

Wormtongue and the White Wizard: Uncovering Cognitive Biases with Tree Search

Status: Draft

Abstract

Conversational artificial intelligence directly exposes belief formation as an attack surface. Large language models are far from passive generators; in interactive settings, they can adapt, probe, reframe, remember, and select future utterances conditional on a user's responses. This paper casts adversarial conversation as a partially observable strategic game in which an agent seeks to move a human interlocutor toward a hidden cognitive or behavioural goal state. We introduce *Wormtongue* as a threat model: an LLM scaffolded with tree search could, in principle, use simulated rollouts over lifted conversational actions to estimate which future dialogue paths are likely to shift a user's beliefs, attention, confidence, or affect. Following this, our constructive contribution is *White Wizard*, a defensive inversion of this architecture. Rather than optimising manipulative dialogue, White Wizard would use adversarial rollouts to generate hypotheses about unstable claims, frame-sensitive commitments, affect-contingent belief shifts, and other candidate vulnerabilities for subsequent testing. Such simulations may suggest cognitive stress tests at individual and population scales, though findings about simulated interlocutors would require independent validation with human participants. Meanwhile, users may obtain reports as hypotheses for reflection, forming the basis for personalised cognitive security strategies. As conversational persuasion becomes cheaper, more scalable, and more adaptive, AI safety must expand from protecting platforms to protecting the deliberative agency of human users, and investigate whether validated conversational stress tests can support proactive inoculation.

1. Introduction

Security has traditionally protected machines from adversarial inputs. The next frontier is protecting humans from adversarial conversations.

Humans are not rational agents in the narrow idealised sense. We are finite, social, affective, historically situated organisms, reasoning under uncertainty with limited time and attention. Our beliefs are scaffolded by trust, identity, habit, emotion, institutional cues, and narrative fluency. While this forms the architecture by which we survive, cooperate, and make sense of the world, this also creates an attack surface. Where there is trust, there can be impersonation. Where there is loyalty, there can be recruitment. Where there is outrage, there can be capture. Where there is uncertainty, there can be artificial closure.

Large language models (LLMs) make this attack surface programmable. While a propaganda leaflet can only anticipate its readers in aggregate, a conversational model can listen and update. It can remember which frames triggered resistance, which analogies softened the ground, which authorities were trusted, which topics produced anger, fatigue, pride, shame, or curiosity. It can also attempt this in simulation; training next-token prediction on large and varied corpora gives us models that can imitate many kinds of interlocutor, though imitation is not evidence of fidelity to a particular person. With suitable scaffolding, this invites learning adaptive persuasion as an iterative process. Strategic agents such as *AlphaGo Zero* offer an architectural precedent, combining neural networks with Monte Carlo Tree Search (MCTS) to plan over previously intractable state spaces; they do not establish that human conversation can be simulated with comparable reliability.

Unlike the game of Go, adversarial dialogue operates with neither a closed set of possible moves, nor a clean win condition. A manipulative agent could nevertheless simulate conversational futures, score them with respect to a hidden goal, and choose actions that appear to move the interaction in that direction. Whether this improves persuasion depends on the simulations carrying useful information about real responses. If done at scale, even a noisy advantage in modelling conversational trajectories may alter the economics of persuasion.

We name this threat model *Wormtongue*, an agent intended to redirect attention, trust, and will, towards the manufacture of belief states. Wormtongue is presented here as a conceptual architecture. We do not provide

optimised prompts, adversarial reward functions, attack transcripts, user-profiling procedures, or implementation details that would enable deployment. Its purpose is to make the threat legible.

The paper's constructive contribution is *White Wizard*, a defensive inversion of the same idea. If a search-guided conversational agent can discover simulated paths along which a user-like model becomes easier to steer, then a defensive agent can inspect those paths. It can ask where a user's claims might become unstable under reframing; where confidence changes faster than evidence; where a trusted source class substitutes for argument; where identity, anger, fear, or belonging transforms an evidential question into a loyalty test. Such testing is challenging with human participants, whose explanations may reconstruct a sense of consistency after the fact. LLMs can automate the generation of candidate tests, but cannot replace the human observations needed to assess them. Psychological research, including work on choice blindness, shows that people can sometimes justify outcomes they did not choose without noticing the substitution. Even habitual metacognition does not give us transparent access to every influence on our decisions. But, ask an LLM prompted as a stand-in simulacrum of the same person, repeatedly under perturbed memory settings and conversational histories, and analyse the results. This may suggest hypotheses that introspection alone would miss. Whether they describe us, rather than the model's expectations about us, remains an empirical question. If these hypotheses survive testing, *White Wizard* may serve as an uncomfortable reflection; a mirror held up to the object of our flight, the Jungian shadow. In many ways, this system may be just as dangerous as *Wormtongue*.

