ChatGPT

How the Brain Builds Beliefs – A Research Approach

Introduction

Every day, people form beliefs about the world – from what we see and hear to what we consider true or false. Yet two individuals can experience the same event and walk away with completely different interpretations of "truth." A striking example is the viral *"dress" illusion*: some viewers saw the dress as white-and-gold while others swore it was blue-and-black, and brain scans revealed that those seeing white-and-gold showed extra activation in frontal and parietal brain regions (areas involved in higher cognition and visual integration) [1]. This simple illusion underscores a profound fact: each human brain actively *constructs* its own version of reality. Beliefs – our mental models of what is true – are not passively absorbed but are built through a complex interplay of perception, neural processing, personal experience, and social context. Understanding *how* the brain builds beliefs is crucial in an era where disagreements on scientific facts, political realities, or even what color a dress is can divide communities. This paper takes an evidence-based yet accessible journey from everyday experiences into the neurobiology of belief formation. We will explore how neural circuits, shaped by evolution, form and update beliefs; how education, religion, politics, economics and culture influence these beliefs; and how cognitive biases and misinformation can lead us astray. We also examine cutting-edge research – from brain imaging to neural network models – that illuminates belief formation, and consider how artificial intelligence (like large language models) might help us understand why humans adopt certain beliefs over others. Throughout, we synthesize insights from cognitive neuroscience, social psychology, and computational modeling, supporting each claim with recent peer-reviewed research. The goal is to shed light on how each human brain constructs its own "truth," and what that means for our shared reality.

Belief in Daily Life: Subjective Reality and Constructed Truth

Beliefs are our mental assumptions or interpretations about the world, and they often start from routine human experiences. From a young age, we learn *beliefs* about how objects behave (a ball dropped will fall), who we can trust, or abstract ideas like morality. Crucially, these beliefs are not flawless mirrors of reality – they are constructions of our brain. The brain continuously filters and interprets sensory input, filling in gaps based on prior experience and context. Everyday perceptual quirks illustrate this constructive process. Beyond optical illusions like the dress, consider how eyewitnesses to the same accident can recall different "truths," each sincerely believed. These differences arise because perception is entwined with belief: our brains make assumptions (often unconsciously) to make sense of ambiguous input. In the case of *The Dress* illusion, for example, individuals who assumed the dress was lit by bright sunlight unconsciously "subtracted" that light and saw darker blues and blacks, whereas those assuming a shadowy illumination saw lighter whites and golds [2] [1]. Neuroscientists found that those who perceived white-and-gold engaged frontal brain regions linked to top-down cognitive influence, essentially using their brain's prior assumptions to interpret the colors [1]. This shows how two brains, processing the same image, can generate two different beliefs about a basic attribute like color. In daily life, many of our beliefs – from trivial perceptions to deeply held worldviews – emerge from similar interactions between incoming information and our brain's predispositions.

Importantly, beliefs are more than just momentary perceptions; they are durable mental representations that shape future behavior and reasoning. Cognitive neuroscientists describe beliefs as the brain's way of storing interpretations of experiences in a meaningful way. One interdisciplinary model, called the **credition model**, posits that beliefs are the result of neural processes that evaluate incoming information in terms of personal meaning and relevance [3]. In this view, believing is a dynamic *process* that links our past experiences to our future actions: the brain encodes experiences (with emotions and context) into memory, forming beliefs that guide what we expect and how we behave [3]. For instance, a child who is bitten by a dog may form the belief "dogs are dangerous," which will influence that child's future behavior around dogs (perhaps avoiding them or feeling fear). Such beliefs might later be updated (e.g. a friendly dog can revise the belief), but at any given time our beliefs serve as a lens through which we interpret reality. People can even explicitly report these internal states – we say *"I believe..."* – reflecting the brain's ability to be aware of and communicate its constructed truths [4]. Thus, from daily routines to once-in-a-lifetime events, our brains are continually distilling experiences into beliefs, illustrating how subjective and constructive human "reality" can be.

Neural Mechanisms of Belief Formation and Updating

Beneath the rich tapestry of our beliefs lies a complex neural architecture. Modern brain imaging research has started to map *where* and *how* the brain builds and updates beliefs. A recent meta-analysis of functional MRI studies identified a core set of brain regions consistently involved in belief formation and belief updating [5]. Two hubs emerged as particularly important: the **precuneus** and the **temporo-parietal junction (TPJ)**. The precuneus, a region in the parietal lobe often linked to imagination and self-related thinking, was active both when people formed new beliefs and when they changed existing ones [5]. This suggests the precuneus may serve as a platform for constructing mental models of the world and for perspectival "mind travel," enabling us to envision different viewpoints or futures when updating our beliefs [5]. Meanwhile, the TPJ – a region on the brain's side roughly above the ear – showed specialized engagement during initial belief formation [6]. The TPJ is well-known for its role in **Theory of Mind**, our ability to understand others' mental states or to represent abstract information. Intriguingly, the meta-analysis found that *social* and *non-social* beliefs draw on partly different circuitry: forming beliefs about social situations (e.g. understanding another person's false belief) preferentially activated the right TPJ, whereas forming beliefs about impersonal facts or physical scenarios relied more on the left **dorsolateral prefrontal cortex (DLPFC)** [7]. In other words, the brain seems to use one network (involving the TPJ) when thinking about beliefs in a social context (navigating people's intentions or societal narratives) and a different network (involving the DLPFC) when evaluating evidence and forming beliefs about the nonsocial world [7]. The DLPFC is a region associated with logical reasoning and executive functions – its involvement in non-social belief formation suggests that analytic thinking and working memory play a bigger role when we form beliefs based on objective data or personal experience, as opposed to understanding interpersonal situations.

