Review

Spontaneous confabulation, reality monitoring, and the limbic system — a review

Armin Schnider*

Department of Rehabilitation, University Hospital, Av. de Beau-Séjour 26, CH-1211 Geneva 14, Switzerland

Abstract

Patients with anterior limbic damage may present a distinct syndrome, spontaneous confabulation: they fail in common memory tests, act on the basis of previous habits rather than currently relevant memories, produce confabulations composed of elements of past true events, are disorientated, and are absolutely convinced about the veracity of their perceived reality. Spontaneous confabulation is independent of other false memories, such as, provoked confabulations or illusory recognition. Studies showed that spontaneous confabulators fail to suppress (inactivate) evoked memories that do not pertain to ongoing reality. Rehabilitation differs from other memory failures. Prognosis depends on the lesion site, but recovery is always associated with recovery of this suppression capacity. Lesions typically involve the posterior medial orbitofrontal cortex or its connections in the basal forebrain. Imaging and evoked potential studies in healthy subjects support the idea that the anterior limbic system provides a reality monitoring mechanism which selects memories of current relevance by suppressing (inactivating) currently irrelevant memories. This mechanism appears to adjust the cortical representation of activated memories before their content is recognised and consolidated. Comparison with animal studies suggests that human reality monitoring is a property of the brain’s reward system. © 2001 Elsevier Science B.V. All rights reserved.

Theme: Neural basis of behaviour

Topic: Learning and memory: systems and functions

Keywords: Confabulation; Human memory; Amnesia; Limbic system; Orbitofrontal cortex; Reality monitoring; Rehabilitation

Contents

1. Introduction ................................................................................................................. 150
2. Clinical presentation of false memories .................................................................................. 151
   2.1. Spontaneous confabulation .......................................................................................... 151
   2.2. Other false memories .................................................................................................. 151
3. Mechanism of spontaneous confabulation .............................................................................. 152
   3.1. Previous interpretations .......................................................................................... 153
   3.2. Confusion of memory traces .................................................................................... 153
   3.3. Failure to suppress irrelevant memories ...................................................................... 154
4. Anatomical basis of spontaneous confabulation and reality monitoring .................................. 154
5. Rehabilitation and clinical course ....................................................................................... 156
6. Reality monitoring in thinking .......................................................................................... 156
7. Reward and reality monitoring — a hypothesis .................................................................... 158
Acknowledgements .............................................................................................................. 159
References .......................................................................................................................... 159

1. Introduction

Memories are the basis of reason and fantasy. They provide our feeling for what day it is, where we are, and
what we want to do. How can the brain allow our spirit to roam in fantasies and at the same time assure that we can always refer our thinking back to current reality? How does it monitor ongoing reality?

Brain-damaged patients with anterior limbic lesions, so-called spontaneous confabulators, fail in this capacity. They enact currently inappropriate habits and justify their behaviour with invented stories based on memories that do not pertain to ongoing reality — spontaneous confabulations. The study of these patients points to an anterior limbic mechanism monitoring reality in thinking, which acts by suppressing activated memories that do not pertain to ongoing reality. Evidence leading to this conclusion, as well as alternative interpretations of spontaneous confabulation, will be reviewed. Recent imaging and electrophysiological studies in healthy subjects indicate that this monitoring mechanism sets in before the content of a memory is recognised and consolidated.

2. Clinical presentation of false memories

2.1. Spontaneous confabulation

Spontaneous confabulators may appear entirely normal to the uninformed person. The account of their recent doings and of their plans for the day may be entirely cogent. Only the informed person may realise that the patient’s discourse is inappropriate for the present hospitalisation and indeed ignores his brain damage. A 58-year-old woman hospitalised following rupture of an anterior communicating artery aneurysm was convinced that she was at home and had to feed her baby; but her ‘baby’ was over 30 years old at the time [56]; a tax accountant with extensive traumatic destruction of the orbitofrontal cortex (OFC) inadvertently left the hospital in the conviction that he had a meeting with the county’s financial director [52,56]; a dentist, hospitalised after aneurysm rupture, repeatedly left the hospital in the conviction that patients were waiting for him at his clinic [40]. Very much like healthy people, most patients have preferred topics in their thinking [6,27,35,50,56,67]; the mother cited above was regularly concerned about her baby, the tax accountant repeatedly thought he had a business meeting, and the dentist was consistently concerned about his patients. Although it may appear that a patient has but one topic of confabulation [6], discussion usually reveals that the patient’s concept of reality and recent past is deranged in a general fashion. Questioning always reveals that the patients are disoriented; they confuse the place and time (they typically anedate it) and are unaware of the reason of the hospitalisation [50,57]. Often, they do not even realise that they have a memory failure. Nevertheless, they may suffer from people insisting on their having a bad memory, and they are permanently at odds with other people’s opinion about where they are and what they should do.

