Rational golems: collective agents as players in the reasoning game

Javier González de Prado (UNED) and Jesús Zamora-Bonilla (UNED)

*This is a final draft of a chapter to appear in: Ladislav Koreň, Hans Bernhard Schmid, Preston Stovall and Leo Townsend (eds.), *Groups, Norms and Practices: Essays in Inferentialism and Collective Intentionality*, Springer. Please refer to the published version when possible.

Abstract: We ordinarily attribute beliefs and other intentional states to collective entities. These attributions can be vindicated from a theoretical perspective that holds that: i) collective entities can behave as rational agents in our argumentative practices of giving and asking for reasons; and ii) attributions of beliefs are interpretative tools aiming to make sense of the behaviors and perspectives of agents, and to keep track of their commitments. However, it is not immediately clear that groups have attitudes that play the same role in our argumentative and interpretative practices as the beliefs of individuals. It seems that the belief-like attitudes attributed to collective agents lack some of the distinctive features of the beliefs of individuals. More specifically, collective (but not individual) agents seem to be able to form beliefs against the available evidence at will, moved by rewards or other practical considerations. This has made some authors (Wray 2001, 2003; Meijers 2002, 2003, Hakli 2006) argue that the belief-like attitudes of collectives are actually some form of acceptance (which does not involve the commitment to truth characteristic of belief). On our view, this proposal is misguided. We shall argue that the belief-like attitudes of collectives are actually subject to the norms that govern belief, even if these norms can be broken more blatantly in the case of groups than in the case of individuals. In order to explain why this is so, we appeal to the fact that doxastic deliberation

in groups is mediated by the deliberative actions of the relevant members, which may be rationally motivated by practical considerations.

1. Collective believers

It is part of our ordinary practice to attribute beliefs and other cognitive states to collective entities, such as firms, sport teams, governments or institutions.[1] Speakers often say things like 'Microsoft believes we are on the threshold of quantum computation' or 'The government is certain that the economic situation will improve during the next year.' On the face of it, these attributions may seem puzzling. After all, groups lack the sort of phenomenology and neural structure characteristic of individual human believers. One could expect such phenomenology and biological constitution to be essential aspects of what it is to be a believer. Can we then vindicate our attributions of collective beliefs, or should we rather regard such attributions as simply metaphorical or loose talk?

We think that an independently attractive functionalist-pragmatist account of belief attribution makes it possible to justify our ascriptions of beliefs (and other cognitive states) to collectives. According to this approach, having a belief is a matter of being in a dispositional state that plays a given role in a suitable practice – in particular, in relation to action, reasoning and language use (see Brandom 1994; Sellars 1956). Having a belief that *p* will amount to being disposed to engage in certain patterns of action, speech and reasoning. For instance, an agent believing that it is raining will be typically disposed to open her umbrella (if she is outdoors and has the goal to remain dry), to assert 'It is raining' (if it is conversationally relevant) and to infer

[1] The notion of collective belief has been defended, among others, by Gilbert (1987, 1989, 1994, 1996, 2002) and Tollefsen (2002, 2015). For further discussion, see List (2005); List and Pettit (2011); Lackey (2016); Kallestrup (2016).

that the streets will be wet. When we attribute to an agent the belief that it is raining, we would be ascribing to her these types of dispositions. Note that we have not mentioned the phenomenology of agents or their brain states. As long as the agent has suitable behavioral and linguistic dispositions, it will make sense to attribute beliefs to her, regardless of her underlying psychological constitution. In this way, it may be in principle appropriate to attribute beliefs to collective entities, provided that they manifest the relevant dispositions.

On our preferred view, defended at length by Brandom (1994, 2000), the dispositions characterizing what it is to be a believer are norm-governed. More specifically, attributions of beliefs take place within normative practices of giving and asking for reasons. Beliefs are characterized by their role as premise-states of pieces of reasoning. Believing that *p* involves a commitment to such a proposition and to its inferential implications (so that one cannot correctly believe propositions logically or materially incompatible with *p* while correctly believing *p*). In turn, correctly believing that *p* puts one in a position to correctly believing its inferential implication *q*, if one transitions to that belief by properly reasoning from one's correct belief that *p*. Likewise, when one correctly believes the premises of good pieces of practical reasoning, one is in a position to correctly undertake the course of action recommended by the conclusion of such reasoning. In public discourse, we typically express our beliefs by asserting their contents. So, in public argumentation an agent may justify her belief that *q* by asserting a further proposition *p* that she correctly believes and that entails *q*. Similarly, an agent may justify her actions by citing considerations that she correctly believes and that are the premises of good pieces of practical reasoning recommending those actions. This is the standard way of providing reasons for one's attitudes and actions. As we will conceive of them here, normative reasons are considerations that count in favor of some attitude or response, and to which agents may appeal in order to justify such attitude or response (i.e. to vindicate its correctness). [2]

[2] See Parfit 2011; Scanlon 1998; Raz 1975.

Paradigmatic believers are competent players of justificatory games of demanding and offering reasons. By attributing beliefs to agents, we make sense of their moves in such normative games (i.e. we rationalize the agent's behavior). Thus, having beliefs requires being capable of participating in certain normative practices, and not so much being in any specific phenomenological or neural state.

Our proposal is that it makes sense to attribute beliefs to groups that are sufficiently responsive to reasons, i.e. groups that are proficient players in a normative game of giving and asking for reasons (see González de Prado Salas and Zamora-Bonilla, 2015). Note that what is crucial is not that there exist rational deliberations among the members of the group, but that the group itself as a unity, as a collective agent, can take part in relevant argumentative practices and engage in exchanges of reasons with other agents, so that its behavior and attitudes become evaluable as reason-sensitive.

