Memory as reconstruction, *how people misremember their own lives*

A famous false-memory study, its critics and its 2023 replication, set beside evidence that freely recalled details of two staged real events were mostly correct.

MITWA LAB, MUMBAI

IN PLAIN WORDS

Psychologists have shown that some people, told that a relative remembers something happening to them as a child, will come to recall it although it never happened. The best-known such study was repeated in 2023 with five times as many people, and about a third were judged to remember the made-up event, though fewer said so themselves. Another study staged two unusual but real events and checked people's later accounts against a record; days to years afterwards, 93 to 95% of the checkable details they gave were correct, though they recalled under a quarter of what happened and most made at least one error. Under clean conditions memory seems mostly accurate about what it keeps, yet someone trusted who keeps suggesting a false past can bend it. How often everyday remembering is that clean is disputed.

A shopping centre that never was

In 1995 the psychologists Loftus and Pickrell gave 24 adults four short stories about their childhood, apparently supplied by an older relative. Three were true, and the fourth, invented by the researchers, placed the participant, as a small child, lost in a shopping mall. After a written booklet and two interviews, six of the 24 (25%) were judged to remember the invented episode in full or in part. The authors presented this as proof that such memories can exist, and said they could not estimate how common they are. We could not open the original paper, so this account follows the description given by Murphy et al. (2023).

The study arrived in the middle of a bitter argument, during the 1990s, over whether suggestive techniques used in some therapy could lead adults to recover memories of abuse that had not happened (Brewin and Andrews, 2017; Murphy et al., 2023). By 2023 it had been cited more than 1,700 times, and it is still raised by expert witnesses in court cases (Murphy et al., 2023). It also drew lasting criticism. Critics pointed to the small sample, to the paper's lack of a clear definition of a full or a partial memory and of any detail on how interviews were coded, and to the fact that the experimenters, rather than the participants, decided whether a memory was present (Murphy et al., 2023; Brewin and Andrews, 2017).

Fifteen per cent or thirty, depending on who counts

Many implantation studies followed, using similar designs with other invented events, such as a hot-air balloon ride shown in a doctored photograph. Brewin and Andrews (2017) gathered 22 datasets from 20 articles and sorted the outcomes by how strict the definition of a memory was. Counting any recollective experience at all, including images and speculation, the weighted average was 47%, with individual studies ranging from 12% to 93%. Requiring clear recollection together with some sign of confidence brought it down to about 15%. In one study where investigator ratings and self-ratings could be compared, investigators credited participants with a memory almost twice as often as participants did (36% against 20%), and many participants doubted what they recalled. The authors concluded that susceptibility to false childhood memories looks

more limited than often claimed, and added that a parent's word cannot always settle that an event never happened, since parents forget and may never have known.

Scoboria et al. (2017) took a different route. Rather than average published percentages, they obtained the interview transcripts from eight implantation studies (423 participants) and recoded all of them with a single scheme built on seven features, among them accepting the suggestion, elaborating beyond it, imagery, coherence and emotion. On that scheme 30.4% of participants had a false memory and a further 23% accepted the event to some degree without remembering it. When the suggestion carried personal details, included an imagination exercise and came without a photograph, the rate reached 46.1%. They criticised Brewin and Andrews for averaging across studies and conditions and for treating 15% as if it were a general limit; in their own pool some conditions were controls meant to produce few memories, so no pooled figure marks a ceiling. They judged that a false memory of abuse is probably harder to create than one of a childhood prank, but pointed out that their participants were undergraduates exposed to brief and fairly mild pressure, and that stronger techniques might lead more people to false memories.

The two reviews agree that implantation happens in a minority of participants and that its rate swings with procedure. Where they part is over what should count as a memory and which conditions belong in an average, and neither side has settled that by data alone.

The replication, with the participants asked as well

Murphy et al. (2023) repeated the mall study in Ireland with 123 participant pairs and preregistered their analyses, declaring in advance that a rate between 15% and 35% would count as a successful replication. Informants confirmed with confidence that the participant had never been lost; 63 further participants were dropped because their relative could not rule it out. Interviews ran over video, a week or two apart.

The coded rate of false memory rose at each stage, from 15% in the booklet to 19% after the first interview and 35% after the second. Most were partial memories (27%), with 8% coded as full. The authors linked the rise to repeated imagining and speculation encouraged by the interviews. Three participants held the memory from start to finish and even after debriefing, and might really have been lost; excluding them lowers the rate to 33%.

What makes the study more than a replication is what was added to it. Asked directly before debriefing, 14% of participants said they remembered being lost and another 52% said they believed it had happened without remembering it. The false memories were thinner than true ones, described in fewer words, rated less clear and backed by less willingness to testify. Agreement between the two coders on the memory categories was modest (kappa 0.60), as the authors acknowledge, so the headline figure depends partly on judgement. In a separate test, 1,006 Irish mock jurors read interview extracts about the invented event without being warned, and judged that the speaker was remembering it in 39% of cases. The authors themselves name the objection they could not test, whether a memory of getting lost can stand in for a memory of abuse, and say it lies beyond a replication.