Confabulations are usually limited in time; they relate to the recent past, the present, and the future [50,56]. General semantic knowledge and remote autobiographic memory is typically, but not always [30], preserved. In rare cases, confabulations may extend over many years: an exceptional patient with extremely severe and productive spontaneous confabulation mixed elements of events from 25 years into the narrative of what he remembered as one single episode [51]. The confabulations thus appeared bizarre and incoherent, difficult to distinguish from psychotic thinking. Nonetheless, similarly with virtually all spontaneous confabulators, the elements of his discourse were ‘within the bounds of the conceivable’ [35] and could be traced back to true events. This feature is typical for confabulations in general [17,27,32,35,50,56,65,67,70].

The patients are absolutely convinced about their conceived reality and about the veridicality of their stories [6,27,35,50,51,56]. Trying to give them the correct date and to convince them that they are in the hospital because of brain injury often upsets them. It appears that the confabulations reflect the patients’ honest conception of ongoing reality. Spontaneous confabulation, as discussed here, constitutes a syndrome of profound derangement of thought, in which the concept of ongoing reality in thinking and planning is dominated by a patient’s past experiences and habits rather than true ongoing reality; the confabulations are simply the verbal manifestation of the thought disorder.

All spontaneously confabulating patients fail in common memory tests, in particular, to recall previously presented information after a delay. Although mostly present, several neuropsychological failures do not distinguish them from other, nonconfabulating patients with a similar defect of free recall. This concerns the production of intrusions in memory tests (words that were not presented), recognition performance, and frontal executive functions [49,51,56]. When pure storage capacity is tested with a conceptually simple task, such as a continuous recognition task (a series of items, among which item recurrences have to be detected, Fig. 1), spontaneous confabulators do not differ from nonconfabulating patients, either. Most notably, some of them demonstrate normal storage capacity [56]. Thus, a ‘gap in memory’ is not necessary for spontaneous confabulation to appear.

2.2. Other false memories

The most important feature of the syndrome of spontaneous confabulation is the inability to refer thinking and actions to ongoing reality; the patient’s concept of ‘now’ in thinking is determined by memories that do not refer to ongoing reality. Recollections of the past, too, are to a various extent inaccurate. The mechanism of spontaneous confabulation described below explains this phenomenon.
Fig. 1. The experiment. In the first run, subjects see a long series of pictures, among which several pictures are repeatedly presented. Subjects have to indicate picture recurrences. This run measures learning and recognition of new information. The second run is composed of precisely the same picture series, arranged in different order. Subjects are asked to forget that they have already seen all pictures and to indicate picture recurrences only within this current second run. d1, d2, first presentation of a picture in run 1 and run 2 (‘distracters’); T1, T2, picture recurrences within run 1 and run 2 (‘targets’). Variations of this task with different numbers of runs, items, and intervals were used in our studies [51,52,54±57].

Nonetheless, false recollection of past events in spontaneous confabulation has motivated many researchers to consider them a more severe form of the same disorder as other forms of false memories: provoked confabulations and illusory recognition.