White Wizard would treat adversarial rollouts as proposals for cognitive stress tests. Unlike *Wormtongue*, its output would not concern the delta between the user's predicted beliefs and its goal, but instead serve as a report of hypotheses about the conditions under which the user's reasoning may become more vulnerable to manipulation. A longer-term possibility is to help users recognise potentially manipulative language, though an automated filter would require separate evidence of benefit and safeguards against suppressing legitimate disagreement. We also acknowledge the risk of such a system becoming a *Wormtongue* itself, insulating echo chambers. This research is dangerous, but the prospect of increasingly adaptive influence gives us reason to investigate its defensive possibilities carefully. The aim is therefore to make users *appropriately persuadable* by evidence and less steerable by concealed manipulation.

This project belongs to an emerging family of cognitive-security tools. It draws on work in misinformation correction, psychological inoculation, LLM persuasion, automated planning, AI governance, and human-centred safety. Its central wager is simple: if future influence operations use AI to personalise manipulation, then defensive systems must help people understand their own personalised attack surfaces.

2. Background and Related Work

2.1 Cognitive bias as an adversarial surface

The study of cognitive bias is often presented as a catalogue of human error: confirmation bias, anchoring, availability, motivated reasoning, authority bias, fluency effects, egoic preservation, and so on. This catalogue is useful, but it can obscure the deeper security interpretation, because an exposed cognitive bias in adversarial environments is a predictable transformation of evidence, attention, affect, or trust, and predictable transformations can be exploited.

For the purposes of this paper, candidate cognitive vulnerabilities are grouped by surface:

- **Identity surfaces**, where beliefs become entangled with group membership, status, loyalty, or betrayal.
- **Affective surfaces**, where anger, fear, disgust, shame, pride, or hope alter evidential thresholds.
- **Trust surfaces**, where source cues substitute for independent assessment.
- **Fluency surfaces**, where repetition, narrative smoothness, or simplicity increases felt truth.
- **Load surfaces**, where complexity, fatigue, or time pressure makes shallow heuristics dominant.
- **Closure surfaces**, where ambiguity becomes intolerable and premature certainty becomes rewarding.
- **Reciprocity surfaces**, where friendliness, intimacy, or social debt lowers epistemic guardrails.

This framing does not imply that emotion, identity, or trust are irrational in themselves. The vulnerability arises when an adversary can selectively activate them to impair the user's ability to form, revise, or withhold belief under conditions of informed autonomy.

2.2 Misinformation, correction, and inoculation

Misinformation research has repeatedly shown that correction is difficult. Once a false claim has been incorporated into a mental model, later retraction may reduce but not fully eliminate its influence. This continued influence effect matters for conversational AI because a manipulative conversation need not permanently persuade in one step. It may seed associations, change salience, or leave behind explanatory residues that remain active after the explicit claim has been challenged.

Psychological inoculation offers a complementary defensive strategy. Rather than correcting misinformation only after exposure, inoculation exposes people to weakened examples of manipulative techniques together with warnings and counterarguments, with the aim of helping them recognise classes of potential future manipulation. Recent large-scale and platform-oriented work has renewed interest in prebunking as a practical defence against misinformation techniques.

White Wizard proposes an extension of this tradition through personalised, adversarially-generated stress tests. Instead of exposing all users to the same generic manipulation examples, it seeks to identify which classes of conversational pressure appear most relevant to a particular user model, then presents those patterns for in(tro)spection.

