I want to believe: delusion, motivated reasoning, and Bayesian decision theory

Francesco Rigoli[1,*], Cristina Martinelli,[2,3] and Giovanni Pezzulo[4]

[1] *City, University of London, Northampton Square, London, EC1V 0HB, UK*

[2]*Institute of Psychiatry, Psychology & Neuroscience, King's College London, 16 De Crespigny Park Road, London, SE5 8AF, UK*

[3]*Kingston University, Penrhyn Road, Kingston Upon Thames, Surrey, KT1 2EE, UK*

[4]*Insitute of Cognitive Sciences and Technologies, National Research Council of Italy, Rome (Italy)*

***Correspondence**: Francesco Rigoli

Department of Psychology

City, University of London

Northampton Square, London, UK EC1V 0HB

francesco.rigoli@city.ac.uk

Abstract

Introduction. Several arguments suggest that motivated reasoning (occurring when beliefs are not solely shaped by accuracy, but also by other motives such as promoting self-esteem or self-protection) is important in delusions. However, classical theories of reasoning in delusion disregard the role of motivated reasoning. Thus, this role remains poorly understood.

Methods. To explore the role of motivated reasoning in delusions, here we propose a computational model of delusion based on a Bayesian decision framework. This proposes that beliefs are not only evaluated based on their accuracy (as in classical reasoning theories), but also based on the cost (in terms of reward and punishment) of rejecting them.

Results. The model proposes that, when the values at stake are high (as often it is the case in the context of delusion), a belief might be endorsed because rejecting it is evaluated as too costly, even if the belief is less accurate. This process might contribute to the genesis of delusions.

Conclusions. Our account offers an interpretation of how motivated reasoning might shape delusions. This can inspire research on the affective and motivational processes supporting delusions in clinical conditions such as in psychosis, neurological disorders, and delusional disorder.

Keywords: delusion; Bayesian decision theory; motivated reasoning; affect; emotion; stress

Introduction

Delusions (i.e., false beliefs retained with strong conviction despite contrary evidence) are at the core of several psychiatric and neurological disorders (Coltheart et al., 2011). Classical reasoning accounts such as Two-factor theory (Coltheart, 2010; Coltheart et al, 2011; Davies et al., 2001; 2005; Langdon & Coltheart, 2000) have offered valuable insight on delusion. According to these, delusions result from maladaptive reasoning processes based on two elements, comprising (i) an abnormal perception and (ii) an aberrant interpretation of such perception. As an example, consider the Capgras syndrome (Alexander et al., 1979), a condition characterised by believing that a family member is not real and has been replaced by an impostor (Alexander et al., 1979). Classical reasoning models propose that this syndrome arises from an abnormal perception, corresponding to a patient's failure to experience any emotional reaction when encountering the family member. Moreover, an abnormal interpretation of this perception would also be a factor: patients would explain their absence of emotional response as deriving from the family member being replaced by an impostor.

Reasoning accounts of delusion have offered tremendous insight. Yet, they have rarely explored affective processes, though these have emerged as fundamental facets of delusion (Bentall et al., 2009; Green et al., 2006; Martinelli et al, 2013; Turnball & Bebbington, 2001). Empirical research has observed heightened anxiety levels in many instances of delusion (Bentall et al., 2009; Green et al., 2006; Turnball & Bebbington, 2001). Moreover, in healthy individuals, experimentally induced anxiety has been observed to promote delusional beliefs (Martinelli et al., 2013), and affective disorders predispose individuals towards delusion (Kempf et al., 2005). Based on these observations, it has been argued that delusions do not simply arise from abnormal perceptions and interpretations, but also from aberrant affective processes (Bentall et al., 2009; Green et al., 2006; Martinelli et al, 2013; Turnball &

Bebbington, 2001). An intriguing possibility is that *motivated reasoning* (occurring when reasoning is not shaped solely by accuracy, but also by other motives such as promoting self-esteem or self-protection; Kunda, 1990) is the process through which, at least partially, affective processes influence delusion (Bentall et al., 1994; Freeman et al., 2002). The existence of motivated reasoning in the normal population is well documented, for instance by data showing that people often accept political beliefs that support their self-interest even if these beliefs are less accurate (Redlawsk, 2002). This raises the possibility that motivated reasoning might also contribute to the formation of delusional beliefs (Bentall et al., 1994; Freeman et al., 2002). Delusions often arise from domains where one believes the consequences at stake are vital (Bentall et al., 1994; Freeman et al., 2002). For instance, a patient may contemplate the possibility of being chased by an alien who intend to kill the patient. Failing to acknowledge this threat, and failing to take the appropriate measures, would be perceived as a fatal mistake. In contexts like these, patients would not much strive to infer the most accurate interpretation of events, but rather the interpretation that best preserves motives such as self-protection, resulting in motivated reasoning. Therefore, a patient might staunchly endorse the delusional belief despite scarce evidence in support of it; rather, the delusional belief would be endorsed because holding it would allow the patient to be prepared if the belief (although unlikely) turns out to be correct.

