Endorsement-logic, simulationism, and the cogito

Benj Hellie

Cogito in Ligerz, 22 April 2017

1 A *cogito*-like argument for Subjects

Δ A *Subject* is an individual with mental properties; a *Cartesian* believes that there have been subjects (indeed, there are some now).

Δ *Autopsychological* sentences are exemplified by 'I (do not) think/doubt/believe/have evidence that/whether φ' or 'I intend to Γ'.

C1. Autoinfallibilism

When the matter of whether ψ is autopsychological, the supposition that so-and-so's belief whether ψ is mistaken is *incoherent*.

C2. Ascriptivism

The mode of presentation of a 'simple' autopsychological belief has the logical form Ψi, where i is a term (denoting some individual, perhaps in a contextually varying manner) and Ψ is a predicate denoting a mental property.

C3. Someone has had (indeed, now has) a simple autopsychological belief.

∴ There have been (indeed, are now) Subjects.

I propose to grant Autoinfallibilism (and the minor premiss); Ascriptivism, however, will be rejected.

2 The *cogito* as evidence?

Δ *Evidence*: k-at-t treats (proposition) e as evidence for j-at-$t' \Rightarrow$ k-at-t believes that:

(i) e is true

(ii) j-at-t' treats e as evidence for j-at-t'

Δ *Egocentricity*: reflecting on what it is like, I note that I perpetually treat a certain human being, BH, as the 'center of the world' (*Center*); and that, at each moment t, I treat t as the 'limit of history' (*Limit*)

Δ *Sensory condition*: an animal's sensory condition over an interval is the course of sensible properties instantiated in its *ecology*—its body in relation to the sensible region of its environment (or, perhaps, some 'narrowing' of that ecology)

Δ *The canonical subject-matter of evidence*: an attentively curated aspect of the sensory condition of the Center (as such) prior to the Limit (as such) (Ryle: 'The sensation is in no sense 'mental' ')

Δ Beliefs held under the *evidential mode of presentation* have a predicate–term logical form; canonically, the term denotes the Center and the predicate denotes the attentively curated aspect of the course of its sensory conditions up to the Limit

Evidentialism

j-at-t has the autopsychological belief that ψ only if j-at-t treats the content of ψ as part of their evidence

- Consequences:
  - Ascriptivism follows directly
  - 'Positive' Autoinfallibilism follows (namely, the restriction to 'simple' ψ);

- The 'negative' direction, however, does not: autopsychological belief is not sensory, so the restriction to canonical subject-matter is lifted, so evidence is potentially limitless; moreover, evidence is 'attentively curated'
- That does not affect the argument, as the minor premiss interfaces only with the positive direction

3 Vehicle–state ambiguity

- An ambiguity in *thinking*
  - *V-thinking*: the auditory imagery serving as the *vehicle* of reasoning through 'talking to oneself'
  - *S-thinking*: the mental *state* made explicit to oneself in reasoning
- A corresponding ambiguity in Evidentialism
  - *V-Evidentialism*: *j*-at-*t* has the V-autopsychological belief *I V-think 'φ'* only if *j*-at-*t* treats it as evident that, at the Limit (for *j*-at-*t*; as such), the Center (for *j*-at-*t*; as such) V-thinks 'φ'

    ✓: the V-belief just involves targeting attention on the verbal imagery running through one's head and thereby having the evidence that one is beset with a course of verbal imagery of that sort
  - *S-Evidentialism*: *j*-at-*t* has the S-autopsychological belief *I S-think that φ* only if *j*-at-*t* treats it as evident that, at the Limit (for *j*-at-*t*; as such), the Center (for *j*-at-*t*; as such) S-thinks that φ

    ×: the canonical subject-matter of evidence is exhausted by sensible features of one's ecology; I find no plausibility whatever in the claim in amongst the colors, tastes, shapes, itches and the like are polkadotted beliefs, doubts, intentions and the like
- Bad news for the Cartesian:
  - It is S-Evidentialism that is needed to establish that there are Subjects
  - For that matter, V-evidentialism is dialectically vulnerable: if the Demon has wiped out my environment and my body, my verbal imagery scarcely stands a chance