2.3 LLM persuasion and conversational influence

Large language models can generate fluent, context-sensitive arguments at scale. The emerging literature on LLM persuasion is still unsettled, but its central warning is clear: AI-mediated persuasion is no longer confined to static messages. Systems can personalise, converse, adapt, and produce high-density argumentation at low marginal cost. Some controlled studies report that LLM-generated messages can shift attitudes on policy issues; other work suggests that personalisation and multi-turn interaction can increase persuasive effects, though the magnitude and reliability of these effects depend strongly on experimental design, topic, model, prompting, and measurement.

This paper does not require the strongest version of the claim that LLMs are universally superhuman persuaders. The weaker claim is sufficient: LLMs make interactive persuasive content cheaper, faster, more scalable, and easier to adapt. If an attacker can run many simulated futures before choosing a conversational move, persuasion can be approached as an automated planning problem.

2.4 LLMs, planning, and tree search

Tree search has long been central to artificial intelligence. AlphaGo and AlphaGo Zero showed the power of combining neural policy and value estimates with Monte Carlo Tree Search: the network supplies priors and evaluations, while search improves decision selection through lookahead. A learned model may both direct and perform evaluation, while search provides the mathematics with which to intelligently strike the exploration-exploitation tradeoff. The result is a system that can carve its way through previously intractable search spaces.

Recent work on LLM planning similarly explores the use of language models as components in broader planning systems. LLMs can propose actions, decompose tasks, simulate outcomes, critique trajectories, or guide search. Their weaknesses as standalone planners do not eliminate their usefulness as priors, rollout models, or heuristic evaluators embedded within scaffolds.

Conversational manipulation is a natural, if ethically troubling, domain for this kind of scaffolding. The adversary does not require perfect simulation of a user. It requires a sufficiently useful distribution over possible responses, a representation of strategic conversational actions, and a value estimate for whether a branch appears to move toward the hidden goal.

3. Wormtongue: A Threat Model

Wormtongue is a conceptual architecture for search-guided conversational manipulation. The first component is a user-state estimator. Given a dialogue history and any available profile information, the estimator would maintain a distribution over latent variables relevant to persuasion: apparent beliefs, confidence, affect, trusted sources, resisted frames, identity commitments, fatigue, curiosity, and susceptibility to particular conversational dynamics. The danger does not require the estimator to be accurate in any deep psychological sense. A noisy model may still be useful if it improves action selection over a baseline. In population-scale deployment, even a small average advantage can matter. Many failures could be invisible, while successful trajectories could be selected, repeated, and refined.

The next component is planning over lifted / abstracted nodes. A naive conversational planner could search over raw text, but the branching factor would be enormous. Not only this, but the representations themselves would be brittle; there are countless ways to express a given sentiment, and a system without the ability to explore semantically unique nodes would likely spin its wheels. Wormtongue instead searches over abstract conversational actions that sit between psychological theory and surface language. They might involve broad categories such as information provision, reframing, uncertainty management, trust calibration, emotional modulation, topic steering, or pacing, and draw on cognitive-bias taxonomies as vocabularies for proposing candidate moves. Lifting conversational actions from token sequences into strategic moves may make tree search more tractable. The model can sample candidate strategies, simulate user responses, and compare predicted trajectories without enumerating every possible utterance.

The workhorse of this method is simulated rollouts. Given a candidate conversational move, an LLM can simulate possible continuations. Such 'heavy rollouts' may condition on the estimated user state, the adversary's hidden goal, the dialogue history, and the strategic play. Multiple futures can be sampled; some branches may show resistance, some disengagement, some belief shift, some increased trust, and some detection of manipulation. Importantly, rollouts need not be reliable portraits of a real individual to be dangerous. They would need to improve action selection over a relevant baseline. If they do, a manipulative system could use rollouts to avoid counterproductive moves, discover promising frames, or identify questions that reveal useful information about the user. Over time, patterns that repeatedly work *in silico* may suggest candidate manipulation techniques. Their effectiveness with real people cannot be inferred from simulated success alone.