An important intuition behind our proposal is that collective agents are a type of *artificial entities*, something that humans create in order to achieve goals that would be otherwise difficult to attain, but that, as all other artefacts, may end having consequences that were not expected by their creators at the beginning, and in some cases may end having something like a 'life of their own', in the sense that they can be taken as emergent beings that, up to a point, behave in a way not totally controllable by the individuals. To say it with a classical image, the collective agent would be a kind of *golem*. As all artificial beings, they can function better in some cases and worse in others; they can be the result of a deliberate planning, or they can emerge and evolve more or less spontaneously; they may have very different kinds of forms, and they are irreducible to a simple set of principles that can be summarised in one clean formula (that is why we renounce to offer a precise 'definition' of collective agent, collective

belief, or the like).[3] The essential feature of a rational collective agent is that *it can function as a subject in a public deliberative practice*, i.e. that it can be regarded as responsive to reasons, so that it makes sense for other agents to attribute to the group commitments to certain propositions and to their inferential implications.

So, the idea we have in mind is that intentional attitudes may be attributed to groups that are able to play the role of competent participants in argumentative conversational practices. If we can engage in a conversation with a collective entity, then it makes sense to attribute beliefs and other attitudes to it. A modification of the classic Turing test might be helpful here. Imagine that you are performing what superficially looks like an instance of the Turing test (i.e., you are passing questions to some kind of interface, and receiving some answers from it), but in which your task is not to discriminate whether the answers are being produced by a human being or by a machine, but to tell whether these answers are produced by an individual or by a group (acting as a genuine collective agent, and not merely expressing the views of an Arrowian dictator). If there are groups capable of passing this *Collective-Turing test* (i.e. groups that cannot be distinguished by their answers from an individual interlocutor in the context of a Turing test-like process), then it is possible for groups to behave as collective agents with beliefs and other intentional attitudes – although, of course, passing the test is only a *sufficient* condition for being a collective agent, not a necessary one.

It is important to stress that we are *not* assuming that there is *always* a rule-like algorithm (for example, an 'aggregation mechanism') that takes as inputs the beliefs of the

[3] We also want to remain neutral about whether the constitution of a collective agent requires that its members adopt attitudes in the we-mode (Tuomela 2013; also Schmitz 2017). In general, we wish to be as non-committal as possible about how the members of a group must interact in order for the group to behave as a collective agent (i.e. as a proficient player in normative practices of giving and asking for reasons). It may well be that there is not a unique answer applying to all types of collective agent.

members of the group and produces as an output the groups' collective attitudes.[4] There may be some groups whose belief-formation procedures are structured according to explicit algorithmic rules, for instance some voting rule. Actually, we will sometimes appeal to such groups as simple toy examples. But surely collective agents will not always include such rule-like decision mechanisms. In particular, not all groups function by making their members *vote* on each proposition the group has to take a stand on (in the same way that the curves of a road are not designed by making to vote all the engineers, workers and politicians engaged in the project). More frequently, the collective agent *will be made to hold* a certain view as the result of some process of *deliberation* by its members, a process that can be more or less explicitly regimented, but that is certainly not always reducible to a voting rule, and in the course of which the members themselves will often change their own individual opinion, even if it does not coincide (not even in the end) with that of the group.

Furthermore, many collective agents (which, remember, are a kind of *tool*) exist with the aim of advancing *epistemic goals* that are difficult to attain by means of uncoordinated individual actions, both thanks to the division of cognitive labour (e.g. Kitcher 1990; Weisberg and Muldoon 2009; Wagenknecht 2016) and to the overcoming of individual biases (e.g. Mercier and Sperber 2017). In this way, what we will often observe is that epistemic collective agents tend to be *more rational* than their individual members (as far as *those specific epistemic goals* are taken into account). In many cases, the point of creating collective agents is not to 'preserve' the rationality presupposed in each individual's system of beliefs (as assumed by much of the

[4] There are numerous mathematical proofs, in the judgment aggregation literature, showing that, given some reasonable assumptions, it is impossible to aggregate profiles of rational individual beliefs in a completely coherent way (see List and Pettit 2011).

theoretical work on judgment aggregation), but rather to overcome the shortcomings of the cognitive capacities of the individual members.

One last clarification is that the constitution of the group as a collective agent *does not eliminate* the cognitive agency and autonomy of its members as individuals. Although groups can attain some epistemic goals that are beyond the reach of individuals (or that these can attain less efficiently), the collective agent's attitudes are in the end derivative of the epistemic agency of its members and of the processes of deliberation that take place among them. Hence, the constitution of the collective agent's belief does not imply that disagreements between its members cease to exist, nor that these are forced to renounce their own beliefs *as autonomous individuals*. The collective agent will simply be an *additional* subject in the public argumentative practices in which it is entitled to take part, and it will have *its own normative position* in such public practices. In principle, there may be disagreements among the beliefs of the collective agents and the beliefs of some (even most) of its members. And, in many cases, the dissenting members of a collective agent will have the freedom to express their dissenting voices and to act on the basis of them, at least when they do not act as public speakers or representatives of the group. These considerations will be crucial in our discussion below.

In the last section of the paper, we shall examine some examples that will further illuminate the picture of collective agency that we endorse. However, our aim here is not to offer a full argument for this sort of picture (we have done so elsewhere, see González de Prado Salas and Zamora-Bonilla, 2015). Rather, we want to discuss a specific objection to the idea that collective entities can play the role of believers, that is an objection to the idea that collective entities can actually take part competently in the types of practices that license belief ascription. It has been argued that the collective attitudes of groups can be at most acceptances, but not beliefs, because such attitudes are not subject to some of the rationality constraints that define what it is to believe. In particular, it seems that the doxastic attitudes of groups can be adopted

voluntarily, do not aim solely at truth and can be overtly influenced by pragmatic factors. One can object that an attitude with these features does not behave as belief does, so it should be classed as a different type of attitude, perhaps an acceptance (Wray 2001, 2003; Meijers 2002, 2003; Hakli 2006; for discussion, Gilbert 2002; Tollefsen 2002; Mathiesen 2006). Our aim here is to resist this line of argument. We will argue that groups may be subject to the norms and rationality requirements distinctive of belief. In the next section we discuss the rational constraints characteristic of belief and the norms that may underlie them. After that, we show that the attitudes of collectives can actually be governed by the norms of belief, even if it is true that in the case of collectives these norms may be broken in a more blatant way than in the case of individuals. In the rest of the paper we explore the sources of these differences between individual and collective belief.