4 Attaching the 'I-think'

Insensitive Nonsaturationism

The logical form of the mode of presentation of the autopsychological belief *I believe that φ* has logical form $\mathsf{B}\varphi$, where B is context-insensitive in its denotation

Truth-logic

- *Objective correctness*: A belief with lf φ held in context *c* is *correct* just if $[\![\varphi]\!]^c$, the *c*-content of φ, is true in *c*; *mistaken* just if $[\![\varphi]\!]^c$, the *c*-content of φ, is false in *c*.
- *Truth-preservation*: $\Psi \vdash \varphi$ just if whenever each Ψ is correct in *c*, so is φ.
- *Contexts as centered worlds*: Fred-at-*t* is in *c* just if *c* represents exactly how Fred is at *t*; let w_c, t_c, and j_c be the world, time, and individual determined by the context *c*.

Truth in a world

Belief-contents are *propositions* (sets of worlds); a proposition *p* is true in *c* just if $w_c \in p$; false in *c* just if $w_c \notin p$.

- *Faultless disagreement*:

  Let Fred have the autopsychological belief *I believe that goats eat cans*, with logical form $\mathsf{B}\gamma$; and let Sam have the autopsychological belief *I do not believe that goats eat cans*, with logical form $\neg\mathsf{B}\gamma$.

  1. By Autoinfallibilism, neither is mistaken.

2. Let Fred's context be c', Sam's c''; let Fred and Sam be world-mates, so that $w^* := w_{c'} = w_{c''}$. $\mathsf{B}\gamma$ is context-insensitive, so $[\![\mathsf{B}\gamma]\!]^{c'} = [\![\mathsf{B}\gamma]\!]^{c''}$; so for any w, exactly one of $w \in [\![\mathsf{B}\gamma]\!]^{c'}$ and $w \in [\![\mathsf{B}\gamma]\!]^{c''}$; so exactly one of $w^* \in [\![\mathsf{B}\gamma]\!]^{c'}$ and $w^* \in [\![\mathsf{B}\gamma]\!]^{c''}$; so exactly one of Fred's and Sam's autopsychological beliefs is mistaken—contradiction.

5 First-personalism

Truth in a centered world

Belief-contents are *centrifugal propositions* (sets of centered worlds—world, time, individual triples); a *properly* centrifugal proposition includes some $\langle w, t, j\rangle$ but excludes some $\langle w, t', j'\rangle$; a centrifugal proposition π is false in c just if $\langle w_c, t_c, j_c\rangle \notin \pi$.

- Relativist semantics for B:
  - For nonpsychological φ, $[\![\varphi]\!]^c$ is improperly centrifugal
  - $[\![\mathsf{B}\varphi]\!]^c$ is properly centrifugal: if Fred's context c determines π_c as Fred's belief-content in w_c at t_c, then $c \in [\![\mathsf{B}\varphi]\!]^c$ just if $\pi_c \subseteq [\![\varphi]\!]^c$
  - If $[\![\varphi]\!]^c$ is properly centrifugal, $[\![\mathsf{B}\varphi]\!]^c = [\![\varphi]\!]^c$—so in particular $[\![\mathsf{BB}\varphi]\!]^c = [\![\mathsf{B}\varphi]\!]^c$ and $[\![\mathsf{B}\neg\mathsf{B}\varphi]\!]^c = [\![\neg\mathsf{B}\varphi]\!]^c$, in line with Autoinfallibilism
- Worries:
  - The notion of *content* gets its job as being that which evolves monotonically when a subject doesn't change their mind, and which is passed around in conversation; Relativism can't allow this; to what then do they anchor their conception of content?
  - Relativism is straightforwardly reparenthesized into Ascriptivism; the above worry generates pressure to do so.
  - How, for Fred's context $\langle w^*, t^*, j^*\rangle$, is j^* related to Fred? Dilemma: (i) If j^* is Fred (or *supervenes on* Fred—is such that, for any w, t, which properties Fred has in w, at t determines which properties j^* has in w, at t), the Cartesian's mental properties and the Relativist's centrifugal contents of autopsychological belief become hard to discriminate;

