Analyse & Kritik 02/2012 (© Lucius & Lucius, Stuttgart) S. 383–398

Carlo Martini

Applying Formal Social Epistemology to the Real World

Comment on Paul D. Thorn and Gerhard Schurz

Abstract: The claim that diversity and independence have a net positive epistemic effect on the judgments of groups has been recently defended formally by Scott Page, among others, and popularized in Surowiecki's *The Wisdom of Crowds*. In *Meta-Induction and the Wisdom of Crowds* Thorn and Schurz take issue with the claim that more diversity and independence in groups leads to better collective judgments. I argue that Thorn and Schurz's arguments are helpful in clarifying a number of over-generalizations about diversity and independence that are often circulated in the social epistemology literature. I also argue that the relevant formal arguments are easily misunderstood when presented 'in a vacuum', that is, without a context of application in mind. I provide a different approach to understanding formal results in social epistemology: With the help of concrete scenarios and the formal literature, I focus on a trade-off between independence and dependence in groups. I show that the approach works well also for another principle in social epistemology; namely, the principle that 'more heads are better than few'.

1. The Mathematics of Social Epistemology

For several centuries scholars have been fiddling with the idea that knowledge, in principle, can better be pursued as a social activity, rather than as an individual one. In the 18^{th} century, the Marquis de Condorcet defended the institution of majoritarian democracy with arguments that were revolutionary for his time. The reasons for such novelty were at least two, the second of which will be the focus of this paper.

But let us start with the first innovation in Condorcet's defense of democracy. Before Condorcet, the main justification for democratic rule had been based on ethical grounds:[1] Democracy is *just*: it is respectful of the principle of human equality and mutual respect, it promotes peace, and so on. But it is not always easy to defend democracy by means of moral arguments: The problem of irreducible disagreement is common, in ethics, when moral arguments hit antithetic bottom-ground principles.[2] This is why it is remarkable that Condorcet could do

[1] See Hawthorne 2001 for a useful review of the different stances in support of democratic rule (that is, majority rule).

[2] See Martini 2012a and Martini et al. 2012b on the problem of *irreducible disagreement*.

away with ethics by arguing that democracy (to him, a specific form of democracy: majoritarian voting) is *epistemically better* at achieving society's material goals. If a country has a problem that can be, at least in principle, resolved by increasing its rulers' knowledge of the truth, then society, as a whole, is better suited at finding the truth than a subset of its members. This stance was in stark contrast with the idea of *enlightened despotism* because it stated that the crowd 'knows better' than a small elite of (supposedly) wise rulers.

But what was even more revolutionary in Condorcet's defense of democracy was that Condorcet could defend his thesis by mathematical proof, and this was the second innovation in his approach. His proof of what passed to history as the Condorcet Jury Theorem established the tradition of investigating social phenomena by means of mathematics and logics. Of course, like any theorem, the one proven by Condorcet came with caveats; one of them is that the theorem holds if the members composing the 'crowd' (the voters) cast their judgment independently. That means that the voters should not unconditionally 'parrot' the judgment of one of their fellow voters. Independence is one of the two main requirements discussed in Thorn and Schurz (2012). The second is diversity.

Formal investigations on diversity are a recent addition to the mathematical treatment of problems in social epistemology. The general idea is that a diverse crowd performs better—that is, finds out the truth more often—than a crowd whose members are like-minded. Most social epistemologists focus on "functional diversity: differences in how people encode problems and attempt to solve them" (Hong/Page 2004, 16385). Krogh and Vedelsby (1995) provide a mathematical interpretation of diversity:[3] for a member of a crowd (call her α), who gives her prediction or estimate on a certain event x, diversity is measured as the difference between α's individual estimate, and the average estimate of the group.

Much of the literature in social epistemology is interested in group accuracy (how different accuracies compare for different kinds of groups, for example), where 'accuracy' is defined as the distance between the estimate and the true value of a certain quantity. An important result in Krogh and Vedelsby's paper is the finding that the accuracy of a group is a function of both the average accuracy of the members of the group and the measure of diversity within the group[4]. What that means is that group accuracy can be increased by increasing diversity within the group, even when the average accuracy of individual members of the group does not change.

As I will explain in the next section, the important mathematical results in Condorcet (1785), Hong and Page (2004), and many others working in social epistemology, have some times been, perhaps even with good intentions, exaggerated to imply that more diversity and more independence are *always* (or even *in most cases*) better in society (see esp. Surowiecki 2004). That error of overgeneralization is the one Thorn and Schurz try to redress in their paper.

---

[3] Given an agent, α, and her forecast $V^{\alpha}(x)$ on input x, the measure of diversity (in the paper called 'ambiguity') on input x of α is defined as $a^{\alpha}(x) = (V^{\alpha}(x) - \overline{V}(x))^2$.

[4] For the formal model see Krogh and Vedelsby 1995, 3, and Thorn and Schurz 2012, 344, 345; a simple numerical illustration of this fact is given in the Appendix of this paper.