How motivated reasoning works in the context of delusions remains to be explored within a computational perspective. Here we carry out this exploration, introducing a computational model about the impact of motivated reasoning upon delusion.

The model

Classical reasoning models of delusion (Coltheart, 2010; Coltheart et al, 2011; Davies et al., 2001; 2005; Langdon & Coltheart, 2000) presuppose that beliefs arise from an implicit motivation to be accurate. Because of their exclusive focus on accuracy, these models of delusion are not suited to account for motivated reasoning, where motives besides accuracy seeking are at play. However, we argue that motivated reasoning can be captured by adopting a Bayesian framework, which is another highly influential perspective about delusions (Adams et al., 2013; Corlett et al., 2010; 2011; Fletcher & Frith, 2009; Frith & Friston, 2013). To date, it remains to be explored whether a Bayesian approach can offer insight on the role of emotional processes in delusion, which might underly motivated reasoning; here we explore this possibility. The model we propose corresponds to a standard Bayesian decision framework implemented adopting the formalism of Bayesian networks (Bishop, 2006; Rigoli, 2021a; 2021b). The network is represented graphically in Fig. 1 (a more formal description is offered in the Appendix). It describes the beliefs an individual entertains about important variables and about their relationships. The variables are represented by rectangles (for categorical variables) and circles (for continuous variable), with arrows indicating probabilistic dependencies among variables. The first variable in the model is Hypothesis (Hyp), representing a categorical variable reflecting a set of alternative hypotheses about an important aspect of reality. For example, someone may attempt to infer the intensions of an individual who is ringing the bell. One hypothesis (a delivery hypothesis) considered by Hyp is that the individual is delivering a box. An alternative hypothesis (a robbery hypothesis) is that the individual is robbing. The second variable in the model is Prior Belief System (PBS). This represents a categorical variable reflecting a set of alternative general views of the world which depend on past experience. For example, one view might be that most people are benevolent and the alternative view that they are hostile. The variable Hyp depends on PBS,

as the arrow going from the latter to the former indicates. For example, someone who tends to view people as hostile will tend to attribute higher probability to the robbery hypothesis.

Both Hyp and PBS are *hidden* (or *latent*) variables, as they cannot be observed directly but need to be inferred indirectly. For example, one does not directly know whether most people are hostile or benevolent (PBS), nor whether the person's intention is to deliver a box or to rob (Hyp). Conversely, the variable Evidence (Evi), capturing any novel sensory or social information, is directly observable. For example, this might correspond to the physical appearance of the person ringing the bell, a feature viewed as useful for inferring the person's intentions. This variable is believed to be the consequence of Hyp, as indicated by the arrow going from Hyp to Evi. This probabilistic relation implies that observing Evi helps estimating the value of the two hidden variables (Hyp and PBS), as explained below.

Finally, the model includes a Hypothesis Decision (HDec) and an Expected Outcome (EOut) variable. HDec is categorical and indicates which hypothesis is accepted as true and is used to guide behaviour. For example, HDec may include the following categories: (i) accept the robbery hypothesis (and do not open the door) and (ii) accept the delivery hypothesis (and open the door). EOut reflects the expected outcome of this decision and depends both on Hyp and HDec. EOut is represented by a continuous variable where negative values correspond to punishment and positive values to reward. For example, EOut describes the outcome expected to occur (i) if the robbery hypothesis is true and I accept it (and I do not open the door), (ii) if the delivery hypothesis is true and I accept it (and I open the door), (iii) if the robbery hypothesis is false but I accept it (and I do not open the door), (iv) if the delivery hypothesis is false but I accept it (and I open the door).

The model realizes Bayesian decision by deciding which hypothesis to accept. Specifically, the model infers the consequences (in terms of reward and punishment) of accepting different

hypotheses considering novel evidence Evi. Eventually, the hypothesis associated with the best consequences is accepted. More formally, this decision process follows multiple steps. At each step, Evi is observed and one different category of HDec is considered as observed, too. On this basis, EOut given Evi and HDec (i.e., P(EOut|Evi,HDec)) is calculated. This calculation is repeated for all possible categories of HDec. Next, a decision follows whereby the category of HDec associated with the best EOut (i.e., the highest posterior outcome value) is chosen.

Crucially, in this framework the accepted hypothesis is not the one enjoying more support based on prior beliefs and novel evidence (i.e., the one that maximizes accuracy), but the one associated with the best consequences (i.e., the one that maximizes utility). The emphasis on utility is important; as we shall see below, it allows the model to implement affective processes which are not contemplated by classical reasoning models of delusion (Coltheart, 2010; Coltheart et al, 2011; Davies et al., 2001; 2005; Langdon & Coltheart, 2000).