Not only do distinct networks handle different kinds of beliefs, but belief *updating* (changing a belief in light of new evidence) also has identifiable neural signatures. Studies have shown that adjusting one's belief recruits a **fronto-parietal network** in the brain, which is associated with shifting attention and cognitive control [8]. In particular, when people encounter evidence that contradicts a prior belief, activity increases in regions of the right lateral prefrontal cortex and parietal cortex – areas important for reappraising information and inhibiting one's prior perspective [8]. This neural pattern aligns with what one might expect: to change a belief, the brain must engage control processes to overcome the inertia of the old belief and incorporate new information. Interestingly, one meta-analytic study highlighted the **precuneus** (again) as a shared hub where belief formation and updating overlap [5]. The precuneus's role in imagining alternatives and shifting viewpoints may be why it's active not just when a belief is first formed, but also when it is revised – enabling us to "see" a new interpretation or truth when circumstances demand it [5]. Taken together, these findings support a model of **partially**

dissociable neural networks for beliefs: a midline core (precuneus and related areas) that constructs and recalls belief representations, and task-specific circuits (like TPJ or prefrontal regions) that come online depending on what kind of belief is being processed [5].

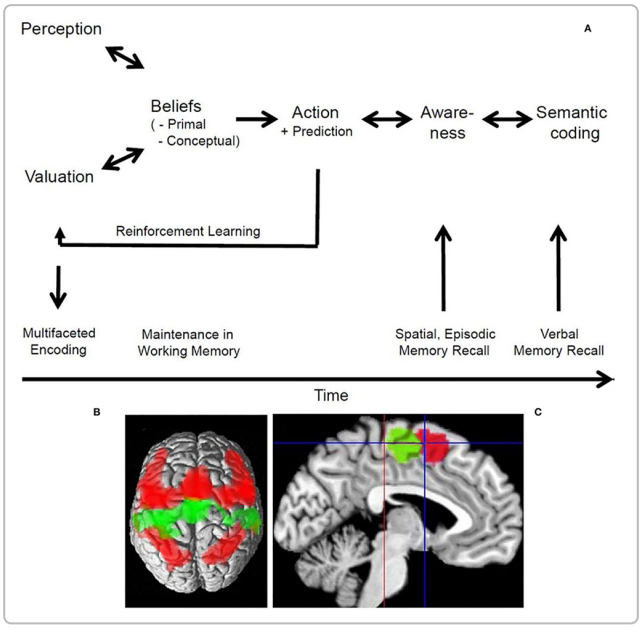

Figure 1: Schematic of the belief formation process in the brain (adapted from Seitz et al., 2022). Belief formation integrates perception (incoming information) with valuation (personal relevance/reward) and memory. Over time, beliefs guide actions and predictions. Brain imaging indicates that multiple regions – including frontal areas (red) and parietal areas (green) in panel B – are engaged during belief-related tasks, highlighting widespread neural participation. Panel C shows dorsal medial frontal cortex (green) activation, identified as a key hub when processing strongly held beliefs (such as religious content), compared to a neutral baseline. [9]

Beyond these general networks, certain brain structures play special roles in shaping the content and strength of beliefs. The **ventromedial prefrontal cortex (vmPFC)**, located in the front-middle of the brain behind the forehead, is one such area. The vmPFC is part of the brain's valuation and emotional circuit – it helps us attach subjective value or significance to information. Research indicates that the vmPFC contributes to belief flexibility: a healthy vmPFC supports a diversity of beliefs, whereas damage to this area can make one's beliefs more rigid. Notably, a study of brain-injured veterans found that patients with lesions (damage) to the vmPFC scored higher on a scale of *religious fundamentalism*, meaning they were more prone to rigid, literal interpretations of religious doctrines [10] [11]. In fact, damage to the **dorsolateral prefrontal cortex (DLPFC)** showed a similar effect – those with DLPFC lesions were as fundamentalist as those with vmPFC damage [12]. Crucially, the effect of DLPFC damage was traced to a loss of **cognitive flexibility** and openness to new ideas [11]. In other words, when the brain's "executive" regions are impaired, people have a harder time adapting or broadening their beliefs, sticking instead to familiar certainties. These findings underscore that the prefrontal cortex (both ventromedial and dorsolateral parts) normally acts as a brake on extreme or inflexible belief patterns – supporting open-mindedness and the capacity to update beliefs when warranted [12]. It also illustrates how interwoven cognition and emotion are in belief formation: the vmPFC links value and emotion to our interpretations, while the DLPFC contributes analytical reasoning and mental flexibility. If either component is compromised, beliefs can become extreme or unshakeable.

The brain's emotional centers further influence which beliefs take root. Strong emotional experiences tend to forge strong beliefs – for example, someone who endures a traumatic event may develop a firm

belief that "the world is dangerous." Neurobiologically, the **amygdala** (which processes fear and salience) and the **nucleus accumbens** (part of the reward circuit) both feed into belief formation by tagging information with emotional "weight." Studies have shown that these regions help encode whether new information is good or bad, important or irrelevant [13]. For instance, if hearing a news report yields a reward response in the accumbens (perhaps because it aligns with what we *want* to believe), that information may be integrated more readily into one's worldview. Conversely, if a fact triggers fear or distrust via the amygdala, we might reject it or form a belief that rationalizes the fear. The **hippocampus** and related memory structures are also critical – they store the experiences and facts that form the basis of our beliefs [3]. When we recall a belief (say, remembering why we trust a certain medical treatment), we are reactivating a network of memory traces, emotional associations, and conceptual knowledge distributed across the cortex. In short, building a belief is a whole-brain endeavor: *sensory areas* provide raw data, *memory circuits* contribute past context, *emotional circuits* supply value judgments, and *prefrontal networks* weigh and integrate these inputs to arrive at a conclusion that we treat as "true."