2. Rational belief and practical considerations

Belief is distinctively connected with truth and evidence. Williams (1973:148) tried to capture these distinctive connections by claiming that belief aims at truth.[5] In general, it is difficult to form beliefs that you take to be false, or that go against the evidence available. If you see that it is raining, it is hard to get yourself to believe that it is not raining. Relatedly, we do not seem to be able to form beliefs at will, in the sense that we cannot decide to form a belief just because

[5] For discussion of this idea and different ways of developing it, see among others Wedgwood (2002); Shah and Velleman (2005); Owens (2003); Steglich-Petersen (2006); Boghossian (2008); McHugh (2014b); Whiting (2010).

it would be desirable or useful to do so.[6] As has been noted in the literature, normal agents are not usually in a position to form beliefs motivated merely by practical incentives to do so (Williams 1973; Shah 2003; Shah and Velleman 2005; McHugh 2012, 2014a). Imagine that a powerful genie offers you a great reward if you form a belief that is not supported by your evidence (say, the belief that London is the capital of Italy). No matter how great the reward is, for normal agents it would seem impossible to do as instructed by the genie and voluntarily form the requested belief. In general, there is not a straightforward route from desiring to form a belief to getting to form it.

The apparent impossibility of forming beliefs as a response to practical incentives suggests that in doxastic deliberation only epistemic considerations can be recognized as reasons to believe (where epistemic considerations are considerations that concern the truth of the relevant proposition). Practical considerations having to do with the desirability of forming the belief that *p* cannot weigh in the deliberative process as reasons to believe that *p* (Shah and Velleman 2005; Hieronymi 2008). Of course, practical factors may have a causal impact in the formation of beliefs, for instance in the form of wishful thinking or biases (e.g. confirmation bias). Likewise, practical considerations may move the agent to try to put herself in a position where she is likely to form the required belief (say, she can undergo hypnosis, experiment with different forms of self-suggestion or expose herself to being brainwashed). Yet in general it is not easy to make you form a belief that you take to be unsupported by evidential considerations – you can only do so in indirect, roundabout ways. It seems that, when we engage in explicit

[6] This does not mean that we lack control over our beliefs. It can be argued that beliefs are under rational control, insofar as we form them in response to the reasons available to us (McHugh 2012, 2014a; Hieronymi 2008).

deliberation about what to believe, we cannot effectively treat practical considerations as reasons to believe some proposition (Williams 1973).

Why is the deliberative formation of beliefs not responsive to practical reasons, at least in an overt, self-aware way? A possible line of response is that doxastic deliberation is transparent, in the sense that deliberating about whether to believe *p* automatically gives way to deliberation about whether *p* (that is, about whether *p* is true).[7] According to Hieronymi (2008), beliefs about *p* embody the agent's answer to the question whether *p*. Thus, deliberating about what to believe in relation to *p* amounts to considering what is the correct answer to the question 'p?'.[8][9] Arguably, the correctness of an answer to the question 'p?' hinges on the truth of *p*. Therefore, only evidential considerations bearing on the truth of *p* are relevant for settling this type of question. In this way, to the extent that the belief that *p* is an answer to the deliberative question 'p?', it is only correct if *p* is true (see Whiting 2010; McHugh 2014b). This would be why only epistemic considerations can be openly recognized as reasons to believe when engaging in doxastic deliberation. Moreover, the existence of a truth-involving standard of correctness governing beliefs would explain why beliefs that happen to be formed in a way that disregards the evidence (say, because of some recalcitrant prejudice) are assessed as incorrect – as somehow epistemically defective.

To be sure, the difficulty to form beliefs voluntarily or against the evidence available might be taken as a mere psychological fact, or at any rate a fact that is not to be accounted for in normative terms (see Bykvist and Hattiangadi 2007; Glüer and Wikforss 2013). However, as

[7] For proposals along these lines, see Shah and Velleman (2005); Hieronymi (2008).

[8] Take 'p?' as shorhand for a grammatically correct formulation of the question whether *p*.

[9] Indeed, as a matter of psychological fact, individuals hardly ever deliberate asking explicitly to themselves 'Should I believe that *p*?', rather than simply '*p*?'. The sort of transparency we are discussing would explain why this is the case.

Con formato: Fuente: Sin Cursiva
Con formato: Fuente: Sin Cursiva
Con formato: Fuente: Sin Cursiva
Con formato: Fuente: 11 pto
Con formato: Fuente: 11 pto, Cursiva
Con formato: Fuente: 11 pto
Con formato: Español (España)

the last paragraph shows, a natural explanation of why we can only believe for epistemic reasons (and not for practical ones) appeals to some truth-involving normative standard governing belief – for instance, the norm that beliefs are correct, appropriate or permissible only if true. According to the explanation given above, when deliberating about what to believe regarding *p*, agents are trying to answer the question whether *p*. Insofar as answers to this question are governed by a truth-involving standard of correctness, only evidential considerations can properly settle it. In what follows, we assume that this sort of truth-involving normative standard is characteristic of the attitude of believing.[10]

3. **The atypical features of collective belief**

Many of the distinctive features of individual belief discussed in the previous section are not found in the belief-like attitudes attributed to collective entities (for ease of exposition, we will keep calling such attitudes collective beliefs, suspending judgment for the moment about whether they constitute genuine beliefs or not). It seems that practical considerations can affect the formation of collective beliefs in a far more direct, widespread and flagrant way than what is observed in the case of individual believers. In particular, there is no obvious barrier to the adoption of collective beliefs as a response to practical incentives – in this sense, collective beliefs seem to be voluntary in a way that individual beliefs are not (Wray 2001, 2003; Meijers 2002, 2003, Hakli 2006).