    (ii) If j^* does not supervene on Fred, then potentially whether $\mathsf{B}\varphi$ is true of Fred at a time does not supervene on which properties Fred has then—undermining the explanatory significance of autopsychological belief as regards, say, avowal-behavior.
- Extending to allopsychological belief:
  - Let Fred's belief *Belkis-at-t' believes that goats eat cans* have lf $\S^{t^*,j^*}\mathsf{B}\gamma$, where t is determined by t' and j^* by Belkis; and let the content in Fred's context be true at $\langle w, t, j\rangle$ just if the content in Fred's context of $\mathsf{B}\gamma$ is true at $\langle w, t^*, j^*\rangle$ (B turns improper to proper, § turns proper back to improper).
  - Let n ('now') denote, relative to c, the Limit for the individual in c, at its time in its world; and let i ('I') denote, relative to c, the Center for the individual in c, at its time in its world.
  - Note that $\S^{n,i}\mathsf{B}\varphi$ ('immanent self-ascription') is improperly centrifugal, while $\mathsf{B}\varphi$ remains properly centrifugal ('transcendent avowal'). While these (I conjecture) entail one another, it does not appear to be the case that $\mathsf{B}\S^{n,i}\mathsf{B}\varphi$ must also be equivalent: existing constraints on B concern only application to properly centrifugal propositions; by immanentizing, autopsychological belief loses its status as in any way distinctive.

6 Endorsement-logic and fundamental self-knowledge

Endorsement-logic

- *Nonobjective correctness*: A belief with lf φ held in context c is *treated as correct in context c'* just if c' endorses $[\![\varphi]\!]^c$; *treated as mistaken in context c'* just if c' antiendorses $[\![\varphi]\!]^c$; when $c' = c$, we say the belief is *reflexively correct/mistaken*

- *Endorsement-preservation*: $\Psi \vdash \varphi$ just if whenever each Ψ is reflexively correct in c, so is φ.
- *Contexts as mental states*: Fred-at-t is in c just if c represents Fred's mental state at t; let p_c, e_c, t_c, and j_c be the belief-content of the subject of c, their evidence, their Limit, and their Center.

Endorsement in a mental state

Belief-contents are *propositions*; p is *endorsed* in c just if $p_c \subseteq p$, *antiendorsed* just if $p_c \subseteq \overline{p}$, otherwise *neutral*

- *The 'Dart' operator*: the c-content of $\triangledown\varphi$ is trivial just if the c-content of φ is believed in c, otherwise absurd (analogously to the familiar *actuality* operator)
  - Observe that φ and $\triangledown\varphi$ are equivalent
  - While $\neg\varphi$ entails $\neg\triangledown\varphi$, the converse is not so
  - Identifying the lf of autopsychological belief *I believe that* φ with $\triangledown\varphi$ therefore directly yields Autoinfallibility

7 Simulation and indexation

- Can't use point-shifters, as $\triangledown$ rigidifies to the context, and there can be no context-shifter; instead, something like the Stalnaker conditional. Rough idea is to capture *simulationism*: my psych-ascription *Belkis believes that goats eat cans* expresses my sentiment that, changing myself around to comport with how I think Belkis is, the result believes that goats eat cans.
  - Belief that *j-at-t believes* φ has lf $B^{t,j}\varphi$; the c-content of $B^{t,j}\varphi$ is trivial just if the c-content of φ is believed in $c \gg t, j$; otherwise absurd.
  - $c \gg t, j$ is the nearest context c' to c among those for which $e_{c'}$ is the strongest proposition the subject of c treats as the evidence of j-at-t.
- *Faultless disagreement*:

  If Belkis and Ruth are interpreting Fred, then even if they believe the same about Fred's evidence, differences in their starting point can lead to differences in what they take Fred to believe. But neither would treat the other as mistaken.