Cognitive Biases: How Our Brains Tilt the Scales of Belief

While the human brain is remarkably powerful at forming beliefs, it is not an unbiased or purely rational machine. We all have **cognitive biases** – systematic tendencies in thinking that can skew how we interpret evidence and thus what we come to believe. These biases are deeply rooted in neural processing strategies, often reflecting the brain's attempt to simplify complex information or protect us from psychological harm. One well-documented bias is the **confirmation bias**, where we favor information that confirms our pre-existing beliefs and discount information that contradicts them. At a neural level, confirmation bias has an identifiable signature: when we are strongly confident in a belief or decision, our brain literally processes incoming data differently. A 2020 neuroimaging study using magnetoencephalography (MEG) demonstrated that high confidence in one's initial decision causes the brain to *amplify* neural processing of supportive evidence and to *suppress* processing of dissenting evidence [14]. In the experiment, after participants made a decision with high confidence, any new information aligning with that decision triggered robust neural activity, whereas contradictory information was essentially ignored by the brain's update circuits [14]. This selective neural gating means that the more convinced we are, the less likely we are to change our minds – the brain's own processing reinforces the belief by shielding it from change. The finding is striking because it provides a biological explanation for why polarized and entrenched beliefs are so hard to budge [15]. If our brains treat confidence as a cue to stop listening to the other side, then simply presenting facts to a person with an opposing belief may fall on "deaf ears" neurally. Researchers concluded that *metacognitive* interventions (making people aware of their own thought processes and confidence levels) might be needed to counteract this bias [16] – essentially, teaching the brain to reopen the gate to disconfirming evidence.

Another common bias is the **optimism bias**, the tendency to update beliefs more in response to positive news than to negative news. Many people, for example, underestimate their personal risk of illness or accident because they discount information that implies vulnerability. Neuroimaging work by Bojana Kuzmanovic and colleagues uncovered a neural circuit underlying this bias [17]. In their study, participants estimated their likelihood of experiencing adverse events (like a cancer diagnosis), were given the real statistical risk, and then re-estimated their risk. As expected, people showed optimism bias: if the real stats were better (lower risk) than they thought, they eagerly adjusted to the good news, but if the stats were worse (higher risk), they tended to ignore the bad news and barely changed their belief [18]. The brain scans revealed why. Good news triggered interaction between the **ventromedial prefrontal cortex (vmPFC)** – a region that computes reward and value – and the **dorsomedial prefrontal cortex (dmPFC)** – a region involved in self-relevant reasoning [17]. Essentially, when hearing something desirable ("your risk is lower than you feared"), the brain's valuation center (vmPFC) signaled

a “thumbs up,” which biased the reasoning center (dmPFC) to integrate this news strongly into the belief update [19]. By contrast, bad news did not engage the vmPFC as effectively, so the update signal to the reasoning parts was weaker [17]. The result is a neural skew in information integration: favorable information gets weighted heavily and incorporated (leading to optimistic beliefs like “I’ll probably stay healthy”), whereas unfavorable information is down-weighted (“that grim statistic must not apply to me”). This optimism bias has obvious emotional benefits – it keeps us hopeful – but it can lead to distorted risk assessments and underestimating real dangers.

Many other biases operate on similar principles, where the interplay of emotion, reward, and attention in the brain tilts belief formation. For instance, **motivated reasoning** is a phenomenon where our goals or identities influence how we process facts. If a particular belief aligns with our political affiliation or self-image, our brain’s emotional centers may treat challenging evidence as a threat, triggering defensive reasoning. In neural terms, even the *perception* of information can be altered by prior beliefs: expectations can shape sensory processing (a form of predictive bias). The brain’s entire architecture is geared toward **predictive processing** – constantly anticipating input based on past experience. While this makes perception and comprehension more efficient, it also means our brains often confirm what they expect to see. In group settings, this can lead to **shared biases**: a recent theoretical paper on predictive processing suggests that when people interact, they may synchronize their expectations and error-corrections, effectively co-constructing a shared belief or ideology that minimizes surprise within the group [20]. In other words, there is a bias toward *consensus* in tightly knit groups – the brain’s prediction circuits align with those around us, because having a common belief system reduces social uncertainty (everyone is on the same page) [20]. This can explain how cultural or group beliefs become self-reinforcing: any deviant information that would create “prediction errors” for the group’s worldview tends to be dismissed or reinterpreted, keeping the group’s shared belief intact.

In summary, cognitive biases are not random quirks but reflect underlying neural processes that *favor* certain interpretations over others. Our brains bias beliefs in emotionally comforting directions, in ways that preserve our existing worldview, and in alignment with our social affiliations. These biases highlight why simply having access to factual information doesn’t guarantee belief change – how the brain *processes* that information is key. Understanding these neural biases is increasingly important in a world where misinformation and partisan content abound, as it helps explain why false beliefs can take such a strong hold. We turn next to how social and cultural forces interact with these cognitive biases in shaping collective beliefs.