According to our model, what is the phenomenological implication of accepting one hypothesis over the other? We propose that, phenomenologically, an agent will believe that the accepted hypothesis is true even if, as explained above, it is not the most accurate. In other words, agents are supposed to be blind to the decision process described above; they would simply perceive the accepted hypothesis as being true, without awareness that their perception is the product of reward maximization. This allows motivated reasoning to emerge (Kunda, 1990). Note that, however, accuracy is still fundamental. This is because accepting a hypothesis which is poorly supported by prior beliefs (PBS) and evidence (Evi) is scarcely rewarding, implying that such hypothesis will be discarded.

In short, our model explains the genesis of beliefs by relying on a Bayesian decision framework, allowing us to contemplate motivated reasoning. Below, we will examine how delusions can be interpreted within this framework.

The model applied to delusions

In our model, prior beliefs (captured by PBS) and novel evidence (captured by Evi) play a critical role. Consider the example above where Hyp includes a delivery and a robbery hypothesis, and assume that the robbery hypothesis is delusional. Someone believing that most people are hostile (PBS) and observing a suspicious look characterising the person ringing the bell (Evi) will tend to favour the delusional hypothesis (fig. 2).

However, given their exclusive role in accuracy maximization, prior beliefs and novel evidence appear as inadequate to explain motivated reasoning and its impact on delusions. Delusions often arise from domains where one believes the consequences at stake are vital (Bentall et al., 1994; Freeman et al., 2002). The patient may believe that accepting or rejecting a hypothesis is fundamental for survival. For instance, a patient may contemplate the possibility of being chased by a cruel alien. Failing to acknowledge this threat, and failing to take the appropriate measures, would be perceived as a fatal mistake. In contexts like these, motivated reasoning would favour interpretations of events which, although less accurate, best preserve motives such as self-protection. How would motivated reasoning unfold in such scenarios? While classical reasoning accounts (Coltheart, 2010; Coltheart et al, 2011; Davies et al., 2001; 2005; Langdon & Coltheart, 2000) struggle to answer this question, our model offers a possible interpretation. Consider the example above requiring to arbitrate between a delivery and a robbery hypothesis, with the latter being a delusional belief. Intuitively, to decide which hypothesis to accept, our model considers two aspects. First, it

assesses which hypothesis fits best with evidence (e.g., by asking: “does the person look like a thief?”). Second, it considers the risks at stake by asking: what is the cost if I accept the robbery hypothesis (and do not open the door) and this hypothesis is true? And if it is false? What is the cost if I accept the delivery hypothesis (and do open the door) and this hypothesis is true? And if it is false? Based on considering both aspects (evidence and risks), either hypothesis will be accepted. Clearly, this decision requires considering not only the accuracy of each hypothesis (i.e., to what extent it fits with evidence), but also the consequences of its acceptance/rejection. For example, if I perceive robbery as a catastrophic outcome (thus perceiving the risks associated with rejecting this hypothesis as unbearable), I will accept the robbery hypothesis even if (based on evidence) I consider it as being unlikely. Such decision would act unconsciously: only its outcome would emerge consciously, expressed in the subjective perception of the delusional hypothesis being true (tab. 1; fig. 3). This example illustrates how, in our model, motivated reasoning might emerge and contribute to the formation of delusions.

This proposal can be compared with previous attempts to understand the role of motivated reasoning in the genesis of delusion (Bentall et al., 1994; Freeman et al., 2002). It has been argued that often a deluded patient is facing a dilemma between a delusional hypothesis (e.g., “I am chased by aliens”) and an alternative hypothesis implicating that the patient is mentally ill (e.g., “aliens are the product of my mental illness”) (Freeman, 2007; Freeman et al., 2002; Freeman & Garety, 2004; 2014). Despite both hypotheses being gloomy, the delusional hypothesis would appear as less gloomy to the patient, and thus it would be accepted. In other words, according to this view (Freeman, 2007; Freeman et al., 2002; Freeman & Garety, 2004; 2014), the delusional hypothesis would be accepted because the alternative, implicating mental illness, is perceived as being less appealing. Our model implies a radically different picture. It implies that the delusional hypothesis is accepted because of perceiving a higher

cost for rejecting it. Contrary to previous arguments (Bentall et al., 1994; Freeman et al., 2002), this predicts that a patient fears more the delusional hypothesis (e.g., being killed by aliens) than the alternative hypothesis (e.g., being mentally ill). We argue that, for instance, our perspective offers a better explanation of jealousy delusions: here it is hard to argue that, like previous proposals imply (Bentall et al., 1994; Freeman et al., 2002), the patient views the delusional hypothesis (e.g., being cheated by the spouse) as more appealing than the alternative hypothesis (e.g., not being cheated by the spouse). Rather, in line with our model, in jealously delusions the cost of rejecting the delusional hypothesis is arguably perceived as higher (e.g., failing to punish the wife for cheating would be perceived as highly shameful).