Provoked confabulations occur when a subject tries to retrieve in detail an imprecise memory. Even healthy subjects pushed to recollect details about a past event (e.g. a holiday) may start to confabulate [7]. Patients with a memory disorder may produce intrusions of words, which were not presented before, when recalling a word list. We found that the number of intrusions positively correlated with the number of retrieved correct words and with a measure of verbal idea production (verbal fluency). Thus, provoked confabulations appear to be the pay-off for increased target retrieval. Provoked confabulations are independent of spontaneous confabulation, and vice versa [56]. Illusory recognition occurs when an item (for example, a word) is semantically closely related to true memories [41,45,47]. Subjects may be convinced about the veracity of such false memories. In contrast to spontaneous confabulations, which only occur after organic brain damage, provoked confabulations and illusory recognition can also be provoked in healthy subjects. They relate to what a subject perceives as events in the past and influence behaviour just like other, accurate memories. Although the content of these memories is erroneous, they do not induce a false representation of ongoing reality in thinking.

3. Mechanism of spontaneous confabulation

Whereas some clinicians suggested a distinction between provoked and spontaneous confabulations [5,29,68], most current theories assume that they represent different degrees of a common disorder with spontaneous confabulation being the more severe type [7,14,16,28]. Even though this assumption has been disproved [56], these theories will be reviewed here, given their massive influence on current concepts of confabulation and reality monitoring.
3.1. Previous interpretations

An early interpretation holds that confabulations represent a tendency to fill gaps in memory [3,38,70]. However, confabulating patients, in comparison with nonconfabulating amnesics, have no increased tendency to answer questions about nonexistent items for which they have a mandatory ‘gap in memory’ (‘where is Premola; who is princess Lolita; what is a waterknube?’) [34,56]. Some confabulators do not even have a ‘gap in memory’; they store normal amounts of information [56].

Another suggestion has been that confabulations are based on the combination of an amnesia with frontal executive failures [9,13,16,64]. However, controlled studies do not support this interpretation: common executive failures do not distinguish spontaneous confabulators from nonconfabulating amnesics [49,51,56], nor do they reliably parallel the course of spontaneous confabulations [52]. More specific theories propose that confabulations are based on a failure to correctly reconstruct a memory [52]. Specific theories propose that confabulations are based on a failure to correctly reconstruct a memory and to monitor its veracity [7,25,33,36,48]. However, these authors did not distinguish between provoked and spontaneous confabulations, explored only provoked confabulations and other illusory memories, or based their theory on observations in healthy subjects. Although these theories may be helpful to understand provoked confabulations and other false memories (false recall or recognition), the assumption that they also explain spontaneous confabulation is unwarranted.

A final view holds that confabulating patients confuse the temporal order and the context of memory acquisition [50,67,68] or that they have an impaired awareness for the flow of time in memory [10,68]. Although spontaneous confabulators typically fail in tasks probing conscious knowledge about the recency or temporal sequence of previously presented information [24,50], this failure is not specific for confabulations; even nonconfabulating amnesics and patients with dorsolateral frontal lesions, who have no discernible memory failure, may fail in such tasks [22,24,31,61]. Thus, lack of the knowledge about when and where information was acquired, does not explain the profound confusion of ongoing reality with currently irrelevant memories which characterises spontaneous confabulation. Nonetheless, spontaneous confabulators as a group, in comparison with nonconfabulating amnesics and healthy controls, have an impaired ability to discriminate short intervals in the range of seconds, the ‘now’. It is doubtful, however, that this failure, albeit typical for spontaneous confabulation, can explain the intrusion of currently inappropriate memories into ongoing thought [49].

A mechanism invoked to explain spontaneous confabulation should satisfy a number of criteria: it should reliably separate spontaneous confabulators from other, nonconfabulating brain damaged subjects, irrespectively of the aetiology of brain damage; it should explain the clinical course of spontaneous confabulation; it should explain the clinical phenomena associated with spontaneous confabulation, that is, the typical disorientation, the profound conviction of the patients that their memories are appropriate, and the observation that spontaneous confabulators’ false memories can virtually always be traced back to real events in their past. The following data from our own studies are based on the comparison of spontaneous confabulators with nonconfabulating, brain damaged patients with a similarly severe retrieval failure in a common verbal memory test. Patients were classified as spontaneous confabulators when they produced confabulated stories and were so convinced about their veracity that they occasionally acted according to them. Aetiologies of brain damage included traumatic brain injury, aneurysm rupture and other types of stroke, herpes encephalitis, olfactory meningioma, hypoxic brain damage, and others. Patients were only recruited once an initial confusional state had cleared.