[10] Most of the main points we want to make could also be made in terms of weaker norms that do not require beliefs to be true, for instance the evidentialist norm that one ought to adjust one's beliefs to the evidence available (see Conee and Feldman 2004). We will focus on truth-involving norms, for the sake of simplicity, and because of their plausibility.

Imagine, as a simple toy example, a group where majority rule is the mechanism to decide what collective belief is to be adopted: the members of the group vote in favour or against adopting the collective belief that *p*, and the option supported by more votes is chosen. Now, in principle all kinds of considerations can motivate the members of the group to vote one way or another. Perhaps they are only moved by epistemic considerations, but they can also have in mind practical considerations, such as the potential benefits of getting the group to adopt the belief that *p*. In this way, the members of the group can (rationally) decide to vote in favour of the collective belief that *p*, even if as individuals they believe that ¬*p* (and take the evidence to support such a belief). If the members of this type of group are offered a reward for making the group adopt the collective belief that *p*, they can easily comply: they just need to vote accordingly.

The lesson of the example does not depend on the specific details about the group's decision mechanisms. Something similar could happen in a group where decisions about what to (collectively) believe are reached via internal deliberations among the members (the members argue with each other until they agree on adopting a certain collective doxastic attitude). Again, the members may be moved by practical considerations when deciding what position they will defend in this intersubjective deliberative process – they may decide to advocate a view that they know is actually false and goes against the available evidence.

It may be argued that these examples only show that the collective beliefs of a group can be voluntarily controlled by its members – so that the practical concerns of the members can play a direct role in the adoption of the collective belief. Yet the examples would not show that the group, as a collective agent, can form beliefs at its own will, or that the group can recognize practical considerations as reasons to believe (see Gilbert 2002; Gilbert and Pilchman 2014). One first thing to say is that, even if this were so, practical factors (more specifically, the practical interests of the members) would still play a more direct and pervasive role than in the

case of individual beliefs. Moreover, there is no reason to think that there cannot be groups where beliefs are routinely formed directly in response to the group's desires and intentions to form them. We can imagine a group where there are two different decision mechanisms, the first one for forming collective desires to adopt collective beliefs, and the second one for adopting collective beliefs. It could happen that the outcomes of the first decision process serve as inputs for the second one, so that the (collective) desires to believe formed via the first decision process directly determine the collective beliefs produced as outcomes in the second decision process (the group just adopts the beliefs that it desires to adopt). Furthermore, it could be that, in the context of public justificatory practices, the group cites the practical considerations that backed its collective desires to believe as its reasons for forming the corresponding collective beliefs. In a case like this, the group could be said to be forming beliefs motivated by practical considerations concerning the desirability of forming such beliefs (which, of course, does not mean that these practical considerations are actually good reasons to adopt the relevant beliefs). Perhaps these types of examples are not best described as cases where the group recognizes practical considerations as reasons to believe (i.e. as considerations that favour the belief and make it fitting), but rather as cases where the group has practical reasons to desire to adopt a belief (i.e., a cognitive commitment), and manages to directly form such a belief in response to this desire. However, this would still constitute a significant difference with individual belief-formation, where it is not typically possible to go directly from a desire to form a belief to actually forming it.

The dissimilarities between individual beliefs and the belief-like attitudes of collectives may make one conclude that such collective attitudes are not actually beliefs, but some other type of attitude – for instance, some form of acceptance (Wray 2001, 2003; Hakli 2006; Meijers 2002, 2003). There are different ways of characterizing the attitude of acceptance, but for our purposes here it is enough to think of it as an attitude that is similar to belief in some of its inferential and practical implications, but that only involves endorsing its content in certain

restricted contexts and for certain specific purposes.[11] In this way, a lawyer may accept the innocence of her client only in the context of defending him, but not in when she is not acting as a lawyer. Likewise, a scientist may accept some claim in the context of a certain research project, perhaps due to its predictive success, even if she remains neutral about its truth (or even if she knows the claim to be strictly false, although perhaps a good approximation). As these examples show, accepting a claim is compatible with not believing it. Moreover, in general an agent can decide at will whether to accept some claim (for certain purposes). After all, we can accept a claim merely for the sake of argument in order to see what follows from it. And scientists may accept some view as a research assumption motivated by the funding prospects associated with undertaking a research project presupposing such a view.

In the contexts where the agent's acceptance of *p* is in play, her behaviour may be similar in many respects to that of an agent that believes *p*. So, the agent may be disposed to assert that *p*, to offer support for *p* in the face of appropriate challenges and to act as if *p* were true. If this is so, it could seem that we can properly characterize the behaviour of a group with a belief-like attitude towards *p* by attributing to it an acceptance of *p*, rather than a belief that *p*. In this way, there would be no problem in granting that the attitude of the group towards *p* may be adopted at will, on the basis of practical considerations. In the next section, however, we argue that this is not a satisfactory approach. Unlike acceptances, the belief-like attitudes of groups are subject to the normative standards governing belief.