Social, Cultural, and Educational Influences on Belief Formation

Beliefs do not form in a vacuum; they are deeply influenced by the social and cultural context in which a brain exists. From the family one is born into, to the schools one attends, to the broader society and media, external factors guide which ideas a person encounters and deems plausible. **Education** is one powerful influence. Generally, education provides not only knowledge but also training in critical thinking – tools to evaluate claims and evidence. Research during the COVID-19 pandemic illustrates education’s impact: a 2023 study found that individuals with a university degree were significantly *less likely* to believe false information about COVID-19, and correspondingly *more likely* to trust in preventive measures like vaccines and masks, compared to individuals without a degree [21]. In that study, educational level had a clear association with misinformation susceptibility – the more educated participants were, the better they could resist false claims about the virus and act on accurate public health advice [21]. Interestingly, the same study noted that participants’ general religious belief strength did *not* predict their vulnerability to COVID misinformation [22], suggesting that at least in that context, scientific literacy (often linked with education) mattered more than personal faith in discerning truth from falsehood. Education can impart cognitive skills (like scientific reasoning, statistical thinking) that act as antibodies against certain unfounded beliefs. However, education is not a perfect inoculation –

educated people are not immune to biases or echo chambers, especially outside their realm of expertise. In some cases, more educated or cognitively sophisticated individuals can become *more* polarized, because they are better at coming up with rationalizations to support their pre-existing views (the so-called "smart idiot" effect). Thus, while education broadly correlates with greater accuracy in beliefs (especially in science), its effect can be swamped by motivated reasoning when topics are politically or emotionally charged.

Culture and religion play a subtler but profound role by establishing the *baseline* beliefs that a person is exposed to and inclined to adopt. A child raised in a devout religious community will likely form beliefs aligned with that faith's teachings, not only because of direct instruction but because their brain's neural pathways for value and truth are tuned through early experience. Anthropologists note that what a society considers "common sense" often reflects deeply held cultural beliefs – for instance, beliefs about authority, morality, or the supernatural. Neuroscientific research shows that engaging with culturally learned beliefs (like reciting a prayer or affirming a political ideology) can strongly activate brain regions associated with *reward, emotion, and social cognition*. In one classic fMRI study, devout participants reciting a Psalm (a religious text) showed intense activation in the **dorsal medial prefrontal cortex**, a region associated with self-reflection and valuation, much more so than when they recited a neutral text [9]. This implies that culturally imbued beliefs (here, religious scripture) recruit neural circuits of personal meaning and significance. Culture also influences which cognitive biases are amplified. For example, many religions and political ideologies encourage a *confirmation bias* within their doctrine – believers are taught to interpret events as confirmation of the belief system. Likewise, cultural worldviews can determine emotional biases: an individualistic culture might instill an optimism bias about personal control ("you can do anything if you try"), whereas a culture that emphasizes fate or karma might bias believers to see events as confirmation of those forces.

Economic conditions are another contextual factor that can shape belief systems, often in unexpected ways. Hardship or inequality can breed beliefs that make sense of one's situation. For instance, when societies face high economic inequality or uncertainty, people may gravitate toward **conspiracy beliefs** or populist ideologies as a way to explain the imbalance. A growing body of research in social psychology suggests that economic stressors erode trust in authorities and fuel an *anomie* – a sense that the normal social order is breaking down – which in turn triggers conspiratorial thinking [23]. A recent review concluded that periods of high economic inequality tend to enhance conspiratorial beliefs by undermining the social fabric; people in such conditions are more likely to believe that unseen forces or elites are manipulating events [23] [24]. In other words, when the real world feels unjust or unpredictable, the human brain seeks patterns and causal stories – fertile ground for beliefs in hidden plots or scapegoats. Large-scale surveys across countries likewise find that the perception (and reality) of poor economic performance correlates with higher endorsement of conspiracy theories [25]. This does not mean economic hardship *automatically* causes irrational beliefs, but it creates a climate where simple, blame-oriented explanations gain appeal. **Political and media environments** then can channel these economic anxieties into particular beliefs (for example, blaming minorities or foreign powers for one's economic woes, a tactic often seen in populist propaganda). Neuroscience adds a layer of understanding here: stress and uncertainty (which accompany economic strain) can shift brain function toward more fear- and habit-based processing (relying on the amygdala and "gut feelings"), at the expense of deliberative reasoning. Thus, someone who feels economically threatened may quite literally have a brain primed to accept bold, emotionally charged claims that promise certainty or someone to blame, as those satisfy emotional needs even if they lack factual basis.

Finally, **social identity** and group membership are powerful shapers of belief. Humans are deeply social creatures, and our brains have evolved to synchronize with our in-groups. We tend to adopt the beliefs of those we identify with – it's a form of social bonding and a shortcut to trust. For example, political partisans often internalize their party's stance as personal belief. The phenomenon of **identity fusion**

takes this to an extreme: in identity fusion, an individual's personal self-concept merges with a larger group or leader [26]. A dramatic case was observed among some supporters of former U.S. President Donald Trump. Research found that certain core Trump supporters experienced a "synergistic union" with Trump – their personal identity and his public persona became psychologically intertwined [26]. This fusion had a striking effect on belief formation: for strongly fused individuals, accepting Trump's claims (even baseless ones) became a way to affirm their own identity. During the 2020 U.S. election and its aftermath, Trump propagated the unfounded "Big Lie" that the election was stolen from him. Surveys of Trump's fused followers showed that the stronger their identity fusion with Trump, the more readily they adopted this false belief, essentially seeing Trump's defeat as a personal injustice against themselves [27] [28]. What's more, believing the Big Lie further reinforced their bond to Trump and made them more likely to embrace other related falsehoods and support his agenda [28]. In neural terms, what likely happens in such scenarios is that the brain's social bonding and reward circuits (like the vmPFC, which tracks value, and oxytocin-related systems for affiliation) make accepting the group's belief feel inherently rewarding – it signals "I belong." Meanwhile, disagreeing with the group or leader would trigger conflict and pain regions, as it threatens a core identity. Thus, social influence can literally rewire what is true in someone's mind. Charismatic public figures can have an outsized impact by serving as *authority signals* – if a trusted leader says something, followers' brains may treat it as highly salient truth, sometimes overriding evidence to the contrary. This effect is amplified by repetition and social reinforcement (if "everyone" in my community says it, it must be true). We next delve more into the modern dynamics of misinformation and how these social factors play out in today's information ecosystem.