3.2. Confusion of memory traces

A striking feature of spontaneous confabulations is that, on careful investigation, they can virtually always be traced back to elements of events in the patient’s past; the patient appears to re-experience an earlier episode as if it were now [17,32,35,50,56,65,67]. An inability to recall, when in the past an event happened, cannot explain this problem. What has to be explained is, how it comes that old, currently irrelevant memories can dominate thought in such a way that they are perceived as representing ongoing reality.

In our studies, we used variations of an experiment which measures the ability to distinguish between memories that pertain to ongoing reality and memories that have no current relevance; no distinction of the recency of last appearance nor knowledge about the temporal relationship among pieces of information is required (Fig. 1) [56]. In brief, the subjects make a first run of a continuous recognition task, that is, they are asked to indicate in a long series of pictures those that re-appear during the run. In the first run, any item that appears familiar can be assumed to be a repetition, that is, a target. Thus, the first run measures the ability to learn and recognise new information. One hour later, subjects are again shown the same picture series, arranged in different order this time. They are asked to forget that they have already seen all pictures and to indicate picture recurrences only within this second test run. The idea behind the experiment is that a subject’s false feeling that an item, which has actually been presented only in the first run yet (a distractor), has already been presented in the second run, reflects a deficient feeling for what memory is currently pertinent (only the memory of items that have already appeared in the ongoing second run).

The result confirmed the idea that spontaneous con-
fabulators fail to distinguish between currently relevant and currently irrelevant memories. Whereas they performed similarly to nonconfabulating amnesics in the first run of the task (learning and recognition), all spontaneous confabulators markedly decreased their performance in the second run. With no exception, they had a higher relative increase of false-positive responses than any nonconfabulating amnesic or healthy control [56]. The reliability of this task in distinguishing spontaneous confabulators was confirmed in all new patients we have since seen. We have encountered only one apparent exception, which resulted from a change in strategy (only ‘no’ responses) when the patient sensed the difficulty in the second run [40]. The biological validity of the task was further corroborated by two additional observations: (1) it perfectly paralleled the clinical course of spontaneous confabulation; only patients who stopped confabulating were again able to maintain performance in the second run (their false-positive rate in the second run no longer increased more than in other nonconfabulating amnesics or healthy controls). No other measure of explicit memory or executive capacities had comparable reliability [52]; (2) performance in the second, as compared to the first, run also explained disorientation, which is always present in early spontaneous confabulators. In a series of amnesic patients, orientation (tested with 20 questions) only moderately correlated with the performance in the first (learning) run. That is, the amount of information that a subject can store is only a weak predictor of his ability to maintain orientation. By contrast, relative performance in the second run (as compared to the first run) very highly correlated with orientation to time, place, and circumstances (r=0.93) [57]. Thus, orientation, too, depends primarily on the ability to sort out the currently pertinent information from memory.

3.3. Failure to suppress irrelevant memories

In a further study, we explored why spontaneous confabulators fail to distinguish currently irrelevant from currently relevant memories. One possible explanation was that the patients fail to represent new, incoming information with normal saliency in memory, so that associations of old, firmly established information would intrude into ongoing thinking [50,56]. A difficulty with this interpretation was that it would not have explained why patients with extremely severe amnesia with no measurable storage capacity do not normally confabulate [53,60] and why occasional patients also confabulate about old events [30,36,51]. An alternative explanation was that spontaneous confabulators cannot maintain a normal contrast between representations of ongoing reality and previously acquired information because they fail to suppress activated, but currently irrelevant, memory traces and mental associations in the face of current reality.

To test these possibilities, we used an adapted version of the continuous recognition task with four runs performed after different intervals [51]. We assumed that a failure to strongly represent incoming information would be mirrored in defective detection of recurring items (targets); that is, spontaneous confabulators should produce fewer hits. However, if the confusion resulted from an inability to suppress currently irrelevant memories, spontaneous confabulators should produce increasingly more false positive responses from run to run than the nonconfabulating amnesics.