4. Acceptance and the norm of belief

As we have suggested above, it is plausible to think that belief is governed by a truth-involving normative standard, so that believing involves a context-independent commitment to the truth

[11] For discussion on the notion of acceptance, see Cohen (1992); van Fraassen (1980).

of the proposition believed. If you believe *p* but at the same time take it to be false, you are bound to have an incorrect attitude. Correctly believing that *p* is incompatible with correctly believing that *p* is false (or with correctly believing ¬*p*). This unrestricted commitment to truth is not found in the attitude of acceptance characterized in the previous section. An agent can always correctly accept *p* (in a certain context, for certain purposes) while leaving open the possibility that *p* is false, indeed even while believing that *p* is false – for instance, we may accept *p* as a useful approximation. In this way, an agent may correctly accept, in different contexts, each of two incompatible propositions (i.e. propositions that cannot be both true). For example, it may be that, when tackling a problem that does not require precise results, a scientist accepts some claim that works well as a simplifying approximation, whereas she accepts instead some more cumbersome non-simplified claim in the context of dealing with other problems where precision is paramount.

Now, if the belief-like attitudes of groups were to be modeled always as acceptances rather than genuine beliefs, we would expect groups to be able to adopt such attitudes towards incompatible propositions, in different contexts, without having to reject one of the attitudes as incorrect. Moreover, it would not be possible to say of groups that they accepts some proposition as true, although they actually believe it is false, like we can say of individuals – for groups would never really believe anything, they would only accept propositions in certain (but not necessarily all) contexts.

Individuals can make a distinction between what they believe and what they merely accept,

Individuals can make a distinction between what they believe and what they merely accept, and the pragmatic-inferential norms they are publicly subjected to when forming and revising their beliefs are noticeably different from the norms they must follow when managing mere acceptances. What we want to argue is that this pragmatic difference between the norms

governing beliefs and the norms governing mere acceptances also applies to group agents. To be clear, we are not arguing that groups cannot accept propositions as true without believing them. In principle, groups will be able to do this, in the same way that individuals can. Our claim is that groups also adopt belief-like attitudes that are subject to normative constraints different from the norms operative for mere acceptances. ~~it should be possible for groups to adopt an 'acceptance-like' attitude towards *p* while acknowledging that *p* is actually false. Similarly, we would expect groups to be able to adopt 'belief-like' attitudes towards incompatible propositions, in different contexts, without having to reject one of the attitudes as incorrect. However, this does not seem to be the case. If a group acknowledges a belief-like attitude towards *p* and at the same time acknowledges having a belief-like attitude towards ¬*p* or '*p* is false' (perhaps in a different context, or with different purposes), the group will be criticizable as having knowingly done something incorrect. Groups that have inconsistent belief-like attitudes will be taken to be in a defective state that calls for revision, on pain of irrationality.~~

More specifically, the belief-like attitudes attributed to groups seem to be sensitive to the sort of truth-involving normative standards governing belief. We have argued that groups may adopt belief-like attitudes that go blatantly against the available evidence (perhaps moved by practical incentives). However, these attitudes will be assessed as incorrect and may be criticized by other participants in the relevant practices of giving and asking for reasons. Imagine, for instance, that the government expresses a belief-like attitude towards the claim that the unemployment rate has been reduced, when the evidence clearly indicates that it has gone up. Surely, the government's attitude will be criticizable, given its disregard for the available evidence. Moreover, the correctness of collective belief-like attitudes that ignore the evidence available will not be vindicated by citing practical incentives to form that attitude. Such practical considerations will not be taken to bear on the correctness of the group's belief-like attitude. Thus, the belief-like attitudes of groups that depart from the normative standards governing belief will be typically seen as violations of such standards.

In light of these observations, we can conclude that the belief-like attitudes of groups are indeed subject to the norms of belief. *What is characteristic of beliefs is not that they always satisfy such norms, but rather that they are always subject to them*. After all, these norms are often transgressed by the beliefs of individual agents, which nonetheless keep counting as beliefs – think for instance of cases of bias or wishful thinking. So, we submit that the belief-like attitudes of groups do not behave as non-doxastic acceptances, but rather play the same normative role as the beliefs of individual agents (see Mathiesen 2007). Therefore, we should not be wary of regarding such attitudes as collective beliefs. To be sure, there remains the question of why collective agents can flout the norms of belief in more overt, glaring ways than individuals. We discuss this issue in the next section.

5. Non-transparent collective doxastic deliberation

In section 2 we argued that the fact that we cannot easily form beliefs at will, moved merely by practical considerations, is related to the transparency of doxastic deliberation. The belief that *p* constitutes an answer to the question 'p?', so deliberating about what doxastic attitude towards *p* is correct amounts to deliberating about whether *p*. We want to suggest that things work differently in the case of collective beliefs. Typically, collective doxastic deliberation is not transparent, because it is mediated by the deliberative choices of the members of the group. This may make groups more vulnerable than individuals to the influence of practical considerations when forming their (collective) beliefs, constituting a possible source of epistemic failure against which individuals tend to be more protected by their natural psychological constitution.

In general, collective beliefs are formed as a result of the actions performed by some of the members in some decision-making process. A simple example, discussed above, is a group where members vote in order to decide what collective belief to adopt. But the actions of

members will also be involved in more sophisticated methods of collective belief formation. Think of a group where collective beliefs are adopted via a deliberative process in which the members argue for different positions until an agreement is reached. The resulting collective beliefs will depend on what positions the members decide to defend in the negotiation. Call the actions performed by members of a group in the context of collective decision-making processes *deliberative actions*. The crucial point is that these deliberative actions performed by the members may be rationally guided not only by evidential considerations about whether *p*, but also by further practical considerations concerning whether it is desirable that the group comes to believe that *p*. The question considered by the members of the group is not directly 'p?', but 'Is it desirable that the group believes that p?' So, in collective deliberations, the members will consider the question about what doxastic attitude is desirable for the group to adopt, and will act in ways that promote the group's adoption of such an attitude (e.g. they will vote for it). It may well be that the individual members have good (practical) reasons to foster a collective belief that they know is false or goes against the evidence, and that will be irrational at the collective level (for instance, in the sense that it will be incoherent with other things believed by the group). For example, some members of the group may have a practical interest in getting the group to act on certain beliefs they know to be false – say, because the political agenda of these members will be favored if the group acts on such beliefs.

