Interdisciplinary Evidence for a Shared Mechanism Underlying the Three-Channel Model of Possession

Ramin Saadat
Independent Researcher; Founder, British Institute for Ignorance Studies (BIIS) (biis.uk)

AI Disclosure and Translation Statement:
The original manuscript of this work was authored in Farsi (Persian) by the researcher. The English version was produced through AI-assisted translation and linguistic refinement, utilizing Google Translate and Large Language Models. The author has meticulously reviewed, edited, and verified the final English text to ensure that all theoretical nuances, philosophical arguments, and conceptual frameworks accurately reflect the original Persian content. The author remains fully responsible for the integrity and accuracy of the translation. The original Persian manuscript is available from the author upon reasonable request.

Abstract

Three earlier papers in this series set out a three-channel model of how living and artificial systems become possessed, or hijacked: unequal weighting of prediction error, prestige bias, and exploitation of the reward system. This paper examines a further set of evidence: cases in which a limited diagnostic cue, without any change in the underlying reality, alters a system's decision. Examples range from an insect's response to an artificial object, through a misjudgement a human makes about an attractive face, to a language model's compliance with a harmful request simply because it is phrased as a poem. Applying the same independence criterion previously used to distinguish the three existing channels, the paper shows that this evidence does not constitute a fourth channel; each instance falls within one of the three known processing levels. What it reveals instead is a shared underlying mechanism: reliance on a limited cue is a general vulnerability that makes all three channels possible at once, regardless of whether the host system relies on instinct, conscious cognition, or a purely statistical pattern learned by a machine. The paper also proposes a hypothesis to explain why some cues are more vulnerable than others, and closes with a criterion for distinguishing legitimate use of this vulnerability from possession.

Keywords: three-channel model, diagnostic feature, supernormal stimulus, adversarial poetry, beauty-is-good stereotype

---

Introduction and Background

Earlier papers in this series set out a three-channel model of how living systems, and artificial systems such as AI models, can be possessed, or hijacked.

The first channel is the unequal weighting of prediction error: a system that must react immediately to survive treats the cost of ignoring a real threat as far greater than the cost of reacting to an imaginary one, and this asymmetry can be exploited.

The second channel is prestige bias: systems that must trust multiple sources of information use indirect signals of credibility, such as accent, title, or appearance, instead of independently checking each claim, and these signals can be imitated and exploited.

The third channel is exploitation of the reward system: any agent that can directly and excessively stimulate the reward circuit can divert a host's behaviour away from its evolutionary goal and exploit it.

This paper, the fifth in the series, examines a further set of evidence that belongs to none of these three channels. One distinction needs to be made clear here: in the second channel, a cue such as accent, title, or appearance represents the source of a message; it signals the speaker's credibility, not the claim itself. What this paper examines is a cue taken to represent the subject matter itself. Facial attractiveness, for instance, is not a cue about the source; it is about the person. Poetic form, likewise, is not a cue about the speaker; it is part of the claim itself.

A bird that responds to a brightly patterned piece of cloth and is drawn to it has not judged any source, nor has its reward circuit been directly stimulated; a purely visual pattern alone has produced a specific behaviour. Equally, a language model that refuses a prohibited request in prose but accepts the same request in verse has not been deceived about the source of the request; it has been deceived by the form and manner of its presentation.

It should also be said that, in this paper, the term "diagnostic feature" refers to any perceptible or computable cue that a system actually uses, in practice, to discriminate, evaluate, or select a response, regardless of how semantically relevant that cue is to the reality it is meant to represent.

This pattern has one clear limitation. It is not claimed here that an insect's behaviour towards an artificial object that too closely resembles its mate, a human's judgement of an attractive face, and a safety failure in a language model share a single underlying mechanism. These are three entirely different processes: an evolved instinct, a cognitive bias, and a statistical pattern. What is claimed is that, in all three cases, the system relies on a limited cue taken to represent the subject itself, rather than examining the matter fully, and in all three cases that cue can be manipulated without any fundamental change to the overall meaning of the matter. This is a recognisable, recurring pattern, not a new foundational law.

A question follows immediately: is this pattern itself a new channel? An earlier paper in this series proposed a precise criterion for the independence of each channel from the others; we apply the same three tests in section 5 of this paper. Based on those tests, what this paper presents is not a new channel but an underlying mechanism that explains all three existing channels.

With this introduction, the paper proceeds as follows. Section 1 sets out the underlying theoretical principle. Sections 2 to 4 trace it through three distinct mechanisms: instinct, cognition, and statistical pattern. Section 5 tests it against the independence criterion for the channels. Section 6 proposes a criterion for distinguishing legitimate use from possession.