In comparison with healthy controls, both spontaneous confabulators and nonconfabulating amnesics had difficulty in detecting target items, but they did not differ from each other. Thus, the failure to saliently represent incoming information (i.e. to learn and recognize new information) is typical for amnesia in general, but it does not explain spontaneous confabulation. In contrast, only spontaneous confabulators, but not nonconfabulating amnesics or healthy subjects, had a steep increase of false positive responses from run to run and failed to suppress this interference even when the interval between two runs was 30 min. Thus, spontaneous confabulation appears to result from a failure to suppress (inactivate) evoked memories that do not pertain to ongoing reality, a failure leading to continued saliency (activity) of currently irrelevant memories [51].

4. Anatomical basis of spontaneous confabulation and reality monitoring

A wealth of studies described provoked confabulations and illusory memories after dorsolateral prefrontal [36,47], medial temporal (hippocampal area) and other brain lesions [53,56], but also in healthy subjects [7,41,46]. That is, provoked confabulations have no anatomic specificity. By contrast, all patients described in the literature with sufficient detail to conclude that they had the syndrome of spontaneous confabulation, had lesions involving anterior limbic structures, in particular the OFC or its connections. Spontaneous confabulation has been extensively documented after basal forebrain lesions (which often include the posterior medial OFC), mostly due to rupture of an anterior communicating artery aneurysm [6,14,24,40,56] or traumatic brain injury [27,51,52,56,57,67]. Isolated damage of the medial OFC can also produce spontaneous confabulation [52,57]. In single cases, the lesions involved the following structures: the right capsular genu, which carries the projections of the dorsomedial thalamic nucleus to the OFC [50]; the amygdala on one side and the perirhinal cortex on the other side [54,56]; or the medial hypothalamus [26,39]. It is likely that damage of the dorsomedial thalamic nucleus (DM) also may produce spontaneous confabulation. There are some reports of long-standing spontaneous confabulation in alcoholic Korsakoff syndrome [30,32], in which damage of the DM has
been suggested to be particularly important [69]. Albeit rarely, paramedian thalamic infarction may also produce spontaneous confabulation [20].

The common denominator of these lesions is the posterior medial OFC, which has direct connections with all cited structures. A simplified interpretation of these data is summarised in Fig. 2. It appears that lesions involving the hippocampal area and its connections induce a failure to store normal amounts of information, that is, common, true amnesia [1,19,53,63]. By contrast, spontaneous confabulation appears to emanate from interruption of the loop connecting the posterior orbitofrontal cortex directly (via ventral amygdalofugal pathways) and indirectly (via the dorsomedial thalamic nucleus) with the amygdala. In addition, the medial, probably anterior, hypothalamus, which has direct connections with the posterior medial OFC, also appears to be important [39].

The system appears to have considerable variability: for all cited lesion sites, cases with impaired memory, but without spontaneous confabulation, have been described. In addition, most spontaneous confabulators eventually stop confabulating, indicating redundant organisation of this system [52].

These data thus indicate that the anterior limbic system provides a mechanism monitoring ongoing reality in thinking by constantly suppressing activated, but currently

---

**Fig. 2.** Anatomy of spontaneous confabulation and amnesia. (A) Lesions that have been reported in patients with spontaneous confabulation. (B) Main connections in this area. The black loop designates the classic Papez circuit (hippocampal loop). The mid-gray loop represents the lateral limbic loop connecting the orbitofrontal cortex with the amygdala. The bright gray connection indicates pathways of both loops with the septum verum. Amy, amygdala; aNc, anterior nucleus of the thalamus; bFb, basal forebrain; CG, capsular genu; Fo, fornix; Hipp, hippocampus; HT, hypothalamus; MB, mamillary bodies; MD, dorsomedial nucleus of the thalamus; MTT, mamillo-thalamic tract; OFC, orbitofrontal cortex; pHpc, parahippocampal gyrus; rSplC, retrosplenial cortex; SV, septum verum (medial and lateral parts).
irrelevant, memories. The outstanding role of the posterior medial OFC for this monitoring mechanism is supported by two additional observations. Firstly, patients with lesions involving the posterior medial OFC and basal forebrain confabulated for much longer periods (several months) than patients with anterior medial OFC lesions [52]. Secondly, in a study using H_2^15O positron emission tomography, healthy subjects performing a similar task as the patients showed circumscribed posterior medial OFC activation during the second run, which requires distinction between currently irrelevant and currently relevant memories. This activation contrasted with parahippocampal activation during the first (learning) run only [54].