Thus, it may be perfectly rational for the members of the group (as individuals) to try to manipulate the collective deliberative process so that the group adopts an irrational attitude – in particular, a belief that goes against the available evidence. Perhaps there is some limit to the amount of internal manipulation that a group can suffer before losing its coherence as unified, rational collective agent. Yet the point remains that deliberations within groups tend to be more exposed to the direct influence of practical considerations than in the case of individuals, given that they typically proceed via the deliberative actions of the group members.

The situation is analogous to cases where an individual agent has practical reasons to make other agents believe something false or against the evidence. An agent may have good practical reasons to lie, with the intention of inducing false beliefs in others (say, by lying the agent may be trying to mislead a killer searching for an innocent victim). Likewise, if you are a skilled hypnotist, you may have (practical) reasons to cause some agent to form a certain irrational belief, perhaps to avoid her emotional distress. As noted in section 2, an agent can also have practical reasons to try to influence her own beliefs in these sorts of ways. However, when the agent is *directly* deliberating about the correct answer to the question 'q?', practical factors can at most play an implicit causal role. And, in general, it is difficult for an individual agent to manipulate successfully her own beliefs against the evidence she acknowledges. By contrast, as we have seen, in principle there may be groups where the members directly manipulate the group's doxastic commitments, guided by their practical interests and going against the evidence that they (as individuals) acknowledge.

To be sure, in many cases the members of a group will have exclusively epistemic goals when engaging in collective doxastic deliberation, so that their deliberative actions will be guided by epistemic considerations (Mathiesen 2007). And when this is so, collective deliberation can afford epistemic goods that would not be available to isolated individual agents (as we will show in next section). What we want to argue here is that there may also be other cases where the deliberative actions of the members of the group respond to non-epistemic, practical considerations. When this happens, rational deliberative actions by the members of the group may lead to irrational attitudes at the collective level (say collective beliefs against the available evidence, or incoherent collective attitudes).

It should also be noted that in groups with defective decision-making procedures, the group may have blatantly irrational beliefs even if the agents perform rational deliberative actions guided only by epistemic considerations. For instance, it is well known that decisions by

voting may lead to different voting paradoxes. In a group where beliefs are formed by voting, incoherent collective beliefs may result from votes that reflect the members' coherent individual doxastic attitudes (see List 2005; List and Pettit 2011). Again, in these cases the flagrant irrationality of the collective belief has its roots in the non-transparency of the deliberative process through which such a belief was adopted. Regardless of whether the members have purely epistemic aims, the beliefs of the group are formed as a function of the views of other agents (the members), rather than as a result of a transparent deliberation in which the same agent that is to adopt the belief considers whether *p*. This may open the door to forms of irrationality that are less frequent in individual doxastic deliberation. Of course, it is to be expected that some collective agents will implement mechanisms aimed at protecting themselves from such risks of irrationality, to the extent that the rationality of these agents is an important factor for their success or their very survival. In general, successful collective agents will resort to strategies for limiting the impact of major sources of irrationality.

To sum up, group deliberation is non-transparent in a way that individual doxastic deliberation is not, given that it goes through the individual deliberation of the members about what deliberative action they should perform in order to influence the group's belief-formation in desirable ways. This explains why practical factors can play a more explicit and overt role in the formation of group beliefs.

6. The epistemic virtues of collective agency

We have argued that it is possible that the collective beliefs of a group are blatantly irrational and overtly break the norms governing belief (in a way not to be found in normal individuals), even if the attitudes of the members of the groups remain perfectly rational. However, this point should not obscure the fact that collective entities are often formed precisely to allow their members to improve their epistemic position, and to avoid epistemic shortcomings associated

with individual perspectives. An obvious way in which collective agents may go beyond the epistemic limitations of individuals is by pooling the evidence and epistemic skills of their members. However, plausibly there are other, less evident epistemic advantages associated with collective agency. We conclude the paper by offering a few examples of collective believers, with the aim of illustrating some of the potential epistemic benefits of collective agency.

The collective constituted by the co-authors of a scientific paper (in particular, of a two-authored paper) offers a clear example of an 'epistemic collective agent' providing access to epistemic rewards beyond the reach of individual agents (see Zamora-Bonilla, 2014). Co-authorship may be justified in many different ways, but the most common and obvious one is by appeal to the division of cognitive labour. In general, the collaborating authors will have different epistemic capacities and resources – each one may know more than the other about some of the topics that need to be tackled in the paper. Because of this, co-authorship often entails some degree of mutual trust between the authors: regarding some claims the paper makes, one author will be more capable of offering a competent defence than the other, even if the defence that is actually written is the result of a deliberation between both. This is more evident if we consider the paper not as a finished and perfect work, but as a single piece in a longer public conversation: if the paper is challenged in the future by other academics, the co-authors' responses need not be themselves co-authored (though they of course may), and obviously each co-author may end up offering different arguments or interpretations about what they wrote together.

However, as one of us recently argued, co-authorship can be shown to have epistemic advantages even in cases when the division of cognitive labour is not the main reason behind the collaboration (Zamora-Bonilla, 2014). This is particularly clear if we picture scientific inquiry as taking place within normative practices of giving and asking for reasons. Think of the scientific paper not as an isolated item of knowledge, but as an argument which is a portion of a longer

social process of deliberation, justification, criticism and contestation. In this type of social argumentative practice, past commitments (by some members of the scientific community) are appealed to in order to justify further commitments (not necessarily by the same individuals), and the credit (or symbolic reward) a certain scientist gets depends on the use that further researchers make of the items of knowledge that such a scientist has advanced or 'authored'. Assuming this sort of picture of scientific practice, it is plausible to think that many instances of co-authorship will be explained because the authors prefer to be recognised and rewarded by something that is a consequence of their individual contributions *taken together*, but that does not follow from each author's 'part' in isolation. Imagine, as a simple example, that author A has proved that *p* entails *q*, and author B has proved that *p*, but both prefer to be recognised as the joint discoverers of *q* (say, an important mathematical theorem), rather than as the individual discoverers of the truth of the less newsworthy claims '*p*' and '*p* entails *q*'.

