is available at: http://www.academia.edu/35032853/acts_and_embodiment.pdf.

in a forthcoming paper (Fine 20XXb).[41] We summarise the basic elements of Fine's view in the next section.

3.4.3 Fine's theory of embodiment

Fine (1999) presents a theory meant to account for timeless and temporary parthood in material objects. Roughly speaking, Fine views material objects as constituted by their timeless parts assembled according to a particular 'relational principle'. The whole material object is for Fine a *rigid embodiment* $e = a, b, c, \ldots /R$, where $a, b, c, \ldots$ are its parts (or elements) and R is their relational principle, the relation in which such parts stand. For instance, a ham sandwich is constituted by two slices of bread and one slice of ham arranged in a certain manner.

> An object of this special sort will be called a *rigid embodiment*, since the "form" R is embodied in the fixed "matter" $a, b, c, \ldots$. Let us agree to designate such an object by the term "$a, b, c, \ldots /R$." The relation R will then be called the *principle of rigid embodiment*, and the operation by which a rigid embodiment is formed from the objects $a, b, c, \ldots$ and a relation R, the *operation of rigid embodiment*. [pp. 65-66]

When the constitution of an object is rigid, we can use the notion of rigid embodiment to conceptualise the object. But Fine also considers material things that can lose or acquire parts throughout their existence, like a car that requires to substitute parts periodically to maintain it in order. These are cases in which it is necessary to talk about a relation of temporary parthood, for which Fine introduces the notion of *variable embodiment*, to represent objects whose constitution can vary in time.

Material things like cars persist through time and can be reidentified through changes or rearrangements of parts thanks to a *functional principle* (F). This functional principle 'picks up', at any moment in time, the rigid embodiments that are *manifestations* of that specific material thing as a variable embodiment.

> In general, we will suppose, given any suitable function or principle F (taking times to things), that there is a corresponding object standing in the same relationship to F as the variable water of the river stands to *its* principle. We call this object the *variable embodiment of* F and designate it by $/F/$. The principle F in $/F/$ will be called a *principle of variable embodiment*, the various objects picked out by the principle F the *manifestations* of the variable embodiment $/F/$, and the *operation* "//" by which $/F/$ is formed from the principle F the operation of variable embodiment. In contrast to the case of a rigid embodiment $a, b, c, \ldots /R$, the matter of a variable embodiment is not given independently of the form or principle, but is itself specified by means of that principle. (p. 69)

In other terms, the variable embodiment $/F/$, where // is the operation of forming an object, is the result of the application of the principle F that picks up the rigid embodiments $e_{t_0} = a, b, c, \ldots /R, e_{t_1} = a', b', c', \ldots /R', e_{t_2} = a'', b'', c'', \ldots /R''$,

[41] The paper is forthcoming in *Metaphysics*. The view of collectives as variable embodiments has also been proposed in Ferrario et al. (2018) and Uzquiano (2018).

etc. which are the manifestations of the variable embodiment at the different times t_0, t_1, t_2, etc. in which the variable embodiment exists. As a consequence, all such manifestations are temporal parts of the variable embodiment and the functional principle F is what provides the identity of the variable embodiment. Fine (1999) defines the identity postulate for variable embodiments as follows:

> (V3) The variable embodiments $/F/$ and $/G/$ are the same iff their principles F and G are the same. (p. 70)

That is, two variable embodiments $/F/$ and $/G/$ coincide if and only if their respective principles pick, for every time, the same rigid embodiments.[42]

3.4.4 Collectives as variable embodiments

The theory of rigid embodiment is designed to capture objects that are rigidly constituted by their components, so it is not a good fit for collectives which can undergo change through time (or some other dimension, for that matter). However, it works well to characterise temporal states of embodied collectives, i.e. the fact that at a certain moment (or period) t, some embodied collective is formed by members $a, b, c, \ldots$ arranged in functional structure R (their relational principle, in Fine's terms).