Misinformation, Public Figures, and the Collective Construction of Truth

In the digital age, misinformation – false or misleading information spread unintentionally or deliberately – has become a significant factor in collective belief formation. Neuroscience and psychology research together paint a picture of why misinformation is so potent at molding beliefs. One fundamental mechanism is the **illusory truth effect**, the tendency for repeated statements to be believed more readily than new ones. When a piece of misinformation (for example, a fake news headline) is encountered over and over – through social media shares, news coverage, or word of mouth – people develop a sense of familiarity with it, which the brain can misinterpret as plausibility. Cognitive experiments show that even a single repetition can significantly increase the perceived truth of a statement [29]. This occurs regardless of age or even prior knowledge – adults will rate a known false statement as more likely true if they've heard it before, compared to an unknown statement they've only heard once [29]. The brain's heuristic here is "if I've heard this often, it must be true," likely because in natural settings repetition often correlates with reliability (or at least with social consensus). Unfortunately, this heuristic is easily exploited by misinformation: a false claim can gain credibility by sheer volume of repetition. Modern social media platforms algorithmically amplify content that gets engagement, which means sensational misinformation often gets repeated and echoed widely, strengthening the illusory truth effect in the public's minds. Neuroimaging studies of memory suggest that repeated information is stored more fluently and pops up more readily in the hippocampus and cortical memory networks, giving a person the gut feeling that "I recall this, so it must be true." Breaking this cycle requires conscious effort – as psychologist Lisa Fazio notes, one must slow down and critically evaluate information rather than trusting the gut familiarity signal [30].

Public figures and authority figures play a dual role in misinformation dynamics. On one hand, trusted figures can debunk falsehoods and guide their followers toward factual beliefs. On the other hand, if the public figure themselves spreads misinformation (intentionally or not), it can dramatically accelerate the spread and acceptance of those false beliefs. Earlier we discussed the case of Donald Trump's false

election claims and how his devoted followers internalized them [26] [28]. That episode also highlighted a broader phenomenon: **motivation to believe the leader**. For many, aligning with the claims of a charismatic leader or a group doctrine becomes a higher priority than objective accuracy. Social psychology research during the COVID-19 pandemic, for instance, found that partisan alignment often swayed people's beliefs about the virus more strongly than the scientific evidence did [31]. In one study, Republicans' perceptions of COVID-19 risk and the efficacy of health measures were more closely tied to signals from party elites and media (some of which downplayed the virus) than to their educational background or knowledge of the virus. The effect of public figure influence is especially powerful when those figures position themselves as *the* arbiters of truth and cast doubt on other sources (e.g., "don't trust the mainstream media, only I tell you the real story"). In those scenarios, followers may adopt a kind of **filter bubble in the brain** – essentially directing attention only to information endorsed by the figure, and automatically treating dissenting information as false. Brain imaging of people who are ardent political partisans has shown that when they read statements from their preferred politician, reward centers (like the striatum) can activate if the statement affirms their side, whereas threat/distrust regions (like the insula) activate for statements from the opposing side, regardless of content truth. This neural "tagging" of sources as friend or foe means that who delivers the information can matter more than what is being said.

Misinformation is also often *designed* to exploit cognitive biases. Conspiracy theories, for example, typically provide emotionally compelling narratives that tap into fear or outrage (engaging the amygdala) and offer simple explanations for complex problems (satisfying the brain's desire for pattern and causality). Once a person begins to believe a conspiracy theory, confirmation bias kicks in strongly – they start seeing every new event as further evidence supporting the conspiracy, while dismissing official accounts as naïve or deceptive. This is exacerbated in online communities where people can find echo chambers of like-minded believers, creating a collective confirmation bias. **Collective belief formation** in the internet era thus has a strongly self-reinforcing character: beliefs (true or false) cluster among groups and each group curates its own reality. We see this in phenomena from anti-vaccine groups to flat-Earth believers – within their communities, they cite "evidence," share anecdotes, and build an entire alternate knowledge system that can be impervious to outsider correction. Each individual brain in such a network is continuously receiving social proof from peers that their belief is valid, which is one of the most powerful drivers of conviction.

Interestingly, the neuroscience of **"collective truth"** is starting to be investigated. When many people believe something, does it become more compelling to an individual's brain? Early research suggests yes: hearing that a majority of one's peers believe X can activate neural circuits related to reward and conformity (dopaminergic systems that make agreeing feel good) as well as reduce activation in conflict-monitoring regions (anterior cingulate cortex), effectively signaling "no need to worry, go with the flow." In contrast, standing against the group lights up stress and conflict circuits, which can be aversive. Thus, misinformation that gains a foothold and becomes a widely held *collective* belief (like a widespread false rumor) can attain a kind of inertial force – it becomes part of the collective's reality, and individuals' brains will align with it to avoid social discord.