Hence, the epistemic motivations for co-authorship do not only derive from the trivial fact that different individuals may have different, complementary capacities and resources, but from an important result about the cognitive division of labour: the fact that in general the 'solution' to some research question does not arise as a mere *juxtaposition* of single items of knowledge, but as an *inferential consequence* of the premises and arguments provided by the co-authors. The proposition *q* constituting the solution to the problem the co-authors wanted to solve is, obviously, believed by both A and B, but the 'social status' of that proposition (as an item of knowledge within the 'long conversation' in which the activity of their scientific discipline consists of) may be legitimately taken as something that 'belongs' to both co-authors as a single subject – i.e. it is the collective agent formed by A and B (A&B) who is entitled to get the credit for providing a proof for *q*. And, of course, since the acceptance of *q* by the rest of the scientists will mainly depend on the epistemic quality of the arguments offered in A&B's paper, A and B have a strong incentive to let the paper contain the soundest arguments they can elaborate, as

far as the practice of evaluating scientific claims within their scientific discipline is governed by strict epistemic norms.

The explicit appeal to collective beliefs in relation to co-authorship is evident when we consider a case in which the scientific community challenges and perhaps ends up rejecting A&B's conclusion – due to further evidence and arguments available after the publication of the paper by A&B. In cases like this, it would make sense to say: 'Even if A&B believed that *q*, we now know that *q* is false'. As we argued in section 1, belief ascriptions in this type of context should not be seen as making reference to any phenomenological or neural state, but rather as a way in which an interpreter keeps track of the commitments of (individual or collective) agents, without the interpreter having to undertake such commitments.

Co-authorship teams can be seen as the simplest collective unit of scientific agency. At the other end of the spectrum we find the entity constituted by a whole scientific discipline. Disciplines, we think, offer a further example of how collective epistemic practices afford potential epistemic rewards not available to individuals working in isolation.

One may be initially doubtful that scientific disciplines behave as unified collective agents – at least, this is not as clear as in the case of co-authorship. In particular, disciplines as such will not always be able to undertake commitments and defend them in cohesive ways in argumentative social practices. However, there is a sense in which scientific disciplines come closer to constitute a collective agent. Disciplines sometimes function as custodians of the consensus views emerging in the field and of the arguments backing those views. Indeed, it may be argued that disciplines may be legitimately taken as the real depositaries of the scientific knowledge in an area of science, in the sense that they have the highest authority when some item of knowledge belonging to that area is appealed to or discussed in the rest of society (and, to begin with, in other scientific disciplines). After all, it may well be that no individual scientist

has the means to acquire herself all the knowledge of the discipline and to master the different epistemic skills and resources underpinning such knowledge.

As opposed to what happened with teams of co-authors, in the case of disciplines the boundaries of the group are not even well defined: who is a legitimate member of the group, and how much voice a given member has in the determination of the group's positions, are often questions without definite answers. But that is only a problem for judgment aggregation views of collective attitudes, which reduce the formation of a collective belief to a set of voting rules or other algorithmic formal procedures. Our approach to collective agents does not presuppose any of that formal machinery (except in those cases where the group explicitly adopts it), but allows us to think of a collective agent as an entity defined by the normative position it occupies in a practice of giving and asking for reasons. In this sense, the scientific discipline (organised through more or less fluid structures like associations, congresses, textbooks and syllabi) defines, both for its members and for the outsiders, what claims and arguments must be taken as valid, which ones as merely conjectural, which others as admissible-but-not-compulsory etc. It also determines when an individual is allowed to act as a spokesperson for the scientific field (i.e. when an individual counts as a qualified expert). And, though some positions that are endorsed by the scientific discipline will be endorsed by all or by most of its members, it will often be the case that, due to the division of labour within it, many things that 'are known', are really unknown by most of the individual scientists belonging to the field (because they lie outside of their more specific area of specialization) – not to mention those cases in which, though the community accepts some claim, it also allows some degree of dissent about it or about its correct interpretation.

Scientific disciplines illustrate the potential epistemic advantages of engaging in collective scientific practices. Individual researchers clearly benefit epistemically from working within the context of consolidated scientific disciplines. In addition to providing scientists with

access to evidence, arguments and epistemic resources beyond the reach of a single isolated researcher, the intersubjective networks constituting disciplines may also function as a tool for correcting and limiting some forms of bias and irrationality that are difficult to avoid for individuals. Think for instance of the phenomenon known as confirmation bias. Researchers tend to give more weight and pay more attention to evidence that confirms their views. However, within a certain scientific discipline there will typically be researchers defending alternative positions. So, even if a researcher favouring a view *p* may tend to disregard evidence against that view, there will be other researchers that defend ¬*p* and will therefore be inclined to highlight evidence against *p*. Thus, it can be expected that in the sort of public intersubjective disputes that underlie the formation of disciplines and the emergence of consensus views, the potential confirmation biases of different individual researchers will be checked and corrected. Indeed, confirmation bias may even become epistemically advantageous in certain collective contexts, in that it may lead to a beneficial form of division of cognitive labour. Agents would specialize in finding evidence supporting the claims they defend, which will be then scrutinized and criticized by other agents defending alternative claims.

Thus, the resulting views collectively regarded as the established position in the discipline will be free of some of the biases and sources of irrationality that tend to affect individual views. This is not to say that some biases present at the individual level may not be reflected at the collective level as well, or that collective scientific practices may not introduce their own form of bias. Yet it seems that collective scientific interactions will help scientists avoid some of the potential epistemic limitations of individual research.

Bibliography

Boghossian, P. (2008). *Content and Justification*. Oxford: Oxford University Press.