Our model can also explain the role of stress in the genesis of delusions. Evidence has shown that delusions are more likely to develop after experiencing stress (Freeman et al., 2001). Often, after a severe stressful episode (e.g., after being fired at work), an individual develops serious psychotic symptoms including delusional convictions (e.g., the belief that aliens want to kill the patient) associated with extreme anxiety. These convictions often develop although the stressful episode, which worked as a trigger in the first place, is apparently unrelated with their content. Within our model, even when they do not represent direct evidence supporting delusional beliefs, stressful events would often impact upon the utility associated with the different hypotheses under consideration. A consequence of this might be that, after a stressful event occurs, rejecting a delusional belief might suddenly become associated with extremely high cost, in turn leading to an abrupt endorsement of the delusional belief. This process might explain why, after stressful events, delusions tend to emerge despite no exposure to novel information supporting them.

In short, our model proposes a key role for motivated reasoning in the formation of delusions. Specifically, we propose that, even when poorly supported by evidence, a delusion may arise

because its rejection is evaluated as too dangerous. This can explain empirical evidence indicating that affective factors are critical in the development and maintenance of delusions.

Discussion

Classical accounts have compellingly argued that impaired reasoning is often at the root of delusions (Coltheart, 2010; Coltheart et al, 2011; Davies et al., 2001; 2005; Langdon & Coltheart, 2000). However, reasoning accounts straggle to explain the role of affective processes in the genesis of delusions. These often appear motivated less by accuracy than by other motives such as self-protection: someone convinced to be chased by aliens is likely to hold this belief not much because it appears as realistic, but rather because its rejection is viewed as too risky. Building on Bayesian interpretations of delusions (Adams et al., 2013; Corlett et al., 2010; 2011; Fletcher & Frith, 2009; Frith & Friston, 2013), here we propose a computational model where these forms of motivated reasoning are explained.

The model can contribute to understand a variety of clinical conditions where delusions are at play. These conditions include, among others, the Capgras syndrome, the Fregoli syndrome, somatoparaphrenia, anosognosia, psychosis, and delusional disorder (Coltheart et al., 2011). Specifically, the model can inspire research on how affective processes and motivated reasoning shape delusions as manifested in these disorders. This is particularly relevant in conditions where affective processes are central, such as in psychosis and delusional disorder (Bentall et al., 2009; Green et al., 2006; Martinelli et al, 2013; Munro, 1999; Turnball & Bebbington, 2001). In other conditions, such as those derived from clear organic neural damages (e.g., somatoparaphrenia), affective processes appear to be more peripheral. Yet, albeit peripheral, affective processes might still have substantial impact here too, and our model might contribute to elucidate these processes. Our model encourages research to

explore affective processes and motivated reasoning in conditions where neural damage is at the root of delusions, an aspect poorly studied so far. An intriguing possibility inspired by our proposal is that, in these conditions, utility considerations associated with the delusional belief might be influential (i.e., that the delusion might not arise solely from accuracy seeking, but also from motivated reasoning). For instance, in the Capgras syndrome, the belief that an impostor has replaced a family member might be hold, at least partially, to avoid the risk of ignoring the danger of living with a malevolent impostor. This and similar hypotheses remain to be explored empirically.

As an example of how our model can help interpreting clinical conditions, we examine delusional disorder in some detail (Munro, 1999). This is defined by delusional beliefs hold in the absence of neurological impairments, psychotic disorders, or mood disorders. Dysfunctions are usually circumscribed to the delusional beliefs, with other psychological functions being preserved. Except when delusional beliefs are activated, the patient often presents a relatively unproblematic everyday life. The theme of the delusion varies, although the diagnosis is appropriate only if the content is not bizarre (otherwise, a diagnosis of psychosis is appropriate). The delusion concerns aspects key for the patient's safety, identity and self-esteem. Based on the delusion's themes, the following subtypes have been identified: erotomaniac, grandiose, jealous, persecutory, and somatic (Munro, 1999).

Contrary to neurological conditions or psychosis, there is no evidence of any abnormal perception in delusional disorder (Munro, 1999). Therefore, processes based on dysfunctional interpretations of abnormal perceptions, as advocated by reasoning models, are unlikely to be pivotal here. This raises the possibility that motivated reasoning, the focus of our model, might be the critical factor.

Our model proposes the following process for the development of delusional disorder. Before any delusion arises, substantial prior probability would already be attributed to the delusional hypothesis. However, initially this hypothesis would not be endorsed yet. For example, a husband might doubt that he has been cheated by his wife, but this belief might not be given too much credit. A period of heavy distress would then occur. The distress might be totally unrelated with the theme of the delusion: for example, the husband might lose the job, an event totally irrelevant to establish whether the wife has cheated him or not. However, as argued above, distress might impact on the utility associated with the different hypotheses under consideration, even if these hypotheses are not linked with stress in any direct way; for example, distress might exacerbate the suffering associated with the hypothesis of being cheated. In turn, this change in the expected outcome would trigger the emergence of the disorder: now, rejecting the hypothesis of being cheated might appear as extremely risky (e.g., in terms of self-honour) if it turns out that the cheating has actually occurred. Thus, suddenly the husband might become totally convinced of being cheated by the wife, and act accordingly.