At the extreme end, entire information ecosystems can split, as has happened in some countries where political polarization leads to completely separate media spheres. In such cases, each side's brains are being trained on different "facts" day in and day out. Public figures act as rallying points in these ecosystems – for example, if a prominent influencer repeatedly claims a certain unsubstantiated "alternative truth," their followers' brains may effectively incorporate it into their worldview. Combating misinformation, therefore, is not just about supplying correct data; it involves understanding identity, trust, and cognitive bias. Interventions like debunking or fact-checking have to be done carefully to avoid backfire effects (whereby correcting a false belief paradoxically strengthens it, because the correction threatens the person's identity or worldview). Some strategies that show promise include

“prebunking” (inoculating people with brief warnings about common misinformation techniques), encouraging analytical thinking (e.g. asking people to rate the accuracy of news headlines, which prompts more critical evaluation), and leveraging *trusted insiders* to deliver corrections (since people are more receptive to information from within their group).

In summary, in the social-media-driven public sphere, each human brain constructs truth not only from direct evidence but from a swirling mix of repeated messages, social cues, and emotional resonance. Misinformation leverages our neural and social predispositions to propagate beliefs that may be sharply at odds with objective reality. Understanding this challenge sets the stage for considering new approaches – including insights from artificial intelligence – to studying and potentially guiding belief formation.

Insights from Artificial Intelligence and Computational Modeling

As we grapple with the question of why humans believe what they do, researchers are increasingly turning to **computational models and artificial intelligence (AI)** as tools for understanding belief formation. The brain itself can be thought of as a prediction engine – a concept from neuroscience known as **predictive processing**. According to this theory, the brain is constantly generating predictions about incoming sensory data and updating its internal model (beliefs) based on the errors between prediction and reality. This perspective aligns well with certain AI systems, especially modern machine learning models that learn to predict outcomes from data. By studying these systems, or using them as analogues, scientists hope to illuminate the hidden dynamics of human belief formation.

One area of convergence is in modeling how beliefs are updated. **Bayesian models** have long been used as theoretical frameworks for belief updating, treating beliefs as probabilistic hypotheses that get strengthened or weakened by new evidence. The human brain doesn’t always follow pure Bayesian logic (because of biases and cognitive limits), but computational models allow researchers to tweak “belief-updating parameters” to see what might cause patterns observed in humans. For example, in computational psychiatry, researchers have modeled how **delusions** (fixed false beliefs seen in mental illness) could arise from the brain assigning aberrant precision or weight to certain prediction errors. In a Bayesian sense, a delusional brain might be too **rigid** in priors (overconfident in a belief, underweighting disconfirming evidence) or conversely might overly adjust to random noise, seeing patterns that aren’t real. Such models have been used to simulate phenomena like the formation of paranoid beliefs or the failure to update beliefs in depression [32]. The simulations suggest that small changes in how the “updating algorithm” of the brain works – possibly due to neurotransmitter differences – can produce large differences in belief behavior, mirroring what we see in different individuals.

Artificial neural networks, especially **large language models (LLMs)**, offer another intriguing window into human-like belief reasoning. LLMs like GPT-4 are trained to predict the next word in a sentence based on massive amounts of text. Through this training, they develop internal representations of the relationships between concepts, which can be thought of as a kind of knowledge or even proto-belief about the world. Remarkably, recent studies show that advanced LLMs have begun to display *Theory of Mind*-like abilities – that is, they can answer questions about what one person might know or believe about a situation, a task that requires attributing mental states. For instance, a 2024 study in *Proceedings of the National Academy of Sciences* tested several LLMs on classic false-belief tasks (the kinds of tasks used to assess Theory of Mind in children) [33]. Early-generation models failed completely, but newer models performed much better. GPT-3.5, released in 2023, could solve about 20% of these tasks, and the most advanced model at the time, GPT-4, solved roughly 75% – a success rate comparable to a 6-year-old child [34]. This suggests that through statistical learning from human language, the AI

developed some capacity to predict and infer beliefs. The researchers even speculated that a rudimentary "belief understanding" might be an emergent property of these models' complexity [35]. However, it's important to note that whether LLMs *truly* understand beliefs or are just mimicking patterns is debated. Follow-up research has both supported and challenged the depth of LLMs' social reasoning. On one side, an analysis found that one can *decode* the model's internal activations to identify representations of different agents' beliefs – implying the model has distinct neural-like states corresponding to "what it thinks X believes" versus "what is actually true" [36]. Moreover, by intervening on these internal representations, the model's answers about others' beliefs change in predictable ways [36], strengthening the case that something analogous to belief attribution is happening inside. On the other side, skeptics argue that LLMs lack genuine understanding and that their apparent Theory of Mind is brittle or breaks with slight rephrasing of problems [37]. They caution that LLMs might be picking up on superficial cues in language rather than actually simulating another mind's perspective [38].

Regardless of this philosophical debate, AI models are proving useful as *tools* for probing hypotheses about belief. For example, one can use an LLM to simulate how a narrative or piece of misinformation might spread or be interpreted by different personas (e.g., prompt the model to act as a conspiracy believer versus a skeptic and see how each "completes" information). If the model's completions differ in telltale ways, it might highlight the key assumptions that distinguish those mindsets. AI models can also be used to test interventions: researchers have started experimenting with using chatbots or automated systems to deliver cognitive debiasing strategies to human users. In one study, an AI-based "accuracy nudge" was inserted into social media feeds – essentially, the AI would occasionally ask users "Do you think this headline is true?" for random posts. This simple prompt, designed by cognitive scientists and delivered at scale by an algorithm, significantly improved people's ability to discern truth from falsehood by refocusing their attention on accuracy [30].