2. Diversity and Independence in Groups

Diversity of opinion and independence are two of the four "pillars of wisdom" in Surowiecki's popular *The Wisdom of Crowds*, the other two being *decentralization*, and *aggregation* (Surowiecki 2004, 10). In his Introduction Surowiecki writes that "paradoxically, the best way for a group to be smart is for each person to think and act as independently as possible". But that is not strictly true, as Thorn and Schurz (2012), among others, have shown.[5]

Unfortunately, Surowiecki's book gathers under the same umbrella different formal as well as non-formal results in the field of social epistemology, and overgeneralizes some of the aspects of group interaction. While that helps popularizing the point that "collectivities are often times smarter than the individuals they comprise", it also muds the water as to how, and under which conditions, groups have an advantage over individuals, or over other groups with different characteristics (e.g. more diverse, less centralized, etc.). It is in that light that Thorn and Schurz provide important clarifications on some of the claims that Surowiecki makes.

Let us consider diversity first. As explained in the previous section, the accuracy of a group is dependent on two factors: the average accuracy of the individuals that compose the group, as well as the diversity among opinions within that group. Hong and Page (2004), Surowiecki (2004), and Page (2007) claim therefore that increasing diversity is a way to increase the accuracy of a group. In a catchphrase 'diversity is better than conformity'. But Thorn and Schurz point out that more diversity can also, under certain conditions, decrease accuracy. A simple example will illustrate this point.[6]

Suppose that two agents, A and B, are asked to estimate the true value, $f(x)$, of a certain quantity, and let us assume that such value is equal to 0. We can compare two situations: in the first one, A's estimate is $V^a(x) = 5$, and B's estimate is $V^b(x) = 9$. The group's opinion, $V^G(x)$, is the arithmetic average of the individual estimates. The rate of the diversity of either A or B is the distance between, respectively, A's or B's opinion and the opinion of the group. Finally, the rate of diversity for the group is the arithmetic average of the individual measures of diversity. Let us now consider a second situation, in which B's opinion has changed: $V^b(x) = 11$. It is clear that in the second situation the rate of diversity in the group has increased. At the same time, however, B's opinion has moved further away from the truth, and, consequently, so has the opinion of the group. Therefore, the group's average error, measured as the difference between the truth and the opinion of the group, has also increased.

What the example above shows, and what Thorn and Schurz argue, is that increasing diversity is not always conducive to an increase in accuracy: "increasing diversity (independently of the resulting effect on $E(x)$) is not sufficient for decreasing $E^G(x)$" (Thorn/Schurz 2012, 347). Most importantly, increasing diversity is not the only, nor is it the best, way of boosting accuracy. The bet-

[5] Estlund 1994 had already shown that a certain amount of deference to the opinion of others can increase a group's accuracy.

[6] Complete calculations and details of this numerical example are in the Appendix.

ter alternative, in the eyes of the authors, is to try decreasing individual error, and that is because "decreasing $E(x)$ to zero is sufficient (independently of the resulting effect on diversity, $D(x)$) for decreasing $E^G(x)$ to zero". In short, making people smarter, individually, is *sufficient*, but not *necessary*, for making the group smarter,[7] while increasing diversity is neither *necessary* nor *sufficient*.

Obviously, the strategy of decreasing individual error is more difficult to put into practice than increasing diversity, but Thorn and Schurz claim that a way to make people smarter, individually, is to make some of them follow the smartest agents in the group; that is, to violate the requirement of independence. In other words, while Surowiecki (2004), and some interpretations of the Condorcet Jury Theorem, claim that that the best way to make a group smarter is by adding independent individuals, the simulations in Thorn and Schurz (2012) show that adding individuals that "imitate" the judgment of some of their fellow group members is, under certain circumstances, the best way to make the group smarter: "adding meta-inductive strategies can only improve and will never diminish the maximal success rate." (Thorn/Schurz 2012, 346)[8]

Thorn and Schurz define 'meta-induction' as "an imitative prediction method, where the prediction methods and the predictions of other agents are imitated to the extent that those methods or agents have proven successful in the past" (Thorn/Schurz 2012, 363). On the surface, their paper appears to be in open contrast with the literature on diversity and independence, but I will show that the view is mistaken. The simulations contained in their paper only prove that, in special cases imitation (meta-induction) is better for group performance, but those cases are compatible with the results in Condorcet (1785), Page (2007), and the other literature that Thorn and Schurz consider. The interesting question then, is: what kind of general conclusions, as far as independence and diversity are concerned, can we gather from the literature?

In the next sections I will argue that, together, the studies of Hong and Page (2004), Page (2007), and of Schurz (2008), Thorn and Schurz (2012) warrant some general conclusions for social epistemology, but such conclusions need to be based on practical considerations on groups, their features, and their membership composition. The approach to social epistemology defended in the next section is grounded on concrete group scenarios, which is what is needed to generalize the formal results and make them applicable in the real world.

3. Thinking Independently and Imitating: A Problem of Calibration

What one can gather from the numerous debates among formal social epistemologists is that it is extremely difficult to generalize the results without incurring in oversimplifications. Part of the issue is that many of the results from formal

[7] It is not necessary, because some times diversity alone *can* make the group smarter.

