deliberative lies

robert e. goodin

Philosophy Program, Research School of Social Sciences, Australian National University, Canberra, ACT 0200, Australia.

E-mail: goodinb@coombs.anu.edu.au

doi:10.1057/palgrave.eps.2210188

Abstract

Deliberative democrats enjoin participants in ideal speech situations never to lie. But game theorists show that people can have purely truth-based motives for strategically misrepresenting information they hold privately when they are deliberating with others. If deliberative democrats want to ensure that every participant in the deliberation fully and truthfully reveals to one another all private information that they hold, that sets some stringent requirements for the nature of the group that deliberates together.

Keywords deliberative democracy; strategic misrepresentation

I am less exercised than Jürg Steiner (2008) about 'concept stretching' as such.¹ My sympathies lie more with Humpty Dumpty. Austen-Smith and Feddersen (2006) are welcome to use the word 'deliberation' one way, and Steiner another, just as long as they keep the subscripts straight: just as long as no one confuses 'deliberation_S' for 'deliberation_{ASF}'. I agree with Steiner that subscripts are messy and, all things considered, it would probably be better if Austen-Smith and Feddersen found a different term. That is not a matter of any great moment, however, to my way of thinking.

I

What matters much more is getting clear on exactly what the game-theoretic representation really tells us about

deliberation. It does not say that 'lying must always occur'. It merely says that 'lying may always be tempting'. That does nothing to contradict deliberationists' injunction 'do not lie!' It merely explains why deliberationists need to vest the injunction with normative force. One might always have reasons of one's own to be tempted to deviate from that normative precept.

By analogy: The moralists' injunction 'do not steal!' only makes pragmatic sense in the context of there being something that one might be tempted to steal. To say 'stealing may always be tempting' is not to deny the normative force of the injunction 'do not steal!' Instead it is to explain why we need to vest that injunction with normative force.

How *much* normative force ought we invest in the injunction? How absolutely

ought we insist upon it? Game theorists show us that a person's lying (strategic misrepresentation of that person's true preferences or true beliefs) can produce better outcomes, not only for that person but also for the group as a whole. So lying might be good, at least occasionally. So too might stealing, at least occasionally.

How occasionally? Proofs concerning what is and is not possible tell us nothing about frequencies. If the frequency of occasions on which lying or stealing is the right thing to do is vanishingly small – if it is possible, but barely possible, that that is the right thing to do – then it is approximately right to treat the injunctions 'do not lie!' and 'do not steal!' as absolute prohibitions. The approximation is only approximate, of course. But it might be as nearly correct as any other by which boundedly rational agents live real lives in the real world.

II

Most game-theoretic work on lying in politics involves the misrepresentation of one's preferences for strategic purposes in order to induce others to do something one prefers that they do. The typical example is insincere voting. You deliberately vote for what you see as the worse of options *A* and *B* in the first round, in order to set up your most preferred option *C* against a weak contender in the second round (Farquharson, 1969).²

In the context of deliberation, however, another sort of strategic behaviour might occur. You might have truth-based as well as preference-based reasons for lying, understood there as strategically misrepresenting your beliefs in order to make others believe what you believe to be true. In the standard game-theoretic set-up (Austen-Smith and Banks, 1996; Feddersen and Pesendorfer, 1998) that typically amounts to acting in a way that misrepresents your private information.

Here is how that might happen. Imagine 100 jurors, who agree that the accused should be convicted if it is 90 per cent likely she committed the crime. Suppose each juror has some private information on that matter, as well as the public information available to all of them; and for the sake of this example suppose jurors are identical to one another in every respect except for this private information. Being savvy game theorists, the jurors immediately perceive that they really only need to think about two possible scenarios. In one scenario their vote is not pivotal (or, if you prefer the language of deliberation to that of voting, the reason they give will not be decisive). In that case it does not matter which way they vote (or speak). The only time it makes a difference how they vote is when their vote is decisive. Now reflect upon the situation that must obtain in order for that to be the case: if the voting rule is that it takes 90 votes to convict, for juror J100 to be decisive it must be the case that 89 other jurors think that the accused is guilty. Juror J100 at that point undertakes some hypothetical Bayesian updating, calculating what he would think the probability of guilt would be conditional on 89 others saying 'guilty'. Depending upon his prior and conditional probabilities, it might well turn out that after that updating he would think it is over 90 per cent likely that the accused is guilty, if 89 others say she is – even though his own private information suggests she is not. So juror J100 strategically votes 'guilty', contrary to his private information suggesting that she is innocent; and the accused is convicted.³

III

Notice why J100 is tempted to vote strategically, rather than act in conformity with his private information. The problem arises because the threshold for collective

acceptance of the proposition 'the accused is guilty' is set somewhere other than J100's threshold for his individual acceptance of that proposition. In the example just given, the collective decision rule – the rule governing the jury's deliberations – was 'it takes 90 votes to convict'. However, the juror's own decision rule (as reflected in his hypothetical Bayesian updating) required only 89 'guilty' votes to make it more than 90 per cent likely that the accused was indeed guilty. Juror J100 voted 'guilty', even though his private information indicated 'innocent', because if there were 89 guilty votes already on the table, that would be sufficient to take him over the '90 per cent certain the accused is guilty' threshold.

So what is the lesson? Simply that we have to make sure the collective (group) decision rule faithfully mirrors the decision rule of individuals within the group, as regards the threshold required for acceptance of a proposition. That is necessary, not only to avoid strategic misrepresentation of beliefs, but also to make sure the collective decision comes out the correct way from the point of view of members of the group.