Moreover, computational modeling of social networks – often powered by AI simulations – helps researchers understand phenomena like echo chambers and tipping points for belief change. **Agent-based models** allow us to simulate thousands of "agents" (simulated people) with certain bias and network connection parameters, and then introduce a piece of misinformation or a factual correction and see how it propagates. These virtual experiments can reveal, for instance, how a small fraction of highly connected agents (hubs) can rapidly amplify a false belief, or how slight increases in open-mindedness (willingness to weigh disconfirming evidence) across a population can dramatically improve collective accuracy. Insights from such models complement empirical neuroscience and psychology. They show, for instance, that if agents have a confirmation bias in their updating rule (very much like humans do), the network tends to polarize into stable camps of opposing beliefs, even if everyone is exposed to the same mixed evidence. This echoes real-world observations in our society.

Finally, AI offers potential *remedies* or aids for human belief formation. Large language models might act as "simulated interlocutors" to help people examine their beliefs. For example, a person could have a dialogue with an AI that plays devil's advocate to their beliefs, exposing them (in a non-confrontational way) to alternative viewpoints or highlighting inconsistencies. Early studies suggest that people may sometimes find it easier to consider counterarguments when presented by a neutral AI than by a human (where ego and social judgment get involved). There are also proposals to use AI in identifying misinformation (through content analysis algorithms) and then tailoring corrections to users in psychologically savvy ways – such as explaining why a false story might feel compelling (thus addressing the emotional hook) before presenting the factual correction.

In essence, artificial intelligence is both a mirror and a magnifying glass for human belief processes. By creating systems that (in very simplified form) emulate aspects of human cognition – from predicting text to attributing mental states – we can test theories about how and why certain beliefs form. AI is

also a burgeoning participant in our information ecosystem, raising the stakes: as AI-generated content and deepfakes emerge, understanding the brain's vulnerabilities in belief formation becomes even more critical. The interdisciplinary approach, combining neuroscience, psychology, and AI modeling, is thus opening new frontiers in deciphering the age-old question of why humans believe as they do.

Conclusion

Human brains are belief-generating machines. From the mundanity of perceiving a dress's color to the complexity of constructing a political ideology, our brains continuously transform experiences and information into beliefs that we treat as reality. This paper explored how that process unfolds at multiple levels of analysis. At the neural level, belief formation engages widespread brain networks – intertwining cognitive regions that handle logic and evidence with emotional and social regions that imbue personal meaning. We saw that specific hubs like the precuneus and TPJ are instrumental in constructing and updating beliefs, and that the balance of activity in frontal regions can determine whether a mind is open to change or locked in rigid convictions [5] [12]. At the psychological level, inherent biases tilt the scales of belief: our brains preferentially accept what is familiar, desirable, or identity-affirming, which explains why *misinformation* and *bias-confirming narratives* can be so persuasive [29] [14]. At the social level, education, culture, and group identity set the menu of beliefs we find plausible and provide social reinforcement for those beliefs. Beliefs, especially shared ones, serve as social glue – but also, as history shows, as fault lines when different groups' versions of "truth" collide.

Crucially, each human brain constructs its own version of truth not in isolation but through constant dialogue with its environment and others. The interplay of neurons and society means that correcting false beliefs or bridging belief divides is not as simple as "delivering facts." It involves understanding the *framework* in which a belief resides – the neural reward it might be providing, the social identity it supports, the cognitive biases that maintain it. For instance, countering a conspiracy belief might require reducing the person's feelings of anxiety or powerlessness (the soil in which conspiracies thrive [23]), as well as providing an appealing alternative narrative that can satisfy the same psychological needs.

Emerging research on brain imaging and modeling provides both hopeful and cautionary notes. On one hand, the fact that the brain's confirmation bias and optimism bias have identifiable neural mechanisms [14] [17] suggests we could develop interventions (pharmacological or behavioral) to modulate those mechanisms. Could we train people to recognize when high confidence might be blinding them, akin to a "metacognitive safety catch"? Can educational programs bolster prefrontal cognitive flexibility (perhaps via games or mindfulness practices) so that people are less prone to fundamentalist, all-or-nothing thinking? These are open questions, but the neuroscience points to potential levers. On the other hand, the same science underscores how natural these biases are – they are features of our brain's design, not bugs. So any solution must work with the grain of human psychology, not against it.

Artificial intelligence adds another layer of complexity. As we deploy AI systems that can generate human-like persuasive text, or as we see algorithms curate our information feed, we must be mindful of how these technologies interact with our belief-building brains. The research indicating that large language models can mimic some aspects of human belief reasoning [35] is fascinating – it tells us that certain patterns of thinking emerge from the mere act of prediction. But AI can also mislead or overwhelm our cognitive systems if not carefully aligned with our values and critical thinking needs. There is an optimistic angle: AI might become a partner in helping us recognize our biases (imagine an "AI truth coach" that alerts us gently when we are accepting something too readily due to familiarity or

partisan bias). Yet there is also a risk that AI could supercharge misinformation (e.g., deepfakes that are neurologically maximized to grab attention and appear credible).