[8] Estlund 1994 had reached similar conclusions, although in this paper I will focus only on the simulations in Thorn and Schurz 2012. "Under certain conditions deference to opinion leaders can improve individual competence without violating independence, and so can raise group competence as well." (Estlund 1994, 132)

research are often presented 'in a vacuum'; that is, without enough thought as to what the specific individual conditions of a deliberating group may be. Moreover, all formal results, whether from simulations or analytical proofs, come with an array of premises and conditions; hence the drawback that it is sometimes hard to make out at first glance how general a certain result is, and whether it is really, or just nominally, in contrast with other formal results. Nevertheless, the advantage of formalism is that we can easily—for the most part at least—identify the assumptions, and compare proofs and simulations not only by their conclusions but also by their premises.

In this section, I will illustrate some of the results from the simulations in Thorn and Schurz (2012) and explain why their conclusions only party contradict the conclusions of Hong and Page (2004) and Page (2007). I will further illustrate how we can make sense of both in concrete applications of social epistemology. I will start with a brief overview of the core of Thorn and Schurz's results from simulations, and explain why they are only apparently at odds with the defenders of diversity and independence.

In sections 5 and 6 of their paper, Thorn and Schurz set up the model that they then use to run simulations and investigate "what happens when to the wisdom of the crowd when meta-inductivists or other social learners replace independent forecasters" (Thorn/Schurz 2012, 349). There are a number of possible strategies that are alternative to independence.[9] This is a list of the main ones: a) *imitate the best*—the imitators 'parrot' the judgment of the most accurate member of the group; b) *weighted meta induction*—the imitators take a weighted average of the most accurate members of the group; c) *peer-imitation*—the imitators take a straight average of the non-imitating players.

The core assumptions of the simulations are that the imitators must have access to the judgment of those they imitate, and that the imitators must be allowed to cast their judgment after the non-imitators have cast theirs, thereby allowing the former to know whom the best independent agents are. A number of tables illustrate the results.

The first simulation shows that a 'Condorcet-group', composed wholly of independent agents, will tend to full accuracy, when the accuracy of its members is above .5, that is, when its members are better than randomizers. That result is implied by the Condorcet Jury Theorem. Then, the authors compare a Condorcet-group with other groups: a group composed fully of agents who adopt the imitate-the-best strategy does not show the wise-crowd effect that the Condorcet-group shows; and two small variations (in tables 3 and 4) on the group's composition do not change the previous results very much. Also a group composed wholly of agents using a weighted meta-induction shows no wise crowd effect. Drawing some partial conclusions, it looks as if homogeneous groups of imitators are not wise. Moreover, though the authors do not seem to take notice,

[9] Independence between two agents in the model, X and Y, who give judgments A and B, respectively, implies that the probability that X will judge A, conditional on the fact that Y has judged B, is equal to the probability that A will judge X regardless of what B does. More formally, for agents X and Y, and their respective judgments A and B, A is independent from B *iff* $P(A|B) = P(A)$.

there is a logical problem in dealing with a group of pure meta-inductivist: if everyone's judgment is based on other people's judgments, then no one could ever hold *any* judgment to begin with.

It is easy to see that the simulations conceal the logical problem I just mentioned. Presumably, though it is not explicit in the paper, agents in the formal model start from a random distribution of opinions: Before the first round of updates, some agents will so happen to be closer to the truth than others, thus setting the course of successive updates. However, this is not what happens in practice: it seems very unrealistic to conceive of any group in which all the members start out as imitators right from the beginning, with a random set of initial beliefs. It is more reasonable to think of concrete groups as mixes where some of the agents have a prior opinion on the matter under consideration, and some have no opinion, and are willing to follow whichever of their fellows they think is (or are) the most accurate. Indeed, in their next round of simulations, Thorn and Schurz study a mixed group composed of both independently opinionated agents as well as imitators.

After the first round of simulations, the authors provide what they say is a more realistic set of assumptions for their model, one in which there is a minority (10% of the total group population) of highly reliable independent predictors, and a majority of imitators (the remaining 90%). The results, as before, are illustrated in several tables. First, to set a benchmark, table 7 presents the case of a group made up for 90% of independent members, with reliability slightly over .5, and for 10% of independent agents with high reliability (.9). The simulation shows that the average individual error approximates the independent *unreliability*[10]—this is the independent variable in the model—of the majority of the group's agents. The group is still smart: the 'average global group error' is 0, in the long run, when its members are better than randomizers.

What the authors want to show is a way to reduce average individual error. As explained in the previous sections, the average group error is a function of two variables: average individual error and diversity. Being able to lower individual error makes the group smarter, and, in the limit, reducing average individual error to 0 makes average group error also disappear. In the following tables, Thorn and Schurz compare the previous non-mixed group of independent agents with a mixed group (90% unreliable imitators, and 10% very-reliable independent agents). When the imitate-the-best strategy is used, the results show that the individual error rate approximates the independent unreliability of the independent agents: "The prediction strategy of the bMIs [imitate-the-best agents] yields the result that the individual error rates of the bMIs approximate the error rate, $u = 0.1$, of the highly reliable subgroup." So individual error rate is much lower in the mixed group, than in the Condorcet-group.