Vivid, confident and drifting

A second line of research needs no deception. On 12 September 2001, Talarico and Rubin (2003) asked 54 students at Duke University how they had heard about the attacks of the previous day, and also to describe an ordinary event from the days just before. Each student was retested once, after one, six or 32 weeks. Both kinds of memory lost consistency at the same rate. What differed was the feeling that came with them; vividness, the sense of reliving and belief in accuracy faded for the ordinary memory and held for the dramatic one.

Hirst et al. (2015) followed this over a decade. A survey sent out within a week of the attacks was repeated about one, three and ten years later (11, 35 and 119 months), and 202 people completed all four. Most forgetting happened in the first year and then levelled off, while confidence stayed high throughout. Once a person's account of how they had heard the news had changed, the changed version tended to be repeated in later surveys rather than corrected. The sample was not representative of the American public, and only about a tenth of the original respondents stayed to the end, but the pattern matches the shorter study in suggesting that the confidence attached to a memory is a poor guide to whether it has stayed the same.

Exceptional memory does not protect against distortion either. Patihis et al. (2013) tested 20 people with highly superior autobiographical memory, identified by exceptional scores on quizzes about dates and public events, against 38 matched controls.

Both groups falsely recognised about 70% of related words that had never been shown, both took in misleading details after viewing a slideshow, and on a questionnaire 20% of the first group and 29% of the controls said they had seen footage of a plane crash for which no footage exists. The groups were small, though the effects were clearly above zero in both.

What recall looks like when it can be checked

Most of the work above measures how far memory can be pushed, whereas Diamond et al. (2020) asked what it does when nobody pushes it. Seventy-four adults went through one of two staged but real events, a hospital mask-fitting procedure or an audio-guided walk around a hospital, and later described it freely, between two days and several years afterwards. What they recalled amounted to less than a quarter of the available detail, a share that fell with age and with time; the rest was forgotten or simply not reported. Of the details they did give that could be verified, though, 93 to 95% were correct in every group, even if most participants made at least one error. When the authors asked memory researchers to predict the result, their median estimate for the most favourable case was 40% accuracy, and academics from other fields guessed 30%. The authors list the limits themselves. Both events were newer and more distinctive than most of daily life and were probably less exposed to discussion and retelling than personal or social experiences, and the two samples held 33 and 41 people.

Wixted et al. (2018) make the general case in their abstract, arguing that eyewitness memory is reliable when it has not been contaminated and is tested properly, and that the first uncontaminated test usually gave sound evidence even in most of the convictions later overturned by DNA. In a reply, Wade et al. (2018) accept the core point but warn against overselling it. They doubt that near-pristine conditions are typical, since witnesses often talk to others before any interview, and they cite studies in which roughly a tenth of what witnesses reported with complete certainty, under near-pristine conditions, was wrong.

Checking adult reports against records made at the time gives a messier picture, though the gap has several causes besides memory. Baldwin et al. (2019) pooled 16 studies with 25,471 participants in which childhood maltreatment had been recorded at the time and later asked about in adulthood. Agreement was poor (kappa 0.19). On average 52% of people with a recorded history did not report it as adults, and 56% of those who reported it had no matching prospective record. The authors are explicit that this does not show adult reports to be invalid, because prospective measures may miss many true cases, and that their results do not extend to allegations brought before the courts. The two measures, in their reading, pick out largely different people.

Why the two literatures can both be right

The disagreement is smaller than it looks, because the two bodies of work measure different things. Diamond et al. (2020) measured the accuracy of what people volunteer about a novel, one-off event that nobody had tried to alter, recalled days to years later. The implantation studies measured what can be added to memory when a trusted source vouches for an event, an interviewer usually encourages imagining it, and the questioning is repeated over weeks. Under the first conditions memory loses a great deal and most of what it does report is correct, whereas under the second, coders judged roughly a third of participants to have formed a false memory in the mega-analysis and in the 2023 replication (30.4% and 35%), though fewer participants said so themselves. This reconciliation is our reading of the evidence. It sits close to the common ground between Wixted et al. (2018) and Wade et al. (2018), who agree that memory can be contaminated and that uncontaminated reports are mostly accurate, and differ mainly over how often conditions are clean. Neither of the common intuitions fits it. In a representative telephone survey of the United States, 63% of respondents agreed that memory works like a video camera (Simons and Chabris, 2011), while the academics surveyed by Diamond et al. (2020) expected most freely recalled details to be wrong.

A newer question is what happens when the trusted source is a machine. Chan et al. (2024) showed 200 people a short silent video of an armed robbery. Some then went straight on to the memory test (the control group); the rest first answered questions, five of them misleading, in a survey, from a scripted chatbot, or from a chatbot running on a large language model that also affirmed wrong answers and elaborated on them. That generative chatbot produced more than three times as many immediate false memories as the control condition and 1.7 times as many as the survey, and a week later participants remained more confident in those false memories than controls. This is a single preprint, and its chatbot was built to mislead.