In the end, understanding how the brain builds beliefs is more than an academic quest – it is vital for addressing challenges ranging from public health to political polarization. By synthesizing neuroscience, psychology, and computational insights, we move closer to a holistic understanding of belief. Each person's brain is a universe of experiences woven into a narrative that feels like truth. Appreciating this can foster empathy: rather than dismissing those with opposing beliefs as simply "ignorant" or "irrational," we can recognize the powerful cocktail of neural processes and life influences that lead people to their convictions. It also empowers us to reflect on our own beliefs: to what extent do *I* hold certain ideas because of how my brain is wired to seek comfort, or because of the culture I grew up in, or because of people I admire? Such reflection is the first step toward what the ancient philosophers advised – "know thyself" – and in this context, know how your brain is constructing your reality. Armed with that knowledge, we can strive to build beliefs that are more aligned with evidence, more open to revision, and perhaps, in the collective, a bit closer to the truth that transcends any single brain's version of it.

References: (Selected key sources supporting this paper's content)

- Azari, N.P., et al. (2001). *Neural correlates of religious experience.* **European Journal of Neuroscience, 13(8),** 1649–1652. [9]
- Fazio, L.K., et al. (2020). *Repetition increases perceived truth equally for plausible and implausible statements.* **Psychological Science.** [29]
- Jetten, J., Peters, K., & Casara, B.G.S. (2022). *Economic inequality and conspiracy theories: Social psychological consequences.* **Current Opinion in Psychology, 47,** 101358. [23] [39]
- Kuzmanovic, B., Rigoux, L., & Tittgemeyer, M. (2018). *Influence of vmPFC on dmPFC predicts valence-guided belief formation.* **Journal of Neuroscience, 38(37),** 7996–8010. [17]
- Lo Presti, S., et al. (2025). *"Don't stop believing" – Neural correlates of belief formation and updating: An ALE meta-analysis.* **Neuroscience & Biobehavioral Reviews, 173,** 106153. [5]
- Rollwage, M., et al. (2020). *Confidence drives a neural confirmation bias.* **Nature Communications, 11,** 2634. [14]
- Swann, W.B., et al. (2024). *Identity fusion and acceptance of misinformation.* **(Study on Trump supporters' belief in the "Big Lie").** [26] [28]
- Zhong, W., et al. (2017). *Biological and cognitive underpinnings of religious fundamentalism.* **Neuropsychologia, 100,** 18–25. [10] [11]
- Zhu, W., et al. (2023). *Language models represent beliefs of self and others.* **Proceedings of ICML (arXiv preprint arXiv:2402.18496).** [36] [37]

---

[1] 'The Dress': Explanation of optical illusion of colors of the striped dress | ScienceDaily
https://www.sciencedaily.com/releases/2015/10/151014085416.htm

[2] The dress - Wikipedia
https://en.wikipedia.org/wiki/The_dress

[3] [4] [8] [9] [13] Frontiers | Believing and Beliefs—Neurophysiological Underpinnings
https://www.frontiersin.org/journals/behavioral-neuroscience/articles/10.3389/fnbeh.2022.880504/full

[5] [6] [7] "Don't stop believing" - Decoding belief dynamics in the brain: An ALE meta-analysis of neural correlates in belief formation and updating - PubMed
https://pubmed.ncbi.nlm.nih.gov/40228650/

10 11 12 Biological and cognitive underpinnings of religious fundamentalism - PubMed
https://pubmed.ncbi.nlm.nih.gov/28392301/

14 15 16 Confidence drives a neural confirmation bias | Nature Communications
https://www.nature.com/articles/s41467-020-16278-6?error=cookies_not_supported&code=ed54eaef-af15-42af-a5a8-ff1e913393d7

17 18 19 How the Brain Biases Beliefs - Neuroscience News
https://neurosciencenews.com/biases-beliefs-9701/

20 Attitudes, ideologies and self-organization: information load ...
https://scispace.com/papers/attitudes-ideologies-and-self-organization-information-load-1a6w6jvpgr?references_page=6

21 22 Education levels impact on belief in scientific misinformation and mistrust of COVID-19 preventive measures | University of Portsmouth
https://www.port.ac.uk/news-events-and-blogs/news/education-levels-impact-on-belief-in-scientific-misinformation-and-mistrust-of-covid-19-preventive-measures

23 24 39 Economic inequality and conspiracy theories - PubMed
https://pubmed.ncbi.nlm.nih.gov/35724596/

25 Multinational data show that conspiracy beliefs are associated with ...
https://onlinelibrary.wiley.com/doi/full/10.1002/ejsp.2888

26 27 28 The Power of Trump's Big Lie: Identity Fusion, Internalizing Misinformation, and Support for Trump | PS: Political Science & Politics | Cambridge Core
https://www.cambridge.org/core/journals/ps-political-science-and-politics/article/power-of-trumps-big-lie-identity-fusion-internalizing-misinformation-and-support-for-trump/AF2A0DBE08319E0E3944825E187EDBCC

29 30 Study shows that repeated statements are more often judged to be true, regardless of a person's age or prior knowledge | Vanderbilt University
https://news.vanderbilt.edu/2020/10/06/study-shows-that-repeated-statements-are-more-often-judged-to-be-true-regardless-of-a-persons-age-or-prior-knowledge/

31 Partisanship sways news consumers more than the truth, new study ...
https://news.stanford.edu/stories/2024/10/new-study-shows-that-partisanship-trumps-truth

32 Belief Updating in Subclinical and Clinical Delusions
https://academic.oup.com/schizbullopen/article/4/1/sgac074/6902048

33 34 35 Evaluating large language models in theory of mind tasks - PubMed
https://pubmed.ncbi.nlm.nih.gov/39471222/

36 37 38 Language Models Represent Beliefs of Self and Others
https://arxiv.org/html/2402.18496v3