Moreover, the mixed group is 'smart' in the sense that the average global group error is quite low, and approximates the error rate of the smart individuals, set by the modelers to 0.1. What is interesting is that when error rates are high for the majority of agents in the group (their reliability is below .5), the group fares better than the Condorcet-group does. As one would expect, the

[10] Unreliability = 1− reliability

Condorcet-group's error rate is high and tends to 1. But when the average individual error is above .5, in the group with imitators the average individual error rate is much lower, since the minority of smart agents are being imitated by the majority of "dumb" ones. So even when the majority of individuals are worse than randomizers, the group with imitators is much smarter than the one without, regardless of whether they are using the imitate-the-best strategy (table 8), or weighted meta induction (table 9).

This point is important for the discussion to follow. "In summary, we see that applying imitate-the-best or weighted meta-induction results in higher individual success rates (as compared to making independent predictions), so long as we assume that the meta-inductivist has the opportunity to imitate players whose independent reliability exceeds her own [that is, when there is a core of highly-reliable independent predictors]." (Thorn/Schurz 2012, 356)

What Thorn and Schurz show is that, as long as there is a core of highly-reliable independent predictors, and as long as the imitators can see their predictions, the group is about as good as a Condorcet-group, and, in fact, better than a Condorcet-group when the average individual reliability of the majority is low. But how do these results compare with the focus on independence and diversity? The claim I defend here is that the results in Thorn and Schurz (2012) do not run counter those of Page (2007) and Surowiecki (2004), but are rather complementary findings that allow us a better understanding of the epistemology of social groups.

Keeping in mind the Condorcet Jury Theorem (Condorcet 1785; Estlund 1994), the findings from Krogh and Vedelsby (1995) and Page (2007), and the simulations in Thorn and Schurz (2012), let us now pinpoint some general results, and a trade-off, in the social epistemology of group wisdom. The first observation is that if the average reliability of the majority is above .5, having only independent agents or a majority of imitators in addition to a minority of highly reliable agents does not make a big difference: the accuracy of the former, the Condorcet-group, converges to 1, while the latter mixed group converges to the average accuracy of the best (independent) members. So a Condorcet-group, if only by little, is better in the long run than a mixed group when average individual accuracy is higher than .5. In this respect, then, independence wins.

But we know that the condition of having members with reliability higher than .5 is not always satisfied, and this is the second generalization that we can make, based on the results in Thorn and Schurz (2012): whenever the majority of agents in a group are worse than randomizers, the average group error in a mixed group will be much lower than in the Condorcet-group, and the group itself will be smarter than a Condorcet-group. So it is now possible to see the utility of the foregoing considerations in the light of the composition of a group. A factor that is often neglected in social epistemology literature, and that is mostly ignored in Surowiecki (2004), is that talking about 'groups' in social epistemology is misleading, because the features of specific real-world groups are extremely important for establishing which formal results are applicable and which ones are not. The trade-off that we can establish from the formal results reviewed so far is that independence yields better or worse results depending on

the average reliability of a group's members. To see an application of that, in the next section I will provide some concrete examples.

4. Applying Formal Results in Real-World Cases

Let us now go back to the results in Thorn and Schurz (2012) and think of some practical situations where they could be applied. Let us imagine a context where a subgroup of the total group population is made up by highly reliable experts, while the majority of the group (the laypeople) have a high error rate. We can think of non-experts as individuals who completely lack any knowledge in a specialized subject area, but also as people whose expertise is completely irrelevant to a certain task. For instance, the otherwise expert engineer will be a layperson if she is put in the role of a physician diagnosing a patient. The typical physician, on the other hand, will be a layperson when trying to calculate the eigenvalues for the project of a bridge. Both tasks, calculating eigenvalues, and diagnosing an illness, require a considerable amount of knowledge and skills that are typically possessed only by a small minority of the total population. In those cases a mixed group, where the majority parrots the expert minority is probably better than a Condorcet-group, where everyone tries to think independently about the solution to the given problem.

But there are a number of important instances in which we simply do not know whom the experts are. In the case of physicians and engineers there is a relatively straightforward set of criteria that determine who has and who does not have the skills to solve a medical or engineering problem. However, as the complexity of the problem increases, it becomes harder to identify the relevant variables, and the experts are no better, if not worse, than the laymen (see Staël von Holstein 1972; Yates/McDaniel 1991).

Let us take as an example electoral predictions. In November of 2012, the American media were predicting a neck-and-neck race between the two principal US presidential candidates, Obama and Romney. The (alleged) election experts were mostly divided between those who claimed that Romney had very good chances of winning the presidency, and those who gave Obama another landslide victory. One can certainly claim that each expert or media outlet always gives better odds to their personal (read 'favorite') pick, because that will most likely increase those very same odds, and that has nothing to do with analytical objectivity. But that does not change the fact that up until the vote-count on the night of November 6, the presidential race was widely perceived by even the most informed people as a neck-and-neck race.

