Risposta ai commenti

Anna Elisabetta Galeotti*

Reply to Comments
Abstract: In my Reply to comments, I take up a rehearsal both of the intentional and of the motivationist-causal account of self-deception. On the one hand, I explain the role of intentions in my account, the instability of the self-deceptive beliefs, and the emotional instability of self-deceivers. On the other, I illustrate how my solution to the selectivity problem works and why I think that the role of agency in the motivationist account is negligible in the causal process leading to self-deception. Finally, I engage with some cognitive theories and, especially, with the theory of argumentative reasoning, which, at first sight, seems to contradict the possibility of collective self-deception. Yet, at a deeper examination, such theory appears to be supportive of the way collective self-deception is produced especially in politics.

Keywords: Intentional model, Instability, Motivationist-causal model, Selectivity, Agency, Argumentative reasoning.

I tre commenti al mio libro sono fra loro complementari in quanto sollevano quesiti diversi in relazione a tre diverse dimensioni filosofiche del trattamento dell'autoinganno. Antonella Besussi mi invita a riflettere sulla natura dell'autoinganno, specificamente con riferimento all'opzione intenzionalista che io scarto per proporre un resoconto – la versione a mano invisibile dell'autoinganno – che si presenta come una revisione del modello causale-motivazionista. Besussi suggerisce che quest'opzione forse perde qualcosa che l'intenzionalismo invece cattura. Patrizia Pedrini si colloca invece entro il perimetro del motivazionismo e, da questa prospettiva, solleva delle obiezioni al mio resoconto che considera come più vicino al modello causale-motivazionista di quanto io non creda. Infine Cristina Meini mi sollecita a confrontarmi con alcune teorie all'interno delle scienze cognitive e della psicologia clinica che, da una parte, amplierebbero la riflessione filosofica e, dall'altra, forniscono evidenze empirico-sperimentali che potrebbero sostenere o falsificare l'analisi filosofica.

Cominciando dai commenti di Besussi, nel mio resoconto le intenzioni non sono estromesse, come Besussi stessa riconosce, e rivestono un ruolo importante che, tuttavia, è assai diverso da quello loro assegnato dagli intenzionalisti. L'intenzionalismo considera l'autoinganno come prodotto intenzionalmente dal soggetto, in parte inconsciamente, per evadere l'evidenza che porterebbe a credere che non-p, contro i desideri del soggetto. L'approccio intenzionale all'autoinganno è foriero di paradossi, e più precisa-

* Professoressa di Filosofia Politica, Università del Piemonte Orientale.

mente, del paradosso statico, per cui il soggetto finirebbe a credere contemporaneamente che p e che non-p, e del paradosso dinamico, relativo al fatto che il soggetto, sapendo che *non-p*, formerebbe intenzionalmente la credenza che *p* corrispondente ai suoi desideri. Detto in altri termini, l'intenzionalismo interpreta l'autoinganno letteralmente come mentire a se stessi, in modo tale che il soggetto risulta contemporaneamente sia l'agente ingannatore che la sua vittima. I sostenitori dell'approccio intenzionale hanno cercato di risolvere questi paradossi in vario modo, attraverso la divisione della mente tra inconscio e conscio (Davidson, 1982 e 1985; Pears, 1984), oppure attraverso una sofisticata quanto irresolutiva considerazione logica (Foss, 1980), o mettendo in dubbio che il risultato dell'autoinganno sia una credenza vera e propria Fingerette, 1969; Rorty, 1972; Audi, 1982), oppure invocando la non trasparenza (Gardner, 1998; Talbott, 1995) e la non linearità (Davidson, 1985). In generale, nessuno di questi tentativi risulta soddisfacente, anche perché alla base non è chiarito quale sia l'intenzione necessaria per definire intenzionale l'autoinganno[1]. Gli intenzionalisti "deboli" alla Fingerette (1998) ammettono che non è l'intenzione di autoingannarsi: ma allora quale intenzione vale? Non può essere l'intenzione di placare l'ansia generata da *non-p*. Perché quella certamente non basta a rendere la falsa credenza che *p* un prodotto intenzionale. E allora quale sarebbe l'intenzione cruciale per definire l'autoinganno un risultato intenzionale? Io ritengo che gli intenzionalisti confondano l'intenzionalità del processo con quella del risultato. Se le due cose vengono propriamente distinte sul piano analitico, allora è possibile salvare l'intenzionalità del processo, riconoscendo al contempo la non intenzionalità del risultato. Il modello di spiegazione a mano invisibile dà propriamente conto del processo e dell'esito non pianificato, spiegando anche l'apparenza teleologica dell'autoinganno e la soddisfazione del desiderio del soggetto nel preservare che p. In altre parole, l'autoinganno è il prodotto di passi intenzionali dell'agente rivolti non formare la credenza falsa che p, ma a trovare una via d'uscita dall'ansia generata dalla evidenza negativa. Il ragionamento dell'agente è pesantemente affetto da "conformity bias" e altre bias cognitive, e alla fine la credenza che p si forma senza bisogno di alcun disegno intenzionale.