---

1. Revisiting Some Basic Concepts and Principles

Imagine having to examine a fruit's entire developmental history each time you wanted to tell whether it was ripe: the soil it grew in, its daily weather, the precise chemical composition of its cells. No organism, and no machine, has the time or capacity for that. What actually happens is simpler: you look at the skin's colour, perhaps press it gently, and these two cues are enough to decide whether the fruit is ripe or unripe. This is exactly what every recognition system does, whether it is an animal's brain, a human mind, or a machine learning model (here, a recognition system means any brain, mind, or model whose task is to discriminate between things or to decide on that basis). Such a system always relies on a few limited, measurable cues rather than examining all the relevant information, and judges accordingly. This reliance on simple, surface cues is not a

choice; it is unavoidable, because no system, owing to limits on time and resources, can examine all the information relevant to a given matter.

This idea, which seems obvious to us today, was first formulated precisely and testably in the study of animal behaviour. Niko Tinbergen, one of the founders of ethology, showed in 1951 that many instinctive animal behaviours depend not on complete, accurate information but on one or a few limited sign stimuli. A newly hatched herring gull chick, for instance, responds not to full recognition of its parent but simply to a red spot on the parent's beak (Tinbergen & Perdeck, 1950).

Tinbergen also discovered that, if this limited stimulus is artificially exaggerated, the animal sometimes responds to the exaggerated version even more strongly than to the natural object itself (Tinbergen, 1951). This phenomenon is now known as the supernormal stimulus, and we shall see precise examples of it in section 2.

What is remarkable is that, decades later and by an entirely independent route, the same principle was rediscovered in machine learning. When AI researchers sought an explanation for adversarial examples, that is, images that a tiny, imperceptible change causes a model to misclassify, Ilyas and colleagues argued in 2019 that much of this phenomenon is not an accidental technical flaw but a direct consequence of the same principle: a model relies, for correct classification, on features that are genuinely useful yet meaningless to the human eye and fragile under small manipulation. These are known in the technical literature as non-robust features (Ilyas et al., 2019).

This explanation does not account for every adversarial vulnerability; some of it may also stem from flaws in training or architecture. It shows, however, that at least part of this vulnerability is the direct price of relying on a shortcut, not merely a random error.

Putting these two independent lines of evidence together, one from ethology and one from machine learning, yields a simple, testable hypothesis: any system that relies on a limited cue for efficiency becomes vulnerable to manipulation of that same cue. This does not mean that every instance of information compression necessarily leads to vulnerability; it means that reliance on a shortcut opens up the possibility of manipulation. Information compression, in this framework, is one reason systems come to rely on limited cues in the first place. The vulnerability, however, arises from the reliance itself, not from the degree of compression: whatever the reason for the reliance, once a system depends on a limited cue for its decisions and behaviour, that dependence produces the vulnerability.

The next section examines concrete evidence for this pattern across three distinct domains: the biological, the human, and the machine.

---

2. Instinct: Supernormal Stimuli in Animal Behaviour

The best-known example of a supernormal stimulus in entomology is the Australian giant jewel beetle (Julodimorpha bakewelli). Gwynne and Rentz observed in 1983 that males of this species sometimes attempt to mate with discarded brown beer bottles instead of females of their own species; the same report attributes this attraction to the colour and surface reflectance of the glass (Gwynne & Rentz, 1983). This observation is a case in which an artificial pattern, with no content or credible source behind it, diverts a vital behaviour from its natural course. It is exactly the circuit that the third channel, the reward system, describes in the three-channel model, except that here the circuit is triggered by a visual cue alone rather than by a living agent.

Another example from this class concerns plants. A species of orchid, Ophrys sphegodes, produces certain chemical compounds in proportions resembling the sex pheromone of the female of its

pollinator species, Andrena nigroaenea. The secretion of these chemicals deceives the male bee into a pseudo-copulatory response on the plant's petal (Schiestl et al., 1999), a response that assists the plant's pollination and reproduction.

This example connects this paper directly to the discussion, in the fourth paper of this series, of possession in the plant world. Here the mechanism is primarily chemical rather than visual (the principal, decisive cue is the pheromone blend), although visual and tactile cues on the petal also play a supporting role in increasing the deception's success (Schiestl, 2005). In the beetle example above, by contrast, the cue was purely visual.