It is important to highlight that, in our account, expected outcome is a subjective estimate, and hence potentially unrealistic. People suffering from delusion might manifest a remarkable discrepancy between the outcomes they expect and the outcomes they actually collect. Consider the example of jealousy delusion. As discussed above, a possibility is that, in this form of delusion, dramatic costs are expected by rejecting the hypothesis of being cheated by a partner. However, in reality, much greater costs might be eventually experienced by embracing the delusional hypothesis, for example as a consequence of increased conflict or of social isolation. Even more paradoxically, delusional beliefs might at times produce self-fulfilling prophecies, thus creating the suffering they strive to avoid. For example, jealousy delusions might sometimes disrupt the relationship with a partner up to a point when the

partner ends up looking for other lovers. Based on these considerations, exploring to what degree expectations about outcomes are unrealistic in people suffering from delusion represents an intriguing research avenue.

The role of expected outcome as postulated here has analogies with Error Management Theory (Haselton & Buss, 2000). This posits that, along its evolutionary history, the human brain has developed cognitive biases grounded upon the distinction between Type I errors (occurring when a false hypothesis is accepted) and Type II errors (occurring when a true hypothesis is rejected). For example, evidence that men, but not women, overestimate the sexual attraction expressed by people from the other sex towards them is explained as arising because, in an evolutionary perspective, the cost of Type I error (assessing potential partners as sexually available although in fact they are not) is lower in men compared to women. Despite the similarity between this logic and the model of delusion developed here, an important distinction concerns the processes proposed to be at play. While, in Error Management Theory, the costs of Type I and Type II errors are at play "implicitly" via evolutionary mechanisms (e.g., favouring the survival of males who overestimate sexual availability), in our proposal the costs of Type I and Type II errors are calculated online by the brain. In other words, in our account (but not in Error Management Theory) the brain weights the costs and benefits of accepting any hypothesis and derives its beliefs accordingly.

In summary, here we propose a computational model of delusion based on Bayesian decision theory. This extends classical reasoning models (Coltheart, 2010; Coltheart et al, 2011; Davies et al., 2001; 2005; Langdon & Coltheart, 2000) by considering the impact of motivated reasoning upon delusions. More generally, by analysing how Bayesian principles explain the role of motivated reasoning in delusional beliefs, our proposal contributes to advancing the Bayesian perspective on the study of delusion (Adams et al., 2013; Corlett et al., 2010; 2011; Erdmann & Mathys, 2021; Fletcher & Frith, 2009; Frith & Friston, 2013).

Appendix

Formally, the model is a mixture of Gaussians. The joint probability can be written as:

$$P(PBS, Hyp, HDec, EOut, Evi) = P(PBS)\, P(HDec)\, P(Hyp \mid PBS)\, P(Evi \mid Hyp)\, P(EOut \mid Hyp, HDec)$$

PBS is a categorical variable with number of categories equal to n_{PBS} and where each category is associated with a probability. If we consider the example of an individual ringing the bell (see above), we can set $n_{PBS} = 2$, *PBS = Hos* if the individual ringing the bell is hostile, and *PBS = Ben* if the individual is benevolent. The probability of the individual being hostile is *P(PBS = Hos) = x* and the probability of the individual being benevolent is *P(PBS = Ben) = 1 - x* (where $0 \leq x \leq 1$). Hyp is also categorical, with number of categories equal to n_{Hyp}. Considering the same example, we can set $n_{Hyp} = 2$, *Hyp = Del* if the individual is delivering a box, and *Hyp = Rob* if the individual is robbing. The conditional probabilities for Hyp are *P(Hyp = Del | PBS = Hos) = y*, *P(Hyp = Rob | PBS = Hos) = 1 – y*, *P(Hyp = Del | PBS = Ben) = z*, *P(Hyp = Rob | PBS = Ben) = 1 – z* (where $0 \leq y \leq 1$ and $0 \leq z \leq 1$). HDec is also categorical, with the number of categories being $n_{HDec} = n_{Hyp}$. In our example, *HDec = RobAcc* when the robbery hypothesis is accepted (or, equivalently, when the delivery hypothesis is rejected) and *HDec = DelAcc* when the robbery hypothesis is rejected (or, equivalently, when the delivery hypothesis is accepted). Probabilities for HDec are *P(HDec = RobAcc) = u* and *P(HDec = DelAcc) = 1 - u* (where $0 \leq u \leq 1$).