We can view the Sunday Afternoon Champions at a point in time t as the rigid embodiment given by Alice, Bob, Charlie, Delilah and Eve kept together by a certain pattern of relations that creates a certain functional whole, in which Alice, Bob and Charlie play the role of players, Delilah that of captain, Eve that of manager, etc., with the concomitant relations of authority, responsibility and so forth.

Suppose now that at some point midway through the championship Charles is substituted with Carl. A different rigid embodiment would result, possibly preserving the structure R (the relational principle). If we want to tally up the score that the Sunday Afternoon Champions have achieved in the course of the championship, we should count the points obtained by Charles, say from t_1 to t_2 (while he was on the team), those obtained by Carl, say from t_2 to t_3 (when he was on the team), and those obtained by Alice, Bob and Delilah at any time, say from t_1 to t_3. Or in short, the points scored at any time by any person that was at that time a team member. That means treating the Sunday Afternoon Champions as a variable embodiment whose manifestation during one period includes Charles and during another period includes Carl.

The changes of the agents who are members of the collective or of the role-assignment are not the only possible transformations an embodied collective can face through time: their organisational structure may also change, by the addition of new roles, the elimination of existing roles, modifications to the powers, duties, etc.

[42] The identity criterion for principles is the usual identity criterion for functions. The identity criterion for rigid embodiments $e = a, b, c, \ldots /R, e' = a', b', c', \ldots /R'$ is given by: $e = e'$ if and only if $a = a'$, $b = b'$, $c = c'$, ..., and $R = R'$. In Fine (1999), an existence postulate of rigid embodiments is also presented. This principle states that whenever a, b, c, ... exist and a certain relation R applies to them, then the rigid embodiment $a, b, c, \ldots /R$ exists. This principle has been criticised in particular by Uzquiano (2018). We notice that, for groups, a principle as such may entail a dangerous proliferation of collective entities. Fortunately, for our treatment, we do not need to commit to this level of generality.

assigned to them. In our bowling example, we could imagine that a new role is created in the Sunday Afternoon Champions, that of 'results communicator', such that some of the responsibilities of the captain are passed to this new role, which can be played either by one of the current players, or by a new (human) agent, who is hired exactly to play this role, or by an artificial agent.

Therefore, both the structure (i.e. R) and the membership (i.e. a, b, c, ...) of a collective may undergo change and the change through time is captured by the functional principle F. For this reason, the notion of variable embodiment provides the suited abstract understanding of collectives in time. Thus, we can rephrase our definition of embodied collectives of Sect. 3.4.2, by viewing them diachronically as variable embodiments. In this case, the structural aspect of the collective is more complex, as it depends on time. The functional principle F is what enables us to select the right structure R for each given time.

The question naturally arises of what makes it the case that one and the same collective survives through a series of changes. At the abstract level, we can see what element of Fine's theory does this: it's the functional principle F that identifies the collective throughout its life, by picking up the different rigid embodiments that manifest the collective at the different times in which it exists. However, this answer is too close to a truism: the principle F simply states that the collective exists at all the times it is manifested. So, in Fine's theory, the question of the persistence of collectives simply boils down to deciding which is the correct principle of variable embodiment for the collective at issue. For instance, deciding whether our bowling team survives the replacement of Charles with Carl amounts to choosing two distinct principles of variable embodiment: a principle F that makes the collective persist or a distinct principle F' that makes the collective cease to exist after the time of the replacement.

Fine offers no details on how to decide the best principles, on what such principles might look like, or on how they are capable of gluing a number of time-slices of collectives together in a reasonable way, in spite of the fact that those manifestations might differ from one another in pretty much any respect.[43] But fortunately, an interpretationist view of variable embodiments can help us understand the selection of the functional principle and the continuity of collectives.