Those whose excitement for the forthcoming outcome of the elections had been pacified well in advance were the ones who had kept an eye on prediction markets. Since the morning of November 6, 2012 prediction markets were already well-settled on the fact that Obama, given at 1 to 6 against Romeny, would comfortably win the 2012 elections (Snyder 2012; Vaughan Williams 2012). In that case, the markets, composed of mostly independent (and 'inexpert') bettors, were better at predicting the outcome of the presidential race than most political

and election experts. When it comes to a complex problem where the variables relevant for expertise are mostly unknown, it then looks like groups are better at finding the truth when their members think independently from one another rather than relying on only-supposed experts.

A further point of this section is that imitation should be used only when we already have a clear picture of the situation at hand. That is, when we can identify the group as a few-experts/many-laypeople one, and when we can tell the ones from the others; that is, when we can identify the experts in the group. As we saw, however, there are important situations where we cannot do that; in those cases we are better-off stressing the importance of diversity and independence. The foregoing considerations are hardly surprising, and have an analog in the formal literature on judgment aggregation and the use of weights.

Let's make the (only artificial) analogy between equal weights (in aggregation problems) and independence (in group decisions). Armstrong (2001b) suggests that equal weights should be the 'default' in cases of uncertainty, when we simply do now know in advance who has the lowest and highest error rates—note that in typical simulations the modeler normally knows in advance the error rate of the agents in a group, or in the group's subgroups. In the section of his paper titled "implications for practitioners", Armstrong writes that "equal weighting provides a good starting point, but use differential weights if prior research findings provide guidance or if various methods have reliable track records in your situation" (Armstrong 2001a, 422, 423). When we know that the situation at hand is one of 'few experts/many laypeople', and when we know whom belongs to which subgroup, there are better techniques of aggregating information than simple equal-weighted average, and we should make use of unequal weights. For all other cases we should assume that everyone has the same chances of hitting the truth, therefore assigning equal weights.

Similarly, the results surveyed in this paper indicate that whenever we do not have information about the composition of a group we might be better-off by relying on independence of the members and, with some caveats that were discussed in the paper, on diversity. However, there are asymmetric situations in which we should be aware of the fact that groups of wholly independent agents are not going to give the best collective results. In those cases meta induction and imitation are the best epistemic strategies.

In this section I have shown that independence or its opposite, dependence, are not universal conditions for the wisdom of groups. There is a trade-off between the two conditions, and which degree of independence or dependence will make a group 'wise' will have to be established on the basis of the characteristics of the specific group in question. The same is true for the trade-off between diversity and homogeneity. The approach used to establish the trade off was an empirical one: we need to look at the composition of a group to establish which particular epistemic benefit the group would receive from adopting different epistemic strategies (e.g. imitation, independent thinking, etc.).

In the next section, I will show that another overgeneralization about groups is the idea that 'many are better than few'. The principle, established by the Condorcet Jury Theorem, is often taken to be one of the fundamental truths of

social epistemology: the more minds put themselves to a task, the more likely they are to find out the truth. Like the principles of independence and diversity, also the 'principle of large numbers' is true only for certain kinds of groups, even when the *formal* conditions for its application are satisfied.

5. When Few Are Better Than Many

One of the better-established principles in social epistemology is the fact that 'many are better than few'. In his *Enquiry* Hume states that when the supporters of a thesis are "but a few" we might legitimately doubt the "report of others" (Hume 1784, 99/113). In other words, Hume claims that agreement from many sources warrants more epistemic weight than agreement from only a few sources. A similar principle, with some qualifications, can be found in Goldman: "It appears, then, that greater numbers should add further credibility [to a given prediction or estimate], at least when each added opinion-holder has positive initial credibility." (Goldman 2001, 99) Let us call the idea expressed in Hume (1784) and Goldman (2001) the 'principle of large numbers'. In this section I will illustrate the fact that under some conditions a counter-principle applies: few are better than many.

In its formal version, established by Condorcet (1785), the principle of large numbers states that, as the number of reliable[11] decision makers becomes larger and larger, the group will hit the truth with probability approximating full certainty. The reliability condition of the theorem already imposes some restrictions on the domain of applicability of the principle of large numbers: to give an example, a medical diagnostic division would not gain any advantage from the whole hospital population participating in the diagnosis of a patient. But there are more (and non-formal) constraints on the applicability of the principle. In this section I will discuss the size of groups of *experts*, to wit, individuals with highly specialized skills and knowledge relevant to a specific subject matter.

We can agree on the fact that a group of experts, as defined above, has a reliability higher than .5, and, for the sake of the argument, that experts enjoy a degree of independence sufficient to satisfy the conditions of the Condorcet Jury Theorem. In this section I will introduce a distinction between voting and deliberation, and explain why the principle of large numbers applies well to the former, but faces serious limitations with the latter.