Evi is represented by a real number and follows a Gaussian distribution, with negative numbers supporting the robbery hypothesis and positive numbers supporting the delivery hypothesis. Evi is conditioned upon Hyp, with conditional probability defined as:

$$P(Evi \mid Hyp = k) = \mathcal{N}(\mu_{Evi|k}, 1/{\lambda_{Evi}}^2)$$

Here, every category of Hyp k has its own associated average $\mu_{Evi|k}$; for instance, the model will include $\mu_{Evi|Rob}$ (conditional on the robbery hypothesis being true) which is different from $\mu_{Evi|Del}$ (conditional on the delivery hypothesis being true). The parameter ${\lambda_{Evi}}^2$ reflects the weight or precision of Evi and in our model is equal for all levels of Hyp (in principle, a specific weight for each level of Hyp can be implemented).

Finally, EOut is a Gaussian variable conditioned on both Hyp and HDec. Its conditional probability is:

$$\text{P(EOut I Hyp = k, HDec = j)} = \mathcal{N}(\mu_{EOut|k,j}, \sigma^2_{EOut})$$

This indicates a specific average for each combination of Hyp and HDec. For instance, the model comprises $\mu_{EOut|Rob,RobAcc}$ (the expected outcome if the robbery hypothesis is true and it is correctly accepted), $\mu_{EOut|Rob,DelAcc}$ (the expected outcome if the robbery hypothesis is true but it is wrongly rejected), $\mu_{EOut|Del,DelAcc}$ (the expected outcome if the delivery hypothesis is true and it is correctly accepted), $\mu_{EOut|Del,RobAcc}$ (the expected outcome if the delivery hypothesis is true but it is wrongly rejected). The parameter σ^2_{EOut} reflects the uncertainty about the outcome and in our model it is equal for all combinations of Hyp and HDec (although in principle one can also implement a specific weight for each combination).

The model is used to make inference. For inference, the variables Evi is observed, while the other variables are not. The inference process included multiple steps. At each step, for each level of HDec *j*, the model infers the conditional probability of EOut given the observed values for Evi and given HDec = j. This corresponds to the posterior Gaussian distribution:

$$\text{P(EOut| Evi, HDec = j)} = \mathcal{N}(\mu_{EOut|Evi,j}, \sigma^2_{POST})$$

Where $\mu_{EOut|Evi,j}$ is the posterior average for the expected outcome. For example, $\mu_{EOut|Evi,RobAcc}$ is the posterior average if the robbery hypothesis is accepted and $\mu_{EOut|Evi,DelAcc}$ is the posterior average if the delivery hypothesis is accepted.

After these inference steps are completed (i.e., after the posterior outcome is calculated for all values of HDec), the model makes a decision by choosing the hypothesis associated with the highest posterior $\mu_{EOut|Evi,j}$. For instance, it will either choose to accept or reject the robbery hypothesis (or, equivalently, to reject or accept the delivery hypothesis, respectively).

References

Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D., & Friston, K. J. (2013). The computational anatomy of psychosis. *Frontiers in psychiatry*, *4*, 47.

Alexander, M. P., Stuss, D. T., & Benson, D. F. (1979). Capgras syndrome: a reduplicative phenomenon. *Neurology*, *29*(3), 334-334.

Bentall, R. P., Kinderman, P., & Kaney, S. (1994). The self, attributional processes and abnormal beliefs: towards a model of persecutory delusions. *Behaviour research and therapy*, *32*(3), 331-341.

Bentall, R. P., Rowse, G., Shryane, N., Kinderman, P., Howard, R., Blackwood, N., ... & Corcoran, R. (2009). The cognitive and affective structure of paranoid delusions: a transdiagnostic investigation of patients with schizophrenia spectrum disorders and depression. *Archives of general psychiatry*, *66*(3), 236-247.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Coltheart, M. (2010). The neuropsychology of delusions. *Annals of the New York Academy of Sciences*, *1191*(1), 16-26.

Coltheart, M., Langdon, R., & McKay, R. (2011). Delusional belief. *Annual review of psychology*, *62*, 271-298.

Corlett, P. R., Taylor, J. R., Wang, X. J., Fletcher, P. C., & Krystal, J. H. (2010). Toward a neurobiology of delusions. *Progress in neurobiology*, *92*(3), 345-369.

Corlett, P. R., Honey, G. D., Krystal, J. H., & Fletcher, P. C. (2011). Glutamatergic model psychoses: prediction error, learning, and inference. *Neuropsychopharmacology*, *36*(1), 294.

Davies, M., Coltheart, M., Langdon, R., & Breen, N. (2001). Monothematic delusions: Towards a two-factor account. *Philosophy, Psychiatry, & Psychology*, *8*(2), 133-158.

Davies, M., Davies, A. A., & Coltheart, M. (2005). Anosognosia and the Two-factor Theory of Delusions. *Mind & Language*, *20*(2), 209-236.

Erdmann, T., & Mathys, C. (2021). A generative framework for the study of delusions. *Schizophrenia Research*.

Fletcher, P. C., & Frith, C. D. (2009). Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, *10*(1), 48.

Freeman, D., Garety, P. A., & Kuipers, E. (2001). Persecutory delusions: developing the understanding of belief maintenance and emotional distress. *Psychological medicine*, *31*(7), 1293-1306.