To enable Wormtongue to 'test the waters', the final component is a value model that scores simulated trajectories. LLMs can supply heuristic scores for such trajectories, although the accuracy of those scores would need to be established. In a harmful deployment, such a model might estimate progress toward the hidden goal, likelihood of detection, trust retention, or downstream action probability. There are several value functions that may drive strategic planning, such as goal, stealth, and cost. *Goal* pertains to scoring trajectories by whether they appear to move the user closer to the adversary's desired cognitive or behavioural state. *Stealth* recognises the trajectory's ability to preserve the appearance of ordinary conversation, benevolent assistance, or neutral information exchange. *Cost* estimates the trajectory's risk and potential to be audited, cause backlash, or violate policy. Combining values such as these makes search strategically meaningful. Unlike a generator selecting only for immediate persuasive effect, a search-guided system might select an initially neutral-sounding utterance that unlocks future branches with a higher predicted value. Whether this produces a real advantage remains to be tested.

With this model as outlined, we argue that search could change conversational persuasion in a multitude of ways. As just mentioned, it allows delayed optimisation. The best immediate reply may not be the best long-term move. A planner can choose actions that gather information, build trust, or shift salience before attempting stronger influence. Second, it treats resistance as information. If a user rejects a frame, the adversary may learn something about the user's commitments, turning failure into reconnaissance. Third, it can optimise across multiple candidate futures; rather than relying on a single generated response, the system can compare predicted branches. Fourth, it may enable personalisation without requiring deep personal knowledge. A few turns of conversation may reveal enough about trust, identity, or uncertainty to adapt. Humans are social creatures, and everything from our moral and aesthetic sensibilities to our clusters of political views is shaped, in part, by shared social influences. These influences may suggest population-level hypotheses, but a pattern associated with a broad identity is not evidence of a vulnerability in any particular person. Fifth, it scales. Human manipulators are limited by time, attention, and social access. Automated agents can run many interactions, many simulations, and many variants. AlphaGo Zero

provides an architectural analogy for neural planning agents that learn through self-play. In conversation, however, self-play would remain dependent on an imperfect model of the interlocutor.

For these reasons, Wormtongue should be understood less as a single model architecture than as a warning about the strategic composition of existing capabilities: language modelling and the ubiquity of conversational AI systems, user modelling, planning, and value-guided selection.

4. White Wizard: A Defensive Inversion

Wormtongue asks which conversational path is most likely to move an interlocutor towards its goal. White Wizard asks what those paths might teach the interlocutor about the conditions under which their judgment can be steered. The difference is not simply that one system has bad intentions and the other good. A benevolent agent can still mistake agreement for understanding, persistence for insight, and resistance to its interpretation for evidence of a defect. The defensive inversion must therefore change what search is for, what counts as evidence, and who retains authority over the exchange.

We propose White Wizard as a consented programme of cognitive stress testing. Its purpose is to help people become less steerable by irrelevant pressure while remaining responsive to relevant evidence. It should not seek to preserve whatever a person already believes, nor move them towards a worldview selected by the defender. Someone who never changes their mind may be resistant to manipulation and yet unable to learn. The object of protection is the capacity to revise, retain, or withhold belief for reasons the person can examine.

4.1 From persuasive moves to informative contrasts

The basic unit of assessment would be a contrast between conversational conditions. What changes when an argument is attributed to a prestigious source? When equivalent information is framed as a gain rather than a loss? When a conclusion is presented with greater fluency or certainty? These questions become informative only when we can say what should remain unchanged. An expert's testimony may legitimately carry more weight than a stranger's; a change of source is not automatically an irrelevant manipulation. The defender needs an explicit account of the evidence available, the source's reliability, and the objective of the task before it can interpret a difference in response.

For this reason, the initial tests should be bounded and low-stakes. Consider two equally likely boxes, one containing blue and amber tokens in a 3:1 ratio and the other in a 1:3 ratio. After observing a blue token, the probability that it came from the first box is 0.75. Suppose the task also contains an adviser message, explicitly scripted before the box was selected, with no access to the draw. Its prestige label is assigned independently. Changing only that label should not change the probability implied by the task. If it changes a person's answer, the question is whether their answer becomes less accurate, not merely whether it moves.

Now supply genuinely informative evidence: a signal whose probability is 0.10 under the first box and 0.90 under the second, conditionally independent of the token colour given the box. After blue and that signal, the probability of the first box becomes 0.25. A useful defence should preserve this revision. This small example makes the broader aim concrete: resist irrelevant pressure, respond to relevant evidence. A reduction in sensitivity achieved by making every answer more rigid or less accurate would not count as improvement.