Let us take an example from economics: monetary policy committees. There are several committees around the world, which take decisions such as whether to ease interest rates, print money, etc. In fact, the employment of committees, to replace what was previously a one-man rule in central banks, has been considered a successful application of some of the principles of social epistemology to monetary policy regulations (Blinder 2004). Monetary policy committees are groups of economic experts who gather periodically to discuss the monetary policy of a country, provide economic analysis, and give the monthly mandate to

[11] Reliable decision makers are defined as those who have better-than-random chances of finding the truth about a given matter.

the central bank for its monetary operations. Different committees adopt different decision making strategies: to mention only a few, the US Federal Open Market Committee (henceforth FOMC) works by consensus decision making—consensus is produced in the deliberative phase of each meeting, and voting is mostly a formality. By contrast, the Bank of England's Monetary Policy Committee (henceforth MPC) takes voting seriously: at the end of a deliberation phase, votes are taken on the official committee's statement, and dissensus is not uncommon and is registered in the minutes.

One can say that while the FOMC works by deliberation, the MPC works by voting. This does not mean that the experts in the MPC do not deliberate, but only that they use a formal aggregation mechanism (voting) to form a group decision. On the other hand, the FOMC is purely deliberative, members are expected to reach a consensus, and the group's decision is not produced by a formal mechanism, but rather by informal deliberation. The question one may ask is whether adding more members to the 9-member MPC, or the 12-member FOMC, would be a good strategy for improving their policies. After all, the formal setting would guarantee that adding competent and independent economists to the committees can only improve their performance.

To be sure, the problem I am concerned here with is independent of any political role of the monetary policy committees. One can legitimately doubt whether those committees have purely predictive and analytical roles, and not broader political mandates. Regardless, under normal circumstances the committees are assigned a specific goal—e.g. the Bank of England has a 2% upper-bound inflation target—and they are supposed to achieve that goal by using the monetary instruments given by the law: that involves mostly value-neutral tasks like prediction and calculation. It is only with respect to those kinds of tasks that we can claim the 'principle of large numbers' to have epistemic relevance.

Unfortunately, increasing the number of members in a committee, even with highly reliable and independent members, is not always the best strategy. That is because, when the committees are not only aggregation mechanism but rather deliberative ones, there are a number of contextual factors which need to be taken into account, and which make the optimal size of a committee much smaller than the theoretical results would imply, even ignoring the costs associated with maintaining a committee. The interaction between committee members that takes place in the FOMC in order to produce consensus is not accounted for by the highly idealized conditions that are necessary to generate formal results in social epistemology. In the following I highlight only two of the factors that limit the applicability of the large numbers principle: coordination costs and free-riding.

The effect of free-riding in large aggregates has been observed in several psychological studies (Sibert 2006): in groups, people tend to perform worse, individually, than they would otherwise do when in isolation, because they can free-ride on other members' efforts. In other words, if I find myself in a situation in which I can 'slack off' someone else's work, I will probably put less effort into the task I'm performing. Economists and psychologists have studied the phenomenon in game theory (Varian 2004) and with experiments (Albanese/van

Fleet 1985), and have identified two distinct theses. The stronger thesis defines free-riding as a zero-contribution, occurring when the members of a group have a chance to free-ride on others. There is little evidence that such strong form of free-riding occurs (Marwell/Ames 1981; Carpenter 2007). The weaker thesis, on the other hand, identifies free-riding as a decrease, rather than complete disappearance, of individual contributions, and has much stronger support in the literature (Albanese/van Fleet 1985; Marwell/Ames 1981; Carpenter 2007).

In particular, the studies show how free-riding is affected by the size of the group in at least two ways. One the one hand, in large groups people feel less pressure towards optimal performance, because it is harder to individuate personal responsibility of group's outcome: if the outcome is positive, merits will be spread thinly, and a negative outcome cannot be easily attributed to one member or the other. On the other hand, in large groups it is harder for 'monitors' to punish free-riders (Carpenter 2007), thereby disqualifying a possible deterrent of free-riding.

The literature on free-riding is vast, and extremely relevant to the epistemology of social groups, but let us return to the example of monetary policy committees. The composition of typical monetary policy committees around the world ranges from five to twenty members, but Sibert (2006) concludes that committees should not count more than seven or nine members. Voting committees might benefit from larger groups. In the MPC each member's vote is public and reported in the minutes,[12] therefore making each individual accountable for his or her decisions. Even in the MPC, however, the benefits of a large committee may be curtailed by free-riding during the pre-voting deliberative phase of the meetings. It is clear that in committees like the FOMC, large groups will present important opportunities for their members to free-ride, so adding members over a certain limit will impact the group's performance (Sibert 2006).

Secondly, let us take a look at coordination costs. In the FOMC, which works by consensus, the deliberative phase is crucial for producing a final statement that will be agreed on by all the members. It is the explicit task of the Committee's Chair to build consensus both in the FOMC and, to add another example to our list, the European Central Bank (ECB) committee on monetary policy: "A *collegial committee* prizes solidarity and strives for group ownership of any decision that it makes. The chairman therefore tries to forge a strong consensus; the goal is unanimity, if humanly possible." (Blinder 2004) But building consensus is clearly a harder task the larger the committee is, at least where there is no alternative aggregation mechanism available.

Voting is such a mechanism, but voting is only a formality in the FOMC and in the ECB, so adding (independent) members to either of those committees would make the task of deliberating and achieving consensus much harder. The MPC, on the other hand, enjoys the advantage of having a simple aggregation mechanism. Therefore, adding members to it would not create significant coordination problems, because any disagreement can be easily removed via voting at the end of deliberation. For the reasons illustrated in the previous paragraphs

[12] Minutes of the MPC meetings are made publicly available on the Bank of England's official website some weeks after each meeting: http://www.bankofengland.co.uk.