Quanto all'instabilità, ampiamente descritta come tipica della fenomenologia dell'autoinganno (Audi, 1989; Pedrini, 1913), essa è in relazione ad alcuni tipi di autoinganno, ma non a tutti. In particolare, gli autoinganni che servono a coprire fallimenti, errori e figuracce, ossia volti a ridurre la dissonanza cognitiva tra l'immagine di sé e l'insuccesso, che io chiamo il modello "uva acerba" (Elster, 1983), rappresentano in generale una condizione più stabile, perché l'evidenza negativa è stata elaborata una volta e sistemata nella spiegazione epistemicamente distorta che risolve il problema psicologico del soggetto. Invece quando l'oggetto dell'autoinganno riguarda qualcosa che sta accadendo, tipo un'infedeltà coniugale, o un problema di droga del figlio, in questo caso la credenza falsa è instabile perché l'evidenza negativa non cessa di arrivare al soggetto, che deve pertanto trovare continue e sempre meno credibili spiegazioni per tacitarla. In questo caso, il soggetto non è in disaccordo con sé stesso, nel senso che non è preoccupato dal suo stato doxastico, bensì dall'inquietudine e dall'ansia provocate dall'evidenza negativa. Non è necessario supporre che il soggetto possegga, a livello inconscio, la credenza che *non-p*, perché l'evidenza negativa è senz'altro percepita, ed è ciò che mette in moto il processo dell'autoinganno, ma viene bloccata a un livello

di rappresentazione inferiore a quello della credenza, dato che l'inferenza da *non-e* a *non-p* viene bloccata. Quindi il soggetto sa che il marito viene a casa sempre più tardi, che riceve telefonate sospette al fine settimana, che ha molte conferenze a cui all'improvviso deve andare, ma queste informazioni non si traducono nell'inferenza dell'infedeltà del marito, inferenza che, in assenza del desiderio, sarebbe ragionevole trarre, ma richiedono piuttosto elaborate e laboriose spiegazioni in linea con la credenza desiderata 'Mio marito mi è fedele'. Il desiderio semplicemente inclina il ragionamento in una certa direzione pregiudiziale; l'intenzione di trovare una spiegazione a questi fatti è in sé legittima; la distorsione si produce nella considerazione selettiva dell'evidenza e nella conclusione, ossia quando il ragionamento si ferma perché ha trovato una spiegazione che riallinea i fatti col desiderio, non importa quanto poco plausibile.