**5. Rehabilitation and clinical course**

The rehabilitation of patients with spontaneous confabulation is a great challenge to any family and rehabilitation team. The patients act according to their perceived reality, tend to walk away from the unit, have no insight into their memory failure, and take corrections about where they are and what day it is as an offence, much like a healthy person would resent being constantly corrected. At the same time, they suffer from their permanent confrontation with the team’s idea about what is current reality.

There is no controlled study on the rehabilitation of spontaneous confabulation. In part, this may have been due to the lack of an empirically founded theory on its mechanism. Based on our studies, we have adopted the following approach: we first determine the level of surveillance that the patient needs with an accompanied visit to town; most patients properly react to the dangers of street life, such as, red lights and cars. Knowing that any cue may activate a memory and provoke a currently inappropriate action, patients receive information in as constant a way as possible. They receive information, mostly in written form, about the circumstances of their brain damage and their hospitalisation, but are not constantly questioned about it. A memory booklet with an agenda and space for notes, that provide feedback to the patient about his own actions, may be helpful [6], although some patients refuse to use it, and most fail to use it in a prospective way. Cumbersome behaviour (arguments with environment; swearing) may respond to self monitoring [12] which, however, does not tackle the basic problem of spontaneous confabulation — the false perception of ongoing reality. Given that this failure is not amenable to conscious control, we try to accept the patient’s reality as far as possible during the confabulatory stage; it is easier for a mother to accept that her baby has already received food than to convince her that her baby is over 30 years old; the tax accountant more easily accepts that his meeting has been postponed than that he is at the hospital because of brain damage. If conflict with the patients can be avoided in such a manner, they usually collaborate in the rehabilitation efforts.

Most patients with spontaneous confabulation eventually stop confabulating and regain correct orientation in time and space and the ability to refer their thinking and acting to ongoing reality [52]. The cessation of spontaneous confabulation is associated with recovery of the ability to suppress currently irrelevant memories. The duration of spontaneous confabulation depends on the lesion site; it may last only a few weeks after isolated damage of the anterior OFC, but months after combined basal forebrain and posterior OFC damage. However, individual patients with continued confabulation after such damage were described [12]. Nonetheless, continued spontaneous confabulation is rare; only one of our patients having extremely extensive OFC damage reaching up to the anterior horn of the lateral ventricles, thus interrupting projections from the DM to the whole prefrontal cortex, continues to have spontaneous confabulation after more than 5 years [52].

Vocational outcome is generally poor. Most patients remain amnesic, but they can live independently. Only those having isolated anterior OFC damage may have complete neuropsychological recovery. Unfortunately, even these patients may fail in private and professional life due to personality changes [15,52].

**6. Reality monitoring in thinking**

The necessity of ‘monitoring’ processes supervising memory retrieval has been emphasized by many authors. Terms like reality monitoring [23,25], self monitoring [4], source monitoring [25], supervisory system [7], strategic search in memory [33], and memory reconstruction [48] have been used almost interchangeably. These processes are thought to be mediated by the prefrontal cortex (PFC), by some authors specifically by the dorsolateral PFC. Their role is to check the veracity and plausibility of an evoked memory. Failure would lead to false memories — confabulation and illusory memories.

The difficulties with these hypotheses to explain spontaneous confabulation are that they are based on studies on phenomena such as provoked confabulations and illusory memories, which are distinct from spontaneous confabulation [56]; that they only consider the confabulations, rather than the profound misinterpretation of ongoing reality characterising the syndrome of spontaneous confabulation; and that they fail to provide an explanation for how the brain determines the veracity and plausibility of evoked memories, that is, how these mechanisms work.