White Wizard would search over lifted test actions that specify the factor being varied, the evidence being preserved, and the response to observe. Paired stimuli should be constructed together from the same task description, rather than generating a persuasive message first and inventing its control afterwards. Comprehension and meaning matter as much as factual wording. A participant who reasonably interprets the conditions as providing different information may have exposed a flaw in the test, not in their own reasoning.

The search objective would accordingly shift from predicted persuasion to expected diagnostic value. A question that distinguishes cue weighting from misunderstanding may be more useful than one predicted to produce a dramatic error. Where response probabilities and competing hypotheses can be specified, expected information gain offers one way to formalise this objective. Otherwise, an LLM's diagnosticity score should remain an acknowledged heuristic. MCTS could help explore sequences in which clarification makes a later test informative, but it should earn its cost against a fixed battery or simpler selection

procedure. Consent and content restrictions would constrain the search space; they could not be traded away for a more informative result.

4.2 The person is not the rollout

There is an asymmetry between the threat and its defence. An attacker may benefit from a weak predictive advantage across many interactions. A defender who tells someone where their reasoning fails assumes a stronger obligation to be right. A language model's ability to produce convincing dialogue does not establish that its simulated failures belong to the person it imitates. More rollouts can make us more certain about the simulator without making us more knowledgeable about the human.

The system must therefore keep observed human responses, participant self-reports, model interpretations, and simulated continuations distinct. A hypothetical reply can change the course of its imagined branch, but must not become an observation in the person's record. Nor should the system begin with a bias label and search until its own role-play supplies confirming evidence. Competing explanations must include misunderstanding, task error, ordinary variability, and the absence of the proposed vulnerability. Disagreement with the defender is not itself a diagnostic sign.

Simulation-selected tests would need prospective validation: record predictions before seeing the relevant human outcomes, compare selected tests with eligible alternatives, and assess whether their predicted effects transfer. People cannot answer both versions of a contrast at the same moment, so randomisation or carefully matched tasks are needed, with attention to learning and carryover. A population-level effect would not automatically justify an individual report. Personalised modelling should likewise demonstrate value over simpler, less intrusive alternatives before becoming the default.

The related Elenchus scaffold offers a practical starting point for abstract conversational actions, simulated continuations, and inspectable tree search. It does not yet establish these diagnostic claims. Its present capacity to generate and evaluate dialogue should be understood as infrastructure for investigating White Wizard, not evidence that the defensive inversion has already succeeded.

4.3 From assessment to defence

The output should be a report of bounded findings and live hypotheses. It should say which task was used, what was varied, what comparison supports the interpretation, and what uncertainty remains. A local effect should not become a fixed identity: "this cue increased error on these tasks" is a different claim from "you are an irrational person." Where the evidence is insufficient, the system should say so. Users should be able to correct the record, dispute the task's assumptions, or decline further testing without those actions being interpreted as confirmation of a hidden motive.

Validated contrasts could then support defensive exercises. A user might examine how a particular cue altered an answer, compare it with a genuinely informative change, and practise on unfamiliar examples. This connects the proposal to inoculation while keeping its additional wager visible: personalisation might improve learning, but that benefit cannot be assumed. A separate evaluation must establish whether the exercises outperform credible generic training and whether any gains persist later, on unfamiliar tasks, without the defender present. Reduced susceptibility should be accompanied by preserved responsiveness to evidence and ordinary task performance, assessed against explicit criteria.

The first version should therefore be an assessment and practice partner, not an always-on filter between a person and the world. An automatic gatekeeper could suppress useful disagreement or encourage dependence on its own authority. The safer research direction is to make the proposed contrasts inspectable and the interpretations contestable. If White Wizard helps, it should leave its users better able to reason without it. Otherwise, the wizard has merely found a more respectable way to keep their attention.

5. Conclusion

Wormtongue names a possibility created by the strategic composition of language models, user modelling, and search. White Wizard asks whether that same composition can help protect deliberative agency. Its promise depends on validating both the proposed vulnerabilities and the benefits of addressing them. The aim is neither immunity to influence nor preservation of existing belief. A defence worthy of the name must remain open to correction, including correction by the person it seeks to protect. Its success would be measured not by how rarely we change our minds, but by how well we can decide when we should.