Dall'altra parte, il modello motivazionista-causale evita i paradossi e offre una spiegazione più semplice e lineare, secondo la quale la motivazione del soggetto attiva bias cognitive che informano il trattamento dei dati e producono la credenza falsa che p, contro l'evidenza disponibile. Dopo tutte le complicazioni per trovare una soluzione al di là dei paradossi, questo resoconto, proposto per primo da Alfred Mele (1997, 2001) è risultato convincente alla gran parte della comunità di studiosi che si occupa di autoinganno ed è in breve diventato dominante in letteratura. Come Pedrini, condivido questo punto di svolta e tuttavia sono convinta che la spiegazione di Mele presenta alcune debolezze. Tra esse la difficoltà a rendere la specificità dell'autoinganno nell'ambito della irrazionalità motivata e il problema della selettività, a cui appunto Pedrini fa riferimento. Il problema della selettività, originariamente evidenziato da Talbott (1995) e poi da Bermudez (2000), fa riferimento al fatto che non tutte le volte che la realtà frustra un nostro desiderio, prendiamo la scorciatoia dell'autoinganno. Se così fosse, questo rappresenterebbe un vero pericolo per la sopravvivenza umana. Ma se l'autoinganno è spiegato da un modello causale, ogniqualvolta un nostro desiderio sbatte contro il muro della realtà, il processo dell'autoinganno dovrebbe attivarsi automaticamente. Invece di regola affrontiamo la realtà che non ci piace e solo talvolta ci lasciamo guidare dai desideri frustrati nella considerazione distorta dell'evidenza disponibile. Un modello non intenzionale, a prima vista, sembra incapace di rendere conto della selettività dell'autoinganno. Come Pedrini nota, io ho cercato di affrontare questo problema aggiungendo al resoconto dell'autoinganno, secondo la mano invisibile, un'analisi delle circostanze in cui è altamente probabile che il processo di autoinganno si inneschi. Tra esse vorrei ricordare due elementi importanti: 1. L'importanza del desiderio che *p* nella funzione di utilità del soggetto; 2. La possibilità per il soggetto di ignorare i costi dell'inaccuratezza epistemica. Studi sperimentali hanno mostrato, per esempio, che quando il soggetto è impotente di fronte all'evidenza negativa, e pertanto non ha nulla da perdere nell'intrattenere una credenza falsa, tende a credere in conformità ai propri desideri (Jervis, 1976). Analogamente, se il soggetto pensa che i costi dell'inaccuratezza non ricadranno su di lui, per esempio perché protetto dalla "deniability clause", come accadde a Kennedy nel caso della Baia dei Porci, o perché non si percepisce come responsabile delle convinzioni e decisioni conseguenti, ecco che il processo di autoinganno trova una via privilegiata. Le circostanze favorevoli, tuttavia, rappresentano solo condizioni necessarie e non sufficienti per mettere in moto

l'autoinganno. Quindi è possibile che, a dispetto di queste circostanze favorevoli, qualche soggetto resista alla scorciatoia e la spiegazione di questo fatto potrà magari venire dalla psicologia cognitiva e dalle neuroscienze. Tuttavia, resto convinta di aver risolto il problema posto da Talbott e da Bermudez, ossia il problema dell'automatismo connesso alla spiegazione causale. Il modello a mano invisibile che io propongo non è causale o lo è solo parzialmente, e il soggetto deve percepire insieme all'evidenza negativa anche i costi per l'inaccuratezza, il che comporta capacità agenziali.

Veniamo così all'altra osservazione che Pedrini fa al mio lavoro. Secondo lei, io interpreterei il modello motivazionista unilateralmente come "anti-agency", mentre, in realtà, i motivazionisti non escludono che nell'autoinganno sia coinvolto un agente. Su questo sono d'accordo e l'ho anche chiaramente affermato nel libro[2]; il problema però è che la produzione dell'autoinganno viene dai motivazionisti esaustivamente spiegata in modo causale, senza il ricorso a mosse dell'agente. Certo non si esclude che queste ci siano e che l'agente faccia delle cose, ma l sua attività non ha alcun peso nella catena causale che porta alla formazione o al mantenimento della credenza falsa che *p* contro l'evidenza. Indubbiamente l'agente considera l'evidenza, dunque mette in moto un ragionamento, ma sono le biases attivate causalmente dal desiderio che generano la credenza ingannevole, di cui l'agente è sostanzialmente vittima impotente.