3.4.5 An interpretationist view of variable embodiments

In introducing our view of embodied collectives, we argued that it is the structure R that instructs the radical interpreter on how to elicit the evidence and actions for the collective. As we argued, the interpreter has to select among candidate structures the one that makes available to it a set of evidence and actions that in turn allows it to best give a rationalising interpretation of the collective. This treatment provides, so to speak, a time-independent abstraction of the work of the interpreter. When considering the interpreter's endeavour diachronically, the task of the interpreter becomes more complex. In this case, we have to postulate the collective as a variable embodiment, manifesting possibly radical changes in its relational structure and its members

[43] The issue of understanding the principle F for social groups is also discussed, and a solution proposed, in Ferrario et al. (2018).

throughout its history. Therefore, to single out the correct evidence and action at each time, the interpreter has to rely on the information encoded in the principle F of variable embodiment, which provides, at each time, the correct structure R. That is, the task of the radical interpreter is now to select among the candidate functional principles of variable embodiment F those that provide a chronicle of evidence and actions that makes the collective, through time, maximally rational.

It is here that the interpretationist has a plausible answer to slot into the theory, at least when the collectives are collective agents, to the open question on how to select the best principles of variable embodiment.

It is the radical interpreter who selects the principle of variable embodiment for a series of time-slices of collectives that, taken together, yield a collective which maximises diachronic rationality. For instance, if our bowling team after the replacement of Charles keeps on behaving as a 'rational' bowling team, pursuing goals, playing matches, tallying up scores, etc., we have reasons to believe that the collective is persisting. This shall be reflected in the choice of a principle of variable embodiment F that makes the bowling team persist after the replacement of Charles with Carl.

This is, in effect, an analogue of a psychological continuity view of personal identity on the individual level. Since the radical interpreter as matter of course ascribes structure to collectives in a rationality-maximising way, the view naturally produces an answer to the continuity question that arises for the theory of collectives as variable embodiments. To be sure, the interpreter would not thereby be settling entirely the contours of the collective's functional principle; there would be various ways in which we can imagine the principle to be varied that would not impact the overall rationality of the collective agent that emerges. But the interpreter nevertheless places a very substantive constraint on the choice of the functional principle in this way.

4 Final remarks

We've done our best to draw attention to some aspects of collective intentionality that, we think, haven't been given enough weight in theorising. We have laid out the architecture of a theory that, besides being generally plausible, accommodates those comparatively neglected aspects of collective intentionality. It's a fairly ambitious undertaking, and accordingly we have only been able to fill out the theory at a certain level of detail. But what we've articulated will hopefully suffice to show that there is an interesting and promising approach here which deserves further attention and development.

Acknowledgements Thomas Brouwer gratefully acknowledges the support of the European Research Council; specifically, the project leading to this publication has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No. 818633). Roberta Ferrario gratefully acknowledges support from the University of Leeds through the non-stipendiary visiting fellowship scheme, as part of this contribution was written during the visiting.

References

Arrow, K. (1951). *Social choice and individual values*. New Haven, CT: Yale University Press.

Bratman, M. E. (1999). *Faces of intention: Selected essays on intention and agency*. Cambridge: Cambridge University Press.

Dietrich, F., & List, C. (2010). The aggregation of propositional attitudes: Towards a general theory. *Oxford Studies in Epistemology*, *3*(215), 34.

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

Epstein, B. (2016). A framework for social ontology. *Philosophy of the Social Sciences*, *46*(2), 147–167.

Ferrario, R., Masolo, C., & Porello, D. (2018). Organisations and variable embodiments. In S. Borgo, P. Hitzler, & O. Kutz (Eds.), *Formal ontology in information systems, proceedings of the international conference FOIS 2018* (Vol. 306, pp. 127–140). Amsterdam: IOS Press. Frontiers in Artificial Intelligence and Applications.

Fine, K. (1982). Acts, events and things. In W. Leinfellner, E. Kraemer, & J. Schank (Eds.), *Language and ontology, Proceedings of the Sixth International Wittgenstein Symposium 23rd to 30th August 1981, Kirchberg/Wechsel, Austria* (pp. 97–105). Vienna: Hölder-Pichler-Tempsky.