The mechanism proposed here can also be considered a ‘reality monitoring’ mechanism, more specifically, a mechanism monitoring ongoing reality in thinking. It is not concerned with the monitoring of the plausibility (reality) of remembered past events. It does not actively select (reconstruct) memories; activation of networks represent-
ing memories is left to associative processes in the cortex [11,18,62]. What this monitoring process does, is to suppress (deactivate) activated memory traces that do not pertain to ongoing reality. Fig. 3 demonstrates the concept of this mechanism and the impact of its failure. In this illustration, every letter symbolises the representation of a piece of information of an event — a memory trace. In a healthy person (Fig. 3, normal), any new, incoming information (the reality of ‘now’) is thought to attain high saliency in cortical representation (‘X’) and to provoke mental associations. Some of these associations may lack connection with ongoing reality; they may be fantasies. The next piece of incoming information (‘P’) again attains high saliency and provokes mental associations. In addition, previous associations, which do not refer to current reality any more, are being suppressed (deactivated). The same happens when a new piece of incoming information is processed (‘Z’). Thus, this mechanism allows the healthy brain to make free associations (to roam in fantasies) but to remain capable of referring thinking back to reality at any time by inactivating those activated memory traces that do not pertain to ongoing reality.

In ‘common’ amnesic subjects, who do not spontaneously confabulate (Fig. 3, nonconfabulating amnesic), a new event attains high saliency and provokes mental associations. Any new, incoming information again attains high saliency and provokes mental associations. In contrast to healthy subjects, however, previously encountered information is not only suppressed, it cannot even be normally retained (consolidated). Thus, ‘now’ is unequivocally represented in thinking, but the information is subsequently forgotten.

In spontaneous confabulators (Fig. 3, spontaneous confabulator), new, incoming information is thought to provoke mental associations, just like in a healthy brain. However, when new pieces of information are processed (P, Z), those associations, which do not refer to the current reality any more, fail to be inactivated and remain active. A chaos of memory traces, their associations, and of new information, which does pertain to ongoing reality, guides thinking and behaviour. This mechanism is compatible with the observation that old, firmly established memories (habits) are more likely to influence thinking than weak or recently acquired memories. But it also explains that patients with severe spontaneous confabulation, who fail to suppress even distant mental associations, may confabulate about old events [30,36]. It is noteworthy that the patient who produced the most bizarre confabulations in our

Fig. 3. Model of ‘now representation’ in thinking. Any letter is meant to indicate an event or memory trace. The size of the letters indicates their saliency in thinking. See explanations in the text.
series, which occasionally encompassed events of 25 years, had the most severe failure to suppress false positive responses in the second to fourth run of our task [51].

The model, as hitherto presented, does not explain how, in the healthy brain mental associations that pertain to ongoing reality (‘this is 2001’) dominate over mental associations that do not pertain to ongoing reality (‘this is 1996’) to such an extent that they represent uncontested truth in thinking and provide the basis for behaviour. The easiest way to explain this would be that, before the content of an evoked memory (mental association) is recognised, it has already been checked and adjusted according to whether it relates to ongoing reality or not. This appears indeed to be the case. We used spatio-temporal analysis of full-scalp evoked potentials while healthy subjects performed a task similar to the one used in patients. Knowing that spontaneous confabulators had specifically failed to suppress false positive responses in the second run of the task (Fig. 1) [51,52], we were particularly interested in the electrical cortical response to these stimuli (‘distracters’ of run 2) as healthy subjects successfully suppress this interference. We found that correct suppression of these stimuli — the requirement for distinguishing between currently irrelevant and currently relevant memories — is associated with distinct alteration of cortical activity after 220–300 ms. This early alteration was characterised by the absence of a specific cortical potential map configuration and continued positivity of a frontal potential, rather than the negative deflection characterising all other stimulus responses (very first presentation of stimuli in the first run, i.e. ‘distracters’ of run 1 and targets of both runs). By contrast, learning and recognition, as evident from differences between the electrical responses to the first and repeated presentation of stimuli, was reflected in cortical amplitude modulation after 400–480 ms [55]. Thus, by the time the content of a mental association is recognized and consolidated, its cortical representation has already been adjusted according to whether it relates to ongoing reality or not. This sequence not only explains the absolute conviction that healthy subjects, but also spontaneous confabulators, have about their interpretation of ongoing reality, it also explains our ability to distinguish between the memory of a true event and the memory of a thought; the representation of these evoked memories had already been adapted as they entered the stage of re-consolidation.