That result bears on our own work. In Technical Note 01 we argued that a companion inventing a memory costs the person more than one forgetting, partly

because an unnoticed invention can enter the person's own memory; Bloom's invented memories went from 127 of 400 checks to 0 of 400 after a change to the system, and how that was achieved is outside the scope of this piece. That figure concerns what Bloom itself recalls, and we have not measured any effect of Bloom on people's own memories.

What is not yet known

- Whether any implantation procedure tells us about memories of abuse. Every implantation study here suggested an event other than abuse, among them being attacked or having a serious accident, and the replication's authors say the question is beyond them.
- Which definition of a false memory should be standard. The gap between 15% and 30.4% turns on coding rules and on which conditions are averaged, and coders in the 2023 replication agreed only moderately.
- How participants' belief without recollection, reported by 52% in the replication, relates to later behaviour outside a laboratory.
- Whether the 93 to 95% accuracy of free recall holds for ordinary, repeated, emotionally significant events that people discuss with others, rather than novel staged ones.
- How often everyday remembering happens under the near-pristine conditions that the reliability argument assumes.
- Whether a conversational system that is not designed to mislead, but simply agrees too readily, changes human memory, and whether the 2024 chatbot result replicates in a larger preregistered study.

References

Baldwin, J. R., Reuben, A., Newbury, J. B., and Danese, A. (2019). Agreement between prospective and retrospective measures of childhood maltreatment: a systematic review and meta-analysis. JAMA Psychiatry, 76(6), 584-593. https://pmc.ncbi.nlm.nih.gov/articles/PMC6551848/

Brewin, C. R., and Andrews, B. (2017). Creating memories for false autobiographical events in childhood: a systematic review. Applied Cognitive Psychology, 31(1), 2-23. https://pmc.ncbi.nlm.nih.gov/articles/PMC5248593/

Chan, S., Pataranutaporn, P., Suri, A., Zulfikar, W., Maes, P., and Loftus, E. F. (2024). Conversational AI powered by large language models amplifies false memories in witness interviews. arXiv:2408.04681 (version 1). https://arxiv.org/abs/2408.04681

Diamond, N. B., Armson, M. J., and Levine, B. (2020). The truth is out there: accuracy in recall of verifiable real-world events. Psychological Science, 31(12), 1544-1556. Read as the authors' accepted manuscript. https://osf.io/ud63x

Hirst, W., Phelps, E. A., Meksin, R., Vaidya, C. J., Johnson, M. K., Mitchell, K. J., Buckner, R. L., Budson, A. E., et al. (2015). A ten-year follow-up of a study of memory for the attack of September 11, 2001: flashbulb memories and memories for flashbulb events. Journal of Experimental Psychology: General, 144(3), 604-623. Read as the authors' manuscript. https://sites.harvard.edu/schacter-memory/files/2022/09/2015-09755-001.pdf

Murphy, G., Dawson, C. A., Huston, C., Ballantyne, L., Barrett, E., Cowman, C. S., Fitzsimons, C., Maher, J., et al. (2023). Lost in the mall again: a preregistered replication and extension of Loftus and Pickrell (1995). Memory, 31(6), 818-830. Read as the authors' accepted manuscript. https://osf.io/nh3zq

Patihis, L., Frenda, S. J., LePort, A. K., Petersen, N., Nichols, R. M., Stark, C. E., McGaugh, J. L., and Loftus, E. F. (2013). False memories in highly superior autobiographical memory individuals. Proceedings of the National Academy of Sciences, 110(52), 20947-20952. https://pmc.ncbi.nlm.nih.gov/articles/PMC3876244/

Scoboria, A., Wade, K. A., Lindsay, D. S., Azad, T., Strange, D., Ost, J., and Hyman, I. E. (2017). A mega-analysis of memory reports from eight peer-reviewed false memory implantation studies. Memory, 25(2), 146-163. Read as the authors' accepted manuscript. https://wrap.warwick.ac.uk/id/eprint/83729/1/WRAP-mega-analysis-memory-reports-Wade-2016.pdf

Simons, D. J., and Chabris, C. F. (2011). What people believe about how memory works: a representative survey of the U.S. population. PLoS ONE, 6(8), e22757. https://pmc.ncbi.nlm.nih.gov/articles/PMC3149610/

Talarico, J. M., and Rubin, D. C. (2003). Confidence, not consistency, characterizes flashbulb memories. Psychological Science, 14(5), 455-461. https://dukespace.lib.duke.edu/server/api/core/bitstreams/db6afb03-2804-4eab-a559-8e75dd2c03ee/content

Wade, K. A., Nash, R. A., and Lindsay, D. S. (2018). Reasons to doubt the reliability of eyewitness memory: commentary on Wixted, Mickes, and Fisher (2018). Perspectives on Psychological Science, 13(3), 339-342. Read as the authors' accepted manuscript. https://publications.aston.ac.uk/id/eprint/32068/1/Reasons_to_doubt_the_reliability_of_eyewitness_memory.pdf

Wixted, J. T., Mickes, L., and Fisher, R. P. (2018). Rethinking the reliability of eyewitness memory. Perspectives on Psychological Science, 13(3), 324-335. Abstract only. https://pubmed.ncbi.nlm.nih.gov/29716454/