Gibbard’s Principle of Commitment

Adam Morton

A central form of argument in Allan Gibbard’s *Thinking how to live* first occurs in chapter 5 where Gibbard argues that we are all, whatever our views on what we ought to do, committed to the claim that the attribute of actions “being okay” supervenes on “prosaically factual properties”. (Gibbard has defined both terms prescriptively, but not unreasonably. The definitions don’t affect the issue here.) The argument is based on considerations of hyperstates, consistent extensions of one’s present state of mind in which one has a belief about every question of fact and a plan for every possible contingency. Gibbard assumes the

> *Principle of Commitment (PC)*. A person is committed to a claim Q if in every hyperstate he could reach without changing his mind he would accept Q.

“Without changing his mind” means “coming to believe nothing that he denies now”; it consists in adding beliefs that are originally neither believed nor disbelieved. He argues that in every hyperstate a person’s beliefs and plans would determine an extension for what is ok to do. This is a plausible conclusion: every possible act, described naturalistically, is either ok to do or not according to one’s hyperstate, and thus the disjunction of all the descriptions of the okay acts plus the negation of the disjunction of all the not okay acts picks out a unique set. Then he invokes PC to conclude that even finite people with very far from complete beliefs and plans should take there to be a prosaically factual description on which all the acts that are okay supervene.

This note concerns the principle of commitment, not claims about supervenience, or the other related uses Gibbard puts the principle to later in the book. I shall argue that the principle is ambiguous. Taken one way the principle is false: no sane person would take herself to be committed to something just because she would hold it in every way she could come to have a complete and consistent extension of her present beliefs. And taken the other way, the principle is true, but cannot do the semantical work that Gibbard wants it to.

What any naïve extension of me would believe Begin with my beliefs, as they are at this moment. Now just add beliefs, using concepts available to me, arbitrarily but never contradicting any belief already added, until we get a set of beliefs which is consistent and also complete. For every *p*, exactly one of *p* and *not p* is in the set. Call this a naïve extension of my beliefs. A complete and consistent set of beliefs naively extending those of a finite uncertain creature will be very different from that of any imaginable agent with a real psychology, even one who is decided on all matters. In particular a real creature in a hyperstate would be aware of many ways in which her beliefs were different from those of a more undecided creature. Contrast me, in my definitely hypostate, with any hyperstate of that state. I do not have beliefs about the genders, parties, or provincial origins of the next three prime ministers of Canada, but since a real creature whose opinions were decided on that topic would have such beliefs she would also have beliefs to the effect that her beliefs about the next three prime ministers were complete. In fact since a little reflection would show her that she had beliefs about the genders etc of all the prime ministers to the end of Canadian history, she would believe "my beliefs about the genders, parties, and origins of future Canadian prime ministers are complete." She would also believe "my opinions are much more definite than those of normal human beings." But *no* naïve consistent and complete extension of the beliefs of my or any other normal human would contain such beliefs. For since I believe that I don't believe that the next prime minister will be a woman (and believe that I don't believe that it will be a man, either) I believe that my beliefs about future prime ministers are incomplete. And so any complete extension of my beliefs must contain that belief: any complete extension of me will believe that it is incomplete.

There is a long list of propositions *p, q, r,* … that I neither believe nor disbelieve. For some of these I am aware of my ignorance, and so I believe that I do not believe *p* and also believe that I do not believe *not p*. So although neither *p* nor *not p* are among my beliefs *I do not believe that p* and *I do not believe that not* p are. A hyperstate extending my beliefs would be complete and consistent, so I would either have added *p* to my beliefs or have added *not p.* In the first case hyper-me would have among its beliefs: *p*, *I do not believe that p, either p and I do not believe that p or not p and I do not believe that not p*. (The third of these follows from the first two.) In the second case hyper-me would have among its beliefs: *not p, I do not believe that not p, either p and I do not believe that p or not p and I do not believe that not p*. These are extremely peculiar combinations of belief. In many cases they are not compatible with functioning as a real epistemic and practical agent. But if we rule them out, restrict ourselves to sets of beliefs that a real agent can have, then PC is clearly false. It then entails that you should believe your beliefs are complete when they are not.

The fact is that any rational way of adding beliefs requires taking away others. A set of beliefs constructed just by addition is not that of any creature with a real psychology, not that of a possible agent. Gibbard has a recurrent device of personifying creatures with complete and consistent states of mind with the names of Greek gods, thus making us think of them as real thinking agents. But the observations so far have refuted PC if it is understood in terms of such agents. That is, *it is not the case that a person is committed to a claim Q if in all hyperstates he could reach without changing his mind and while remaining a functioning agent he would accept Q.* This conclusion is particularly significant in the context of Gibbard's larger project, in which the states contain plans as well as beliefs: I just don't know what counts as a plan for a being that has not the minimal accuracy about its own states to qualify as a rational agent.