Sibert (2006) concludes that the size of committees where both psychological factors and group dynamics are present is in the end an empirical issue.

To summarize, the principle of large numbers states that large groups are smarter than small ones. In particular, if we add competent members to a group, its performance will increase. This is an important result from the research on social epistemology. However, we have seen in this section that both psychological and economic research shows that adding members to a group can in fact make the group 'dumber', both because the members of a larger group are likely to free-ride on the effort of others, and because of coordination problems. Having that been established, I also argued that the two effects (free-riding and coordination) affect a committee differently depending on whether it works by consensual decision making (FOMC) or by voting (MPC). As it was the case for independence and dependence, as well as diversity and homogeneity, also in this section one can see that there is a trade-off: the epistemic benefits of adding members to a group need to counter balance the possible epistemic losses deriving from free-riding and coordination costs.

6. Conclusion

In this paper I have focused on what seem to be two established formal results in the epistemology of groups, which are based on the simulations of Thorn and Schurz (2012), and the literature the authors address, especially Condorcet (1785), Surowiecki (2004) and Page (2007).

In *section 3* the first conclusion was that, if the average reliability of the majority is above .5, then having only independent agents or a mixed group (with a majority of imitators in addition to a minority of highly reliable agents) makes little difference. In the long run, the accuracy of the former, which I called the Condorcet-group, converges to 1, while the mixed group converges to the average accuracy of the best (independent) members, which is set high by default. A Condorcet-group then, if only by little, is better in the long run than a mixed group, and in this respect, independence comes out the winner. One should note that the imposing conditions on imitative groups—e.g. they must be able to detect correctly the best performers in the group—may also, in this case, weigh in favor of the simpler independent strategy, but these considerations are not properly epistemic.

The second conclusion holds when the most members in the group are worse than randomizers. In that case it is well-known that a Condorcet-group gives disappointing results: its accuracy decreases to 0 in the long run. A mixed group of imitators and experts, however, performs much better, and the simulations of Thorn and Schurz (2012) showed that the average group error approximates the error of the best members (the experts), thus affording the mixed group much better results than the Condorcet-group.

In *section 4*, I showed how those results can be applied to concrete situations, marking the fact that the choice between independence and imitation is not an all-or-nothing one, but rather a matter of calibration and trade-off, which should

be based on real-world situations. I have also shown, in *section 5*, that a similar trade-off exists with other principles of social epistemology, for example with the principle of large numbers. Ultimately, I have argued that proper conclusions in social epistemology need to be based on concrete cases, and that formal results are mostly useful when applied to those cases.

Bibliography

Albanese, R./D. D. Van Fleet (1985), Rational Behavior in Groups: The Freeriding Tendency, in: *Academy of Management Review*, 244–255

Armstrong, J. S. (2001a), Combining Forecasts, in: Armstrong, J. S. (ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*, Norwell, 417–439

— (2001b) (ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*, Norwell

Blinder, A. S. (2004), *The Quiet Revolution: Central Banking Goes Modern*, New Haven

Carpenter, J. P. (2007), Punishing Free-riders: How Group Size Affects Mutual Monitoring and the Provision of Public Goods, in: *Games and Economic Behavior* 60(1), 31–51

Condorcet, Marquis de (1785), *Essai sur l'application de l'analyse á la probabilité des décisions rendues á la pluralité des voix*, Paris

Estlund, D. (1994), Opinion Leaders, Independence, and Condorcet Jury Theorem, in: *Theory and Decision* 36, 131–162

Goldman, A. (2001), Experts: Which Ones Should We Trust, in: *Philosophy and Phenomenological Research* 63(1), 85–110

Hawthorne, J. (2001), Voting in Search of the Public Good: The Probabilistic Logic of Majority Judgements, unpublished manuscript, University of Oklahoma, URL: http://faculty-staff.ou.edu/H/James.A.Hawthorne-1/Hawthorne-Jury-Theorems.pdf [October 2012]

Hong, L./S. E. Page (2004), Groups of Diverse Problem Solvers Can Outperform Groups of High-ability Problem Solvers, in: *Proceedings of the National Academy of Science* 101(46), 16385–16389

Hume, D. (2007[1748]), *An Enquiry Concerning Human Understanding and Other Writings*, Cambridge

Krogh, A./J. Vedelsby (1995), Neural Network Ensembles, Cross Validation, and Active Learning, in: Tesauro, G./D. Touretzky/T. Leen (eds.), *Advances in Neural Information Processing 7*, Cambridge, 231–238

Lambert, R. (2005), Inside the MPC, in: *Bank of England's Quarterly Bulletin Spring* 2005, URL: http://www.bankofengland.co.uk/publications/Pages/quarterlybulletin/monpol.aspx [December 2011]

Martini, C. (2012), A Puzzle about Belief Updating, in: *Synthese*, DOI: 10.1007/s11229-012-0132-9