A questo punto, l'invito a considerare gli apporti della teoria cognitiva, come suggerisce Meini risulta quanto mai appropriato. Ho cercato di farlo nel libro, ma le tre teorie da lei menzionate completano la filosofia dell'autoinganno, mettendo in evidenza fenomeni correlati analizzati con metodi sperimentali. In particolare le prime due da lei citate, la *relevance theory* e la teoria del *working self,* forniscono una conferma su base sperimentale, dell'impatto della motivazione sulla cognizione, sia per quanto riguarda l'elaborazione di informazioni, sia per quanto riguarda la memorizzazione e il richiamo dei ricordi. Questo sostegno è importante perché c'è invece un filone delle teorie cognitive che tende a negare l'interferenza motivazionale, affidando la spiegazione di ciò che appare come irrazionalità motivata alle sole bias e a meccanismi "freddi" (Gilovich, 1991). Perché, si chiede Meini, i ricordi coerenti col sistema doxastico della persona sono favoriti rispetto a quelli che corrispondono al vero? La risposta potrebbe venire dalla psicologia clinica, da cui emergerebbe che tale preferenza corrisponde a tratti deliranti della personalità. A questo proposito, vorrei però sottolineare che l'autoinganno, a differenza della *self-delusion*, è sì un processo distorsivo della cognizione, tuttavia non è considerato un fenomeno patologico, proprio perché colpisce la quasi totalità degli individui in particolari circostanze. Individui, che per il resto, esibiscono normali capacità e abilità cognitive, e che, in genere, emergono dal loro autoinganno senza bisogno di trattamento clinico. Particolarmente interessante è infine la teoria argomentativa del ragionamento, avanzata da Mercier e Sperber (2011, 2017), che sostiene che il ragionamento è emerso evolutivamente non in funzione cognitiva, ma in funzione comunicativa: abbiamo imparato a ragionare per convincere gli altri delle nostre ragioni e, pertanto, la struttura del nostro ragionare è argomentativa, volta alla persuasione degli interlocutori. Sulla base di solidi risultati sperimentali, Mercier e Sperber sono così in grado di mostrare che in gruppo ragioniamo meglio, perché siamo in grado di mettere a confronto le diverse ipotesi, di vagliarne la giustificazione, scartare quelle sbagliate e ritenere

quelle migliori. A questo punto, la loro teoria sembrerebbe contraddire una tesi da me sostenuta, ossia che l'autoinganno avviene anche in gruppo e che anzi entro certi gruppi trova un terreno fecondo per prodursi. Tesi, per la verità, già ampiamente illustrata da Janis (1982) relativamente al fenomeno del *groupthink*, contro cui Sperber e Mercier polemizzano. Tuttavia, l'effetto positivo del ragionamento collettivo rispetto alla soluzione di problemi e alla produzione di conoscenza si manifesta solo in certe circostanze che Sperber non indaga nel dettaglio, ma che sintetizza nella motivazione condivisa dei partecipanti al gruppo verso la verità. Gli esperimenti che illustrano il fenomeno del successo del ragionamento di gruppo sono in effetti disegnati in modo che i partecipanti non solo siano motivati esclusivamente a trovare la soluzione del problema posto, del genere di quelli proposti nel Cognitive Reflection Test (Chaiken & Trope, 1999), ma sono anche in una posizione reciprocamente paritaria. Questi due aspetti rappresentano in realtà proprio due fra le condizioni ideali identificate dalla teoria deliberativa della democrazia per giungere a decisioni migliori (Habermas, 1996). Sperber e Mercier notano poi che se il contesto è invece quello di un dibattito avversariale, la motivazione non è più quella di risolvere il problema, ma di difendere il proprio punto di vista e attaccare quello dell'altro. In questo contesto, si manifesta un ragionamento motivato affetto da *confirmity bias* quando si tratta di sostenere le proprie tesi e da *disconfirmity bias* quando si tratta di attaccare le tesi dell'avversario, che amplifica la polarizzazione delle posizioni. Le abilità nel ragionamento qui si dispiegano come produzione di giustificazioni per le credenze pregresse nei soggetti. La cosa interessante è che in questo contesto non vince l'argomento migliore, ma quello sostenuto più abilmente da una delle due parti. Questo dato mette in questione la tesi secondo cui la motivazione interviene nel sistema intuitivo e automatico della mente, il cosiddetto S1, generando esiti affetti da bias e scorretti. Al contrario il sistema riflessivo del pensiero analitico, il più lento ma più affidabile S2, sarebbe esente da rischi di errori e da autoinganno (Chaiken & Trope, 1999). In realtà la teoria di Sperber, sostenuta anche da ricerche sperimentali di Kahan (2013), afferma che il ragionamento motivato si dispiega all'interno di S2 e richiede notevole capacità di ragionamento per trovare giustificazioni convincenti alle proprie opinioni e per attaccare le opinioni altrui. Ciò confermerebbe la tesi che io e Pedrini, tra gli altri, sosteniamo, secondo la quale l'autoinganno richiede una buona sofisticazione argomentativa, senz'altro viziata da bias – e quindi substandard se misurata sulla razionalità epistemica – ma in ogni caso capace di trovare giustificazioni complesse per le proprie credenze[3]. Questa elaborazione rappresenta un tratto distintivo dell'autoinganno rispetto per esempio al *wishful thinking* dove un desiderio induce direttamente a credere che p sia vero, senza la ruminazione e la produzione di controargomenti per neutralizzare l'evidenza negativa che sono proprie dell'autoinganno. Alla fine, la teoria del ragionamento argomentativo porta acqua alla tesi dell'autoinganno come processo controevidenziale a sostegno di credenze distorte dai propri desideri.