7. Reward and reality monitoring — a hypothesis

The lesion analysis of patients with the spontaneous confabulation syndrome [51,57] and our finding that correct performance in the second run of our task activates the posterior medial OFC of healthy subjects [54] indicate that the alteration of cortical activity after 220–300 ms is mediated by the anterior limbic system, in particular the posterior medial OFC. Animal experiments have shown that this area has a central role in the processing of reward [42,59]. It would appear to be a waste of resources if the reality monitoring mechanism discussed here represented an independent faculty of this area. The failure of spontaneous confabulators is characterised by an inability to adapt behaviour to ongoing reality; the patients act on the basis of currently inadequate memories, that is, on the basis of expectations that justly motivated their behaviour in the past, but which have no current relevance. A similar failure was described in animals. Monkeys with ablations of the posterior medial OFC continue to react to stimuli which are no longer rewarded, that is, they have a severe deficit of extinction [8]. In other words, monkeys with posterior medial OFC ablation continue to act on the basis of the now irrelevant memory that a cue was previously followed by reward. The posterior medial OFC of monkeys contains neurons that specifically increase their firing rate when an expected reward fails to be delivered; that is, they specifically fire in extinction trials [44,66]. In other words, these neurons signal the inconsistency of a memory with ongoing reality. If one accepts the idea that human behaviour, too, aims at the achievement of predicted goals, this model may be applied to human thinking. The assumption is that the goal of any action plan is the correlate of reward in animal experimentation. Of course, human action plans are much more complex. It is likely that at any time, myriad action plans, spanning disparate periods of time, have the potential of being elicited by environmental cues and thoughts. Meaningful behaviour is only possible when action plans evoked by external or internal cues can be suppressed when there is no chance of achieving the goal, that is, when the action plan does not pertain to ongoing reality.

How might such neurones in the posterior medial OFC suppress the influence of currently irrelevant memories on thinking and behaviour? Assuming that activated memories are cortically represented as the synchronous activity of neuronal populations [18,62], suppression of currently irrelevant memories might be exerted by simple desynchronisation of these populations [62]. Our finding that successful suppression of distracters in run 2 of our task by healthy subjects was associated with absence of a specific electrical cortical map configuration (which was present in response to all other stimulus types) is compatible with this idea [55]. But how, through what anatomical routes, might the OFC induce desynchronisation of cortical neuronal networks; how does it communicate with the neocortex? One suggestion has been that frontal–subcortical loops, which connect frontal cortex with distinct portions of the striatum, pallidum and substantia nigra, thalamic nuclei, and projections back to the cortex [2], might allow communication between the OFC and the neocortex [43]. Whereas the loops emanating from different cortical areas appear to be fairly segregated down to the level of the striatum [21], massive cross-communication and conver-
gence of fibres seem to be present at the level of the substantia nigra [37]. The OFC, which initially projects to the ventral striatum [21], might thus influence activity of wide areas of the neocortex. It is noteworthy that the non-delivery of an expected reward, which elicits increased firing of select neurones in the OFC, has been shown to inhibit firing of select neurones in the ventral striatum and the substantia nigra at the time when the reward is expected to be delivered [58]. This finding would be compatible with the idea that the OFC may actively suppress neuronal activity in subsequent sites of the loop connecting it with the neocortex.

The theory suggested here on how the brain monitors ongoing reality can be summarised as follows: it appears that the anterior limbic system, in particular the posterior medial OFC, is the essential mediator of the ability to refer thinking and behaviour to ongoing reality [54,56,57]. It exerts this role by constantly inactivating (suppressing) evoked memories that do not relate to ongoing reality [51,52]. This suppression influences the cortical representation of memories before their content is recognised and consolidated [55]. Findings from animal studies are compatible with the idea that human reality monitoring reflects a property of the brain’s reward system.

Acknowledgements

My studies were supported by the Swiss National Science Foundation (grants no. 4038-44052, 3238-62769, and 32-50882.97). Many people contributed to the work reported here; I thank in particular K. Gutbrod, A. Buck, C.M. Michel, and their co-workers for their support. K. Gutbrod, T. Landis, and M. Badan made helpful comments on the manuscript.

References

[33] B. Melo, G. Winoeur, M. Moscovitch, False recall and false