IV

So far I have been assuming everyone in the group is identical in respect of everything except the private information that they possess. Among other things, I have been assuming that their 'acceptance threshold' is the same. If it is, then it is easy in principle to see how to implement the injunction to make the group acceptance threshold the same as the individual acceptance threshold.

If individuals differ in that respect however, then that recommendation is not straightforwardly implementable, and serious problems might arise in consequence. In Austen-Smith and Feddersen's (2006) example, jurors J1 and J2 have

individual acceptance thresholds requiring three out of three votes to convict, whereas juror J3 has an individual acceptance threshold requiring only two out of three votes to convict. Reasoning game theoretically, juror J3 once again focuses purely on the case in which he is pivotal. If either J1 or J2 (much less both) say the accused is innocent, she is acquitted whatever J3 does. The only case in which J3's vote matters is that in which both J1 and J2 vote 'guilty': then if J3 also votes 'guilty' the accused is convicted; if on the other hand J3 votes 'innocent' the accused will be acquitted, since J1 and J2 think all three must vote 'guilty' to convict her. Remember, however, J3 thinks only two 'guilty' votes should be required to convict; and in the only case that matters (the case in which he, J3, is pivotal) the other two jurors will have voted 'guilty'. So J3 should strategically vote 'guilty', in order to secure the conviction that by his (lower) standards is warranted, even if J3's own private information indicates the accused is innocent.⁴

'Guilty' truly would be the right verdict from J3's point of view, thinking as he does that two votes should suffice to convict. But if J1 and J2 had known that J3's private information was that the accused was innocent, they would have insisted on acquittal, thinking as they do that it should take three votes to convict.

V

The upshot seems to be that only certain kinds of groups can be expected to deliberate together successfully. Doubtless there are many requirements for that. One of them, the above example shows, is that the members of the group must have an agreed standard, common to every member of the group, as regards what the 'acceptance threshold' for a proposition should be (and, further, the group's collective 'acceptance threshold' must be the same as that 'acceptance

threshold' that is common among all members of the group).

Who would have thought otherwise? Voting rules have to be settled before we can decide what the outcome of a vote is. If some members of the group think anything that gets a bare majority wins, whereas others think a two-thirds majority is required to win, then any outcome between half and two-thirds is a contested outcome from the point of view of some members of the group. The group as such has not made a decision at all in such cases.

Similarly in a group that deliberates rather than votes: if some members of the group think the standard of proof is 'more likely than not' while others think it is 'beyond a reasonable doubt', then any outcome between those two standards is a contested outcome from the point of view of some members of the group. Once again, the group as such has not made a decision at all in such cases.

If we imposed a two-thirds rule on such a mixed group, then those who privately thought half should be enough would of

course be tempted to strategically misrepresent their beliefs in precisely the way game theorists show. Ditto if we imposed a 'beyond a reasonable doubt' standard on a mixed group, those who privately thought that 'more likely than not' should be enough to settle the issue would be tempted to strategically misrepresent their beliefs. But that is simply to say that, from a deliberative point of view, the group was not well-formed in the first place.

To be well-formed from a deliberative point of view, groups thus seem to need to be composed of individuals with identical 'acceptance thresholds' (and collective and individual acceptance thresholds should be identical). The good news is that, so long as such groups operate according to simple majority rule, no one would be tempted to lie. The bad news from the deliberationist point of view is that that sets the standard of what is required for a well-formed deliberative group substantially higher than most real-world groups can realistically aspire to achieve.

Notes

1 I am grateful to Christian List for discussions of these issues.

2 We might distinguish 'strategic misrepresentation of preferences' of that sort from 'strategic compromise'. Suppose in a situation of plurality voting you prefer $A > B > C$, but you think A has no chance of winning even with your vote, and C would win unless you voted for B . Voting for B instead of A in those circumstances is strategic voting, but of an innocuous sort (Dowding and Van Hees, 2007).

3 Alternatively, juror J100 might have private information indicating 'guilty' but vote to acquit because, given J100's prior and conditional probabilities, he would not think it is 90 per cent likely that the accused is guilty if there were nine jurors out of the 100 reporting 'not guilty'. That might be more common in practice (Davis *et al*, 1988).

4 And in Austen-Smith and Feddersen's (2006) two-step procedure he should also strategically lie about his private information in the first round, to get what he sees as the right/true result in the second round.

References

- Austen-Smith, D. and Banks, J.S. (1996) 'Information aggregation, rationality and the Condorcet Jury Theorem', *American Political Science Review* 90(1): 34–45.
- Austen-Smith, D. and Feddersen, T.J. (2006) 'Deliberation, preference uncertainty and voting rules', *American Political Science Review* 100(2): 209–217.
- Davis, J., Stasson, M., Ono, K. and Zimmerman, S. (1988) 'Effects of Straw Polls on group decision-making: sequential voting pattern, timing and local majorities', *Journal of Personality and Social Psychology* 55(6): 918–926.

- Dowding, K. and Van Hees, M. (2008) 'In praise of manipulation', *British Journal of Political Science* 38(1): 1–16.
- Farquharson, R. (1969) *Theory of Voting*, Oxford: Blackwell.
- Feddersen, T. and Pesendorfer, W. (1998) 'Convicting the innocent: the inferiority of unanimous Jury verdicts under strategic voting', *American Political Science Review* 92(1): 23–36.
- Steiner, J. (2008) 'Concept stretching: the case of deliberation', *European Political Science* 7(2): 186–190.

About the Author

Robert Goodin is Distinguished Professor of Social and Political Theory and of Philosophy in the Research School of Social Sciences at Australian National University. He is founding editor of *The Journal of Political Philosophy* and general editor of the 10-volume series of *Oxford Handbooks of Political Science*. He is author of, most recently, *What's Wrong with Terrorism?* (Polity, 2006).