Note

[1] Ho discusso del modello intenzionale e delle modalità di uscita dai paradossi nel cap. I di *Political Self-Deception*, pp. 25-31.

[2] Cfr. *Political Self-Deception*, p. 38.

[3] Questa tesi è sostenuta anche da Michael & Newen (2010) sulla base della ricerca sperimentale di Wentura & Greve (2005).

Riferimenti bibliografici

Audi, R. (1982), "Self-Deception, Action and the Will", *Erkenntnis*, 18, pp. 133-158.

Audi, R. (1989), "Self-Deception and Practical Reasoning", *Canadian Journal of Philosophy*, 19, 2, pp. 247-266.

Bermudez, J.L. (2000), "Self-Deception, Intention and Contradictory Beliefs", *Analysis*, 60, 4, pp. 309-319.

Chaiken, S., Trobe, Y. (1999), *Dual Process Theories in Social Psychology*, New York: Guilford.

Davidson, D. (1982), "Paradoxes of Irrationality", in R. Wollheim, J. Hopkins, *Philosophical Essays on Freud*, Cambridge: Cambridge U.P., pp. 289-305.

Davidson, D. (1985), "Deception and Division", in E. LePore, B. McLaughlin, *Actions and Events: Perspectives on the Philosophy of Donald Davidson*, Oxford: Blackwell, pp. 138-148.

Elster, J. (1983), *Sour Grapes. Essay on Rationality and Irrationality*, Cambridge: Cambridge U.P.

Fingerette, H. (1969), *Self-Deception*, London: Routledge & Kegan Paul.

Fingerette, H. (1998), "Self-Deception Needs No Explaining", *Philosophical Quarterly*, 48, pp. 289-301.

Foss, J. (1980), "Rethinking Self-Deception", *American Philosophical Quarterly*, 17, pp. 237-243.

Gardner, S. (1993), *Irrationality and the Philosophy of Psychoanalysis*, Cambridge: Cambridge U.P.

Gilovich, T. (1991), *How Do We Know What Isn't So?*, New York: The Free Press.

Janis, I. (1982), *Groupthink: Psychological Studies of Policy Decisions and Fiascoes*, Boston: Houghton Mifflin.

Jervis, R. (1976), *Perception and Misperception in International Politics*, Princeton: Princeton U.P.

Mele, A. (1997), "Real Self-Deception", *Behavioral and Brain Sciences*, 20, pp. 91-102.

Mele, A. (2001), *Self-Deception Unmasked*, Princeton: Princeton U.P.

Michael, C., Newen, A. (2010), "Self-Deception as Pseudo-Rational Regulation of Beliefs", *Consciousness and Cognition*, 19, pp. 731-744.

Pears, D. (1984), *Motivated Irrationality*, Oxford: Oxford U.P.

Pedrini, P. (2013), *L'autoinganno. Che cos'è e come funziona*, Bari-Roma: Laterza.

Rorty, A.O. (1972), "Belief and Self-Deception", *Inquiry*, 15, pp. 387-410.

Talbott, W.J. (1995), "Intentional Self-Deception in a Single, Coherent Self", *Philosophy and Phenomenological Research*, 55, pp. 27-74.

Wentura, D., Greve, W. (2003), "Who Wants to Be Erudite? Everyone! Evidence for Automatic Adaptations of Trait Definition", *Social Cognition*, 22, pp. 30-53.