- *Autoinfallibility*:

  The lf of autopsychological belief *I believe that* φ, in this scheme, is $B^{n,i}\varphi$.

  Suppose every context is unexceeded in similarity to itself; and recall that j-at-t has evidence e just if j-at-t treats j-at-t as having evidence e: then $c = c \gg t_c, j_c$. Recall moreover that the c-denotations of n and i are t_c and j_c: then $B^{n,i}\varphi$ has trivial c-content just if the c-content is believed in $c \gg t_c, j_c = c$—and is therefore equivalent to $\triangledown\varphi$.

8 Morals

1. The V-cogito does not establish the existence of Subjects; and while V-Evidentialism is prima facie plausible and would support both V-Autoinfallibility and V-Ascriptivism, it appears to be on a par with other principles about my evidence in their vulnerability to the Demon
2. An S-cogito would establish the existence of subjects. But the straightforward approach, going by way of S-Evidentialism, is unavailable, because the latter is highly implausible.
3. Moreover, with S-Evidentialism out of the picture, it becomes a nice question what could undergird S-Autoinfallibility. If the infallibility is not of the 'empirical' sort accorded by evidence, the natural alternative is the 'rational' sort accorded by connections of meaning. A good place to start is with the idea of 'attaching the *I-think*' to a thought: a simple operator that transforms the meaning of its operand in such a way that the two are equivalent.
4. An initial try crashes immediately into the *faultless disagreement* problem. To resolve it, we move to a relative-truth approach. That turns out

to be unpromising: it is hard to make sense of prima facie, hard to distinguish from Cartesianism prima facie, and the best bet for making a distinction collides with psychological explanation. Moreover, extending to allopsychological belief yields a clash between transcendental avowal and immanent self-ascription.

5. Instead, I propose, we should work with an *endorsement-logical* conception of meaning on which belief can be assessable for correctness or mistake *directly* on the basis of the mental state one occupies, without any involvement from the world. With such a conception, we may think of belief about one's mental state as getting its distinctive significance from the range of mental states that make it correct, without also thinking of the belief as imposing any specific condition on the world. If the belief imposes no condition on the world, there is nothing surprising in its infallibility; there is, in particular, no motivation for rolling it in with evidence. Armed with the endorsement conception, we give sense to a conception of mentality on which it is essentially self-revelatory, but also not in any sense an aspect of the world. In particular, the $\triangledown$-operator models a fundamental sort of self-revelation of the proximal mental state, entirely unlike empirical knowledge, and untainted by extraneous demands for comparison with distal mental states.

6. Now of course we can compare proximal and distal mental states. This requires some sort of indexation by particulars in order to net the distal state in the first place, or for that matter to mark it as recollected or anticipated, or as the state of the other. Fortunately, this indexation requires no enrichment of our ontology: egocentricity serves as a 'pivot' between the objective point of view on an individual at a time and the view of a mental state taking that individual and time as center and limit.

7. Calling this 'indexation' is not a euphemism for 'predication'. The origin in simulation of a consequent nonobjectivity makes it clear that in mental ascription, I do not return to imposing a condition on the world or one of its constituents. Rather, the meaning is exhausted by its display of my commitment to a specific manner of empathy with the indexing individual at the indexing time.

8. Summing up: S-Autoinfallibility is plausible, and undergirded by the correct *logic*, which demands of mentality that it be self-revelatory. But at the most fundamental level, there is no trace of S-Ascriptivism, as my mental state's self-revelation is unconcerned with anything else. The prospect for cross-comparison does not restore S-Ascriptivism, as it involves indexation to animals of broadly modal operators rather than predication to Subjects of mental properties.