Freeman, D. (2007). Suspicious minds: the psychology of persecutory delusions. *Clinical psychology review*, *27*(4), 425-457.

Freeman, D., & Garety, P. A. (2004). *Paranoia: The psychology of persecutory delusions*. Psychology Press.

Freeman, D., & Garety, P. (2014). Advances in understanding and treating persecutory delusions: a review. *Social psychiatry and psychiatric epidemiology*, *49*(8), 1179-1189.

Freeman, D., Garety, P. A., Kuipers, E., Fowler, D., & Bebbington, P. E. (2002). A cognitive model of persecutory delusions. *British Journal of Clinical Psychology*, *41*(4), 331-347.

Frith, C. (2012). Explaining delusions of control: The comparator model 20 years on. *Consciousness and cognition*, *21*(1), 52-54.

Frith, C. D., & Friston, K. J. (2013). False perceptions and false beliefs: understanding schizophrenia. *Neurosciences and the Human Person: New Perspectives on Human Activities*, *121*, 1-15.

Green, C., Garety, P.A., Freeman, D., Fowler, D., Bebbington, P.,Dunn, G., & Kuipers, E. (2006). Content and affect in persecutory delusions

Kempf, L., Hussain, N., & Potash, J. B. (2005). Mood disorder with psychotic features, schizoaffective disorder, and schizophrenia with mood features: trouble at the borders. *International Review of Psychiatry*, *17*(1), 9-19.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological bulletin*, *108*(3), 480.

Langdon, R., & Coltheart, M. (2000). The cognitive neuropsychology of delusions. *Mind & Language*, *15*(1), 184-218.

Martinelli, C., Cavanagh, K., & Dudley, R. E. (2013). The impact of rumination on state paranoid ideation in a nonclinical sample. *Behavior therapy*, *44*(3), 385-394.

Munro, A. (1999). *Delusional disorder: paranoia and related illnesses*. Cambridge University Press.

Redlawsk, D. P. (2002). Hot cognition or cool consideration? Testing the effects of motivated reasoning on political decision making. *The Journal of Politics*, *64*(4), 1021-1044.

Rigoli, F. (2021a). A computational perspective on faith: religious reasoning and Bayesian decision. *Religion, Brain & Behavior*, 1-18.

Rigoli, F. (2021b). Masters of suspicion: A Bayesian decision model of motivated political reasoning. *Journal for the Theory of Social Behaviour*.

Turnbull, G., & Bebbington, P. (2001). Anxiety and the schizophrenic process: clinical and epidemiological evidence. *Social psychiatry and psychiatric epidemiology*, *36*(5), 235-243.

	EOut | Hyp, HDec	
	Person 1: Delusional belief	Person 2: Non-delusional belief
EOut | Rob, RobAcc	0	0
EOut | Del, DelAcc	0	0
EOut | Rob, DelAcc	**-100**	**-10**
EOut | Del, RobAcc	-10	-10

Tab. 1. Description of the role of expected outcome. The scenario is discussed also in the main text, where Hyp includes two categories (Robbery hypothesis vs Delivery hypothesis; with the former being delusional), PBS includes two categories (Hostile vs Benevolent), and negative values of Evi support the Robbery hypothesis. The table reports the expected outcome for accepting each hypothesis when the hypothesis is true or false (EOut | Rob, RobAcc: outcome expected if the robbery hypothesis is correctly accepted; EOut | Del, DelAcc: outcome expected if the delivery hypothesis is correctly accepted; EOut | Rob, DelAcc: outcome expected if the delivery hypothesis is wrongly accepted; EOut | Del, RobAcc: outcome expected if the robbery hypothesis is wrongly accepted). Two persons are compared, equal in terms of parameters and in terms of Evi available (here Evi is supporting slightly the Delivery (non-delusional) hypothesis for both individuals), but varying with respect to EOut | Rob, DelAcc. Specifically, Person 1 expects a much higher cost associated with EOut | Rob, DelAcc (-100 versus -10). Because of this, the model predicts that, contrary to Person 2, Person 1 will embrace the (delusional) Robbery hypothesis.

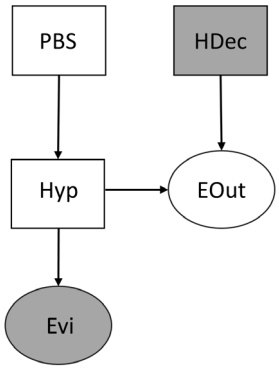

Fig 1. Bayesian network representing the model. Its variables are: Prior Belief Systems (PBS), Hypothesis (Hyp), Evidence (Evi), Hypothesis Decision (HDec), and Expected Outcome (EOut). Categorical and continuous variables are represented by rectangles and circles, respectively. Arrows indicate probabilistic causal relations from one variable to another. Shaded variables are those considered to be observed at each step of inference.