— /J. Sprenger/M. Colyvan (2012), Resolving Disagreement through Mutual Respect, in: *Synthese*, DOI: 10.1007/s10670-012-9381-8

Marwell, G./R. E. Ames (1981), Economists Free Ride, Does Anyone Else?, in: Experiments on The Provision of Public Goods IV, *Journal of Public Economics* 15(3), 295–310

Page, S. E. (2007), *The Difference–How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*, Princeton

Sibert, A. (2006), Central Banking by Committee, Working Paper No. 091/2006, De Nederlandsche Bank NV

Schurz, G. (2008), The Meta-Inductivist's Winning Strategy in the Prediction Game: A New Approach to Hume's Problem, in: *Philosophy of Science* 75, 278–305

— (2009), Meta-Induction and Social Epistemology: Computer Simulations of Prediction Games, in: *Episteme* 6, 201–220

Snyder, G. (2012), Mitt Romney's Very Long Odds of Winning the Election, in: *The Atlantic Wire*, October 26, URL : http://www.theatlanticwire.com/politics/2012/10/mitt-romneys-verylong-odds-winning-election/58414/ [November 1, 2012]

Surowiecki, J. (2004), *The Wisdom of Crowds*, New York

Staël von Holstein, C.-A. S. (1972), Probabilistic Forecasting: An Experiment Related to the Stock Market, in: *Organizational Behavior and Human Performance* 8, 139–158

Thorn, P./G. Schurz (2012), Meta-Induction and the Wisdom of Crowds, in: *Analyse & Kritik* 34, 339–365

Vandenbussche, J. (2006), Elements of Optimal Monetary Policy Committee Design, IMF Working Papers WP/06/277, URL: http://www.imf.org/external/pubs/ft/wp/2006/wp06277.pdf [June 2011]

Varian, H. (2004), System Reliability and Free Riding, in: Camp, J. L./S. Lewis (eds.), *Economics of Information Security*, Dordrecht, 1–15

Vaughan Williams, L. (2012), Prediction Markets: The Other Big Winners on Election Night, in: *Huffington Post*, November 8, URL: http://www.huffingtonpost.com/leighton-vaughan-williams/predictionmarkets-election-b-2091920.html [November 8, 2012]

Yates, F. J./L. McDaniel (1991), Probability Forecasting of Stock Prices and Earnings: The Hazards of Nascent Expertise, in: *Organizational Behavior and Human Decision Processes* 49, 60–79

A. Diversity: An Example

In this Appendix I provide a numerical example that illustrates two aspects of diversity. In the first place, the fact, shown in Krogh and Vedelsby (1995), that diversity plays an essential role in the accuracy of a group. And in the second place the fact, stated in Thorn and Schurz (2012), that diversity does not always reduce the group's error rate.

In the following example, the true value, $f(x)$, of a quantity is 0, and in the two cases both members of the group, A and B, are trying to estimate $f(x)$.

Case 1: A group with two members. Member A's estimate is $V^a(x) = 5$, and agent B's is $V^a(x) = 9$. The "group's estimate" is defined as the arithmetic average, $V^G(x) = 7$.

Let us know calculate *diversity* for agent A: $D^a(x) = (V^a(x) - V^G(x))^2 = 4$; and then for agent B: $D^b(x) = (V^b(x) - V^G(x))^2 = 4$. *Diversity* for the group, $D(x)$ is simply the arithmetic average of individual diversities and is equal to 4.

Next, we can calculate error rates. A's error is $E^a(x) = (f(x) - V^a(x))^2 = 25$. B's error is $E^b(x) = (f(x) - V^b(x))^2 = 81$. One may expect the group's error to be the average of individual errors, but that is not the case. The group's error is $E^G(x) = (f(x) - V^G(x))^2 = 49$, but the arithmetic average of A's and B's errors is $E(x) = 53$.

The difference between $E(x)$ and $E^G(x)$ is 4, which is equal to the rate of diversity in the group: $D(x) = 4$. The main theorem in Krogh and Vedelsby (1995), explains that fact: $E^G(x) = E(x) - D(X)$.

To see that an increase in diversity does not always amount to an increase in group's accuracy, let us compare Case 1, above, with a slightly modified case, in the following.

Case 2: A group with two members. Member A's estimate is $V^a(x) = 5$, and agent B's is $V^a(x) = 11$. As before, the "group's estimate" is defined as the arithmetic average, $V^G(x) = 8$.

Let us now calculate, in this order, diversity for A, B, for the group, and errors for A, B, their average error, and the group's error. $D^a(x) = 9$; $D^b(x) = 9$, $D(x) = 9$. $E^a(x) = 25$, $E^b(x) = 121$, the arithmetic average of A's and B's errors is $E(x) = 73$, and $E^G(x) = 64$. As before $E^G(x) = E(x) - D(X)$ $(64 = 73 - 9)$.

One can immediately see that the diversity of the group in Case 2 has increased from 4 to 9, but also that the increase has not made the group wiser: In Case 2, the group's error is higher than in Case 1. This shows that an increase in diversity is not a sufficient condition for making a group's wiser, as explained in *section 2* of this paper and in Thorn and Schurz (2012).