Fine, K. (1999). Things and their parts. *Midwest Studies in Philosophy*, *23*(1), 97–105.

Fine, K. (20XXa). Acts and embodiment. In A. Hinshelwood (Ed.), *Being and doing*. Oxford: Oxford University Press.

Fine, K. (20XXb). The identity of social groups. *Metaphysics*.

Gilbert, M. (1989). *On social facts*. Princeton, NJ: Princeton University Press.

Harris, K. (2020). How individuals constitute group agents. *Canadian Journal of Philosophy*, *50*(3), 350–364.

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

Lewis, D. K. (1974). Radical interpretation. *Synthese*, *27*, 331–344.

List, C. (2014). Three kinds of collective attitudes. *Erkenntnis*, *79*(9), 1601–1622.

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

List, C., & Puppe, C. (2009). Judgment aggregation: A survey. In *Handbook of rational and social choice*, Oxford: Oxford University Press.

Ludwig, K. (2014). Proxy agency in collective action. *Noûs*, *1*(48), 75–105.

Masolo, C., Vieu, L., Bottazzi, E., Catenacci, C., Ferrario, R., Gangemi, A., & Guarino, N. (2004). Social roles and their descriptions. In D. Dubois, C. Welty, & M. Williams (Eds.), *KR 2004*: *Proceedings of the 6th international conference on the principles of knowledge representation and reasoning, June 3–7, 2004, Whistler, British Columbia, Canada* (pp. 267–277). Palo Alto, CA: AAAI Press.

Masolo, C., Guizzardi, G., Vieu, L., Bottazzi, E., & Ferrario, R. (2005). Relational roles and qua-individuals. In G. Boella, J. Odell, L. van der Torre, & H. Verhagen (Eds.), *AAAI fall symposium on roles, an interdisciplinary perspective, November 3–6, 2005, Hyatt Crystal City, Arlington, Virginia* (pp. 103–112). Palo Alto, CA: AAAI Press.

Neander, K. (2017). *A mark of the mental: In defense of informational teleosemantics*. Cambridge, MA: The MIT Press.

Porello, D. (2018). Logics for modelling collective attitudes. *Fundamenta Informaticae*, *158*(1–3), 239–275.

Porello, D., Bottazzi, E., & Ferrario, R. (2014). The ontology of group agency. In P. Garbacz & O. Kutz (Eds.), *Formal ontology in information systems, proceedings of the international conference FOIS 2018* (Vol. 267, pp. 183–196). Amsterdam: IOS Press. Frontiers in Artificial Intelligence and Applications.

Quinton, A. (1975). Social objects. *Proceedings of the Aristotelian Society*, *76*, 1–27.

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

Ritchie, K. (2020). Social structures and the ontology of social groups. *Philosophy and Phenomenological Research*, *2*, 402–424.
Searle, J. R. (1995). *The construction of social reality*. New York: The Free Press.
Tollefsen, D. (2002). Organizations as true believers. *Journal of Social Philosophy*, *33*(3), 395–410.
Tollefsen, D. (2006). From extended mind to collective mind. *Cognitive Systems Research*, *9*(2), 140–150.
Tuomela, R. (1995). *The importance of us: A philosophical study of basic social notions*. Stanford, CA: Stanford University Press.
Tuomela, R. (2000). *Cooperation: A philosophical study*. Dordrecht: Kluwer Academic Publishers.
Uzquiano, G. (2018). Groups: Toward a theory of plural embodiment. *Journal of Philosophy*, *115*(8), 423–452.
Varzi, A. C. (2006). A note on the transitivity of parthood. *Applied Ontology*, *1*(2), 141–146.
Williams, J. R. G. (2016). Representational scepticism: The bubble puzzle. *Philosophical Perspectives*, *30*(1), 419–442.
Williams, J. R. G. (2020). *The metaphysics of representation*. Oxford: Oxford University Press.
Wray, B. K. (2001). Collective belief and acceptance. *Synthese*, *129*(3), 319–333.