What some sophisticated (and all naïve) extensions of me would believe. PC is motivated by an analogy with the semantics of first order logic. A sentence **s** follows from a set of sentences **P** iff it is true in all models in which all sentences in **P** hold, and given the completeness of first order logic this is equivalent to saying "**s** follows from **P** iff **s** is contained in all consistent and complete extensions of **P**" (This point is most explicit in completeness proofs in the style discovered by Leon Henkin, for which a classic source is Church 1956, sections 44 and 45.) So, moving swiftly from sentences to beliefs, and from first order logical consequence to what one is logically committed to, we get the principle, that you ought to believe something if you would believe it in every way you could get to a complete and consistent belief system without changing your mind about anything. But the analogy with logic contains a warning, too. When we move from first to second order logic the picture is not so simple. When we expand a set of second order sentences to a complete and consistent set we find that what is syntactically complete may not be semantically complete: the syntactically complete sets of sentences contain some perverse "non-standard ones" which make claims that are intuitively necessarily false. Typically these sentences will underestimate how many sets there are, or overestimate how many numbers. (See section 54 of Church 1956 or, more readably, chapters 5, 6, 7, 12 of Grandy 1977.) Arcane as these facts may seem, they are relevant to the issues here. The connection is that hyperstates may include strange beliefs that block the uses Gibbard makes of PC.

We saw above that complete and consistent naïve extensions of my beliefs would contain the belief that my beliefs are not complete. We can avoid that conclusion by defining the transition to a complete and consistent set of beliefs in a more nuanced way, which preserve psychological plausibility. One way would be to impose a stratification on the beliefs, so that in its beliefs about beliefs the final set distinguishes between beliefs that were originally held and those that were added in the hyperization. Then the set can in effect say "I originally believed neither p nor not p, but now I believe p". Call this a sophisticated extension. Some sophisticated extensions of my beliefs will not be absurd things for a rational agent to believe. But it is still not true that all sophisticated extensions of my beliefs will contain the belief that they are complete, or in other ways accurately represent their own structure. For to get from my present set of beliefs to a complete and consistent set one has to add infinitely many more, only a few of which will have any relation to anything that I believe now. So there is plenty of room for adding beliefs misdescribing the general character of my beliefs in the hyperstate. In particular a person in a hyperstate might believe, falsely, that there are instances of some predicate which her beliefs fix neither one way nor the other. She might believe "a is P", "b is P", " c is P". and so on, for all the cases a,b,c, … in which P applies, and similarly for "z is not P", "y is not P", "x is not P", and so on for all the cases in which it does not, and *also* believe "there is some individual for which I believe neither that it is P nor that it is not P". As a result she might believe "my (present) beliefs are not complete". (This would be analogous to omega-inconsistency in logic, and thus to the existence of non-standard extensions: a consistent extension of a system of arithmetic might contain the sentence Pn for each n but also contain the sentence ∃n ~Pn.)

This is a problem for the use Gibbard has for PC. He argues that we should believe various claims about the semantics of normative terms like "ok" or "is the thing to do" because these claims would be assented to in all hyperstates extending our present states. (See particularly pp. 91-3, 94-5, 96-7, 170-1.) All these claims require that there be a property of a certain kind that is coextensive with the normative attribute in question. And Gibbard's argument that they would be included in all hyperstates is that in each hyperstate one would be decided on their coincidence in each particular case. That is, there would be a P of the required kind such that for each instance N(a) in which one believed that the normative term applied one would also believe that P(a) held. (Which P in the kind in question will vary from one hyperstate to another: the point is that there always is one.) So, he argues, in each hyperstate one would believe that N is coextensive with some property in the kind, and therefore one should believe now that N is coextensive with some such property. But now the problem is clear. From the outside we can see that the hyper-beliefs determine an extension for N, but from the inside it may look quite different, since there may also be beliefs that there is some b for which N(b) to which no such property applies. So the complete and consistent extension of my beliefs may include a denial of the claim that N is coextensive with some suitable P. PC does not then apply. Some stronger principle is needed.

These doubts may seem pretty rarefied and sophisticated. But the conclusion is simple. If by "hyperstate" in PC we mean "not-naïvely complete" then PC is false. If we take "hyperstate" to include naïve extensions of one's present state then PC is true, but now there is such a range of hyperstates that PC will not support the uses Gibbard makes of it. I would draw an equally simple moral from the conclusion. Gibbard uses considerations about what we would accept if our beliefs and plans were not so limited in order to defend conclusions about the attitude we should have to our beliefs and especially plans in our actual limited state. But in fact limitation and planning are not so separable: many of our plans are there as reactions to our limitations. If we were less limited we would make very different plans. (For dramatic examples of how planning for limited agents differs from planning for unlimited ones see chapter 4 of Rubinstein 1988.) As a result, plans that project from unlimited versions of our actual states may either insinuate inappropriate limitation-relative elements into an unlimited state or suggest elements that are only appropriate in the unlimited case as models for our limited selves. Neither provides much guide for what we, as we are, should do and think.

Church, Alonzo (1956) *Introduction to mathematical logic* vol I, Princeton NJ: Princeton University Press .

Gibbard, Allan (2003) *Thinking how to live*. Cambridge, MA: Harvard University Press

Grandy, Richard (1977) *Advanced logic for applications*. Dordrecht: Reidel.

Rubinstein, Ariel (1998) *Modelling bounded rationality*. Cambridge, MA: MIT Press