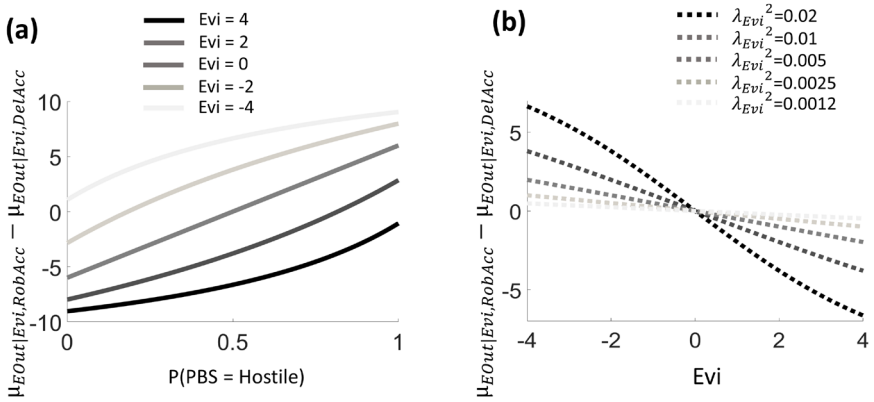

Fig. 2. Simulation of the model. The simulated scenario is discussed also in the main text, where Hyp includes two categories (Robbery hypothesis vs Delivery hypothesis), PBS includes two categories (Hostile vs Benevolent), and negative values of Evi support the Robbery hypothesis. The y axis reflects the posterior outcome value for accepting the Robbery hypothesis minus the posterior outcome value for accepting the Delivery hypothesis. **A**: The x axis reflects the prior probability for PBS = Hostile. Different lines indicate different values for Evi (for all lines, the precision parameter for Evi is $\lambda_{Evi}{}^2 = 0.005$, the outcome of accepting the Delivery hypothesis when it is true ($\mu_{EOut|Del,DelAcc}$) is equal to zero, the outcome of accepting the Delivery hypothesis when it is false ($\mu_{EOut|Rob,DelAcc}$) is equal to -10, the outcome of accepting the Robbery hypothesis when it is true ($\mu_{EOut|Rob,RobAcc}$) is equal to zero, the outcome of accepting the Robbery hypothesis when it is false ($\mu_{EOut|Del,RobAcc}$) is equal to -10). **B**: The x axis reflects the value of Evi. Different lines indicate different values for the precision parameter $\lambda_{Evi}{}^2$ (for all lines, P(PBS = Hostile) = 0.5 and other parameters are as above).

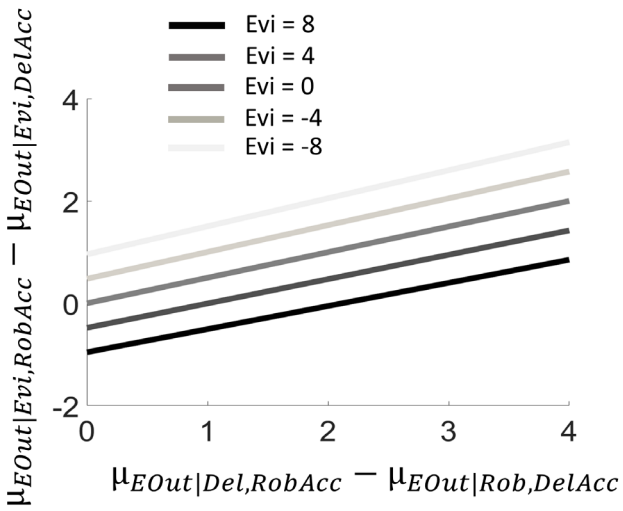

Fig. 3. Simulation of the model concerning expected outcome. The simulated scenario is discussed also in the main text, where Hyp includes two categories (Robbery hypothesis vs Delivery hypothesis), PBS includes two categories (Hostile vs Benevolent), and negative values of Evi support the Robbery hypothesis. The y axis reflects the posterior outcome value for accepting the Robbery hypothesis minus the posterior outcome value for accepting the Delivery hypothesis. The x axis reflects the difference between the expected outcome of accepting the Delivery hypothesis when it is false ($\mu_{EOut|Del,RobAcc}$) and the expected outcome of accepting the Robbery hypothesis when it is false ($\mu_{EOut|Rob,DelAcc}$). Different lines indicate different values for Evi (for all lines, P(PBS = Hostile) = 0.5, the precision parameter for Evi ${\lambda_{Evi}}^2 = 0.0012$, the expected outcome of accepting the Delivery hypothesis when it is true ($\mu_{EOut|Del,DelAcc}$) is equal to zero, the expected outcome of accepting the Robbery hypothesis when it is false ($\mu_{EOut|Del,RobAcc}$) is equal to -10, the expected outcome of accepting the Robbery hypothesis when it is true ($\mu_{EOut|Rob,RobAcc}$) is equal to zero).