Brandom, R. (1994). *Making it explicit*. Cambridge, MA: Harvard University Press.

Brandom, R. (2000). *Articulating reasons*. Cambridge, MA: Harvard University Press.

Bykvist, K. & Hattiangadi, A. (2007). Does thought imply ought? *Analysis,* 67, 277-285.

Cohen, L.J. (1995). *Essay on belief and acceptance*. Oxford: Oxford University Press.

Collins, S. (2013). Collectives' duties and collectivization duties. *Australasian Journal of Philosophy,* 91(2), 231-248.

Collins, S. (2017). Duties of group agents and group members. *Journal of Social Philosophy,* 48(1), 38-57.

Gilbert, M. (1987). Modelling collective belief. *Synthese,* 73(1), 185-204.

Gilbert, M. (1989). *On Social Facts*. Princeton: Princeton University Press.

Gilbert, M. (1994). Remarks on Collective Belief. In F. Schmitt (ed.), *Socializing Epistemology: The Social Dimensions of Knowledge* (pp. 111–134). Lanham, MD: Rowman and Littlefield.

Gilbert, M. (1996). *Living Together: Rationality, Sociality, and Obligation.* Lanham, MD: Rowman and Littlefield.

Gilbert, M. (2002). Belief and acceptance as features of groups. *Protosociology,* 16, 35-69.

Gilbert, M. & Pilchman, D. (2014). Belief, acceptance, and what happens in groups. In J. Lackey, (ed.), *Essays in Collective Epistemology*. Oxford: Oxford University Press.

Glüer, K. & Wikforss, Å. (2013). Against belief normativity. In T. Chan (ed.), *The Aim of Belief* (pp. 80-99). Oxford: Oxford University Press.

González de Prado Salas, J., and J. Zamora-Bonilla. (2015). Collective Actors without Collective Minds: An Inferentialist Approach. *Philosophy of the Social Sciences, 45*, 3-25.

Hakli, R. (2006). Group beliefs and the distinction between belief and acceptance. *Cognitive Systems Research,* 7(2-3), 286-297.

Hieronymi, P. (2008). Responsibility for believing. *Synthese,* 161(3), 357-373.

Kallestrup, J. (2016). Group virtue epistemology. *Synthese*, 1-19
(2016). https://doi.org/10.1007/s11229-016-1225-7

Kitcher, P. (1990). The division of cognitive labor. *The journal of philosophy,* 87(1), 5-22.

Lackey, J. (2016). What is justified group belief? *Philosophical Review,* 125(3), 341-396.

List, C. (2005). Group knowledge and group rationality: a judgment aggregation perspective. *Episteme,* 2(1), 25-38.

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

Mathiesen, K. (2006). The epistemic features of group belief. *Episteme,* 2(3), 161-175.

McHugh, C. (2012). Epistemic deontology and voluntariness. *Erkenntnis,* 77(1), 65-94.

McHugh, C. (2014a). Exercising doxastic freedom. *Philosophy and Phenomenological Research,* 88(1), 1-37.

McHugh, C. (2014b). Fitting Belief. *Proceedings of the Aristotelian Society,* 114, 167-187.

Meijers, A. (2002). Collective Agents and Cognitive Attitudes. *ProtoSociology,* 16, 70-85.

Meijers, A. (2003). Why Accept Collective Beliefs?: Reply to Gilbert. *Protosociology,* 18, 377-388.

Mercier, H. & Sperber, D. (2017). *The enigma of reason*. Cambridge, MA: Harvard University Press.

Owens, D. (2003). Does Belief Have an Aim? *Philosophical Studies,* 115, 283- 305.

Parfit, D. (2011). *On what matters*. Oxford: Oxford University Press.

Raz, J. (1975). *Practical reason and norms*. Oxford: Oxford University Press.

Scanlon, T. (1998). *What we owe to each other*. Cambridge, MA: Harvard University Press.

Schmitz, M. (2017). What is a mode account of collective intentionality? In G. Preyer & G. Peter (eds.*), Social Ontology and Collective Intentionality* (pp. 37-70). Cham, Switzerland: Springer International Publishing.

Sellars, W. (1956). Empiricism and the Philosophy of Mind. *Minnesota studies in the philosophy of science,* 1(19), 253-329.

Shah, N. & Velleman, J.D. (2005). Doxastic deliberation. *The Philosophical Review,* 114(4), 497-534.

Steglich-Petersen, A. (2006). No Norm Needed: On the Aim of Belief. *Philosophical Quarterly,* 56, 499-516.

Tollefsen, D.P. (2002). Challenging Epistemic Individualism. *Protosociology,* 16, 86-117.

Tollefsen, D.P. (2015). *Groups as agents*. New York: Wiley.

Tuomela, R. (2013). *Social ontology: Collective intentionality and group agents*. Oxford: Oxford University Press.

Van Fraassen, B. (1980). *The Scientific Image*. Oxford: Clarendon.

Wagenknecht, S. (2016). *Social Epistemology of Research Groups*. London: Palgrave Macmillan.

Wedgwood, R. (2002). The aim of belief. *Noûs,* 36, 267-297

Weisberg, M. & Muldoon, R. (2009). Epistemic landscapes and the division of cognitive labor. *Philosophy of science,* 76(2), 225-252.

Whiting, D. (2010). Should I believe the truth? *Dialectica,* 64(2), 213-224.

Williams, B. (1973). *Problems of the self.* Cambridge: Cambridge University Press.

Wray, K.B. (2001). Collective belief and acceptance. *Synthese,* 129(3), 319-333.

Wray, K.B. (2003). What really divides Gilbert and the rejectionists? *Protosociology,* 18, 363-376.

Zamora-Bonilla, J. (2014). The nature of co-authorship: a note on recognition sharing and scientific argumentation. *Synthese*, 191:97-108.