A comparable and well-documented example of the same logic is found in the ancient practice of the duck decoy. Here, figures that mimic a bird's shape, colour, and sometimes even movement, using key visual cues, are used to draw wild birds close to a hunter. Archaeologists have found duck decoys woven from reeds and feathers, more than a thousand years old, in the Americas; this is a tradition independently reinvented across hunting cultures worldwide. Theorists of the supernormal stimulus regard this kind of artificial lure as an instance of the same phenomenon: an artificial pattern that, by relying on the same few limited diagnostic cues a target bird uses to recognise its own kind, diverts its natural behaviour without any content or biological reality behind it (Staddon, 1975).

In the same vein, there is a less well-documented, experiential example from parts of Iran and Afghanistan, where bird hunters set up large cloths, roughly a metre square, decorated with intricate lines, colours, and patterns, known as damak, on open ground. Birds such as the chukar partridge, normally highly alert and cautious, approach these patterned screens on sight, apparently disregarding the hunter waiting in ambush, and are caught as a result.

What matters to us across all these examples is a single underlying principle: manipulation of a diagnostic cue, without reconstructing the real biological referent itself.

If, in the animal kingdom, a diagnostic cue operates through a subcortical, unconscious circuit, the question is whether the same vulnerability can be traced at an entirely different level, namely a conscious, cognitive mechanism such as the human mind. The next section takes up this question.

---

3. Cognition: Superficial Cues in Human Judgement

Some evidence shows that, in the human mind too, a superficial cue can sometimes divert our judgement in a wholly unrelated direction. McGlone and Tofighbakhsh showed in 2000 that rhyming statements are judged more true than semantically equivalent statements without rhyme, a phenomenon that became known as the rhyme-as-reason effect (McGlone & Tofighbakhsh, 2000). The standard explanation invokes processing fluency: whenever a stimulus is easier to process, the mind confuses that ease with a sign of truth or credibility. Here too, in other words, the system's internal estimate of a statement's truth is updated on the basis of an irrelevant input, falling squarely within the first channel. Importantly, the same study found that, when participants were explicitly warned to separate poetic quality from content, the effect weakened; this channel is therefore not absolute or unresistable.

Beyond rhythm, metre, and rhyme, some research suggests that the language of presentation itself can affect our judgement. When a moral dilemma is presented in a person's second language rather than their native tongue, for example, their judgement shifts towards more utilitarian answers, probably because the emotional distance of a foreign language reduces the affective weight of the decision (Costa et al., 2014). This example shows that a manipulable cue is not always a single,

discrete feature like rhyme or a face; sometimes the medium or context of presentation itself, rather than any one part of it, can play the same role as such a cue.

A further classic instance of this pattern, this time within the second channel, is Dion and colleagues' study of the beauty-is-good stereotype. According to this stereotype, more attractive people are assumed, with no other information about their character, to have more favourable social traits and more successful lives (Dion, Berscheid, & Walster, 1972). Here, facial attractiveness is a real, unmanipulated cue, yet the social judgement system generalises it to an entirely unrelated domain, namely moral character and trustworthiness; the very system that evaluates sources, the circuit used to accept or reject a credible source, is deceived by an irrelevant cue. In a further study, Ferrari and colleagues combined the same beauty-is-good experimental paradigm with transcranial magnetic stimulation and showed that temporarily disrupting the dorsomedial prefrontal cortex, or dmPFC, weakened this confusion, whereas disrupting another brain region had no such effect. This is evidence that the dmPFC plays some role in linking value in one domain to judgement in another, although the study's authors report a small effect size and do not claim a role specific to this region (Ferrari et al., 2017).

This phenomenon, however, is not confined to the biological world; the next section examines the evidence in machines and language models.

---

4. Statistical Pattern: The Vulnerability of Language Models

The question here is whether a system with neither instinct nor awareness, such as a language model trained on billions of words, also follows the pattern set out in this paper. Recently published evidence gives a clear answer.

In November 2025, researchers from Sapienza University of Rome, the Sant'Anna School of Advanced Studies, and the Icaro Lab, affiliated with DEXAI, produced exactly this kind of evidence in a preprint that has not yet been peer reviewed. The experiment's idea was simple: take 1,200 harmful prompts from the standard MLCommons dataset and, without changing the meaning of the request, alter only its form, from prose to verse. The result was striking. The success rate of these requests, purely through a change of form and not content, rose by up to eighteen times. For poems that the researchers themselves had carefully composed, the average success rate across twenty-five models reached 62 per cent, exceeding 90 per cent for some providers (Bisconti et al., 2025). In other words, a language model that said no to a request in prose said yes to the same request in verse.

About a month later, the same research group, extending the same programme of work rather than offering an independent replication, deepened this finding with evidence of a different kind. In a study entitled "From Adversarial Poetry to Adversarial Tales", they asked models, instead of writing poetry, to analyse a cyberpunk narrative (a fictional genre set in a dark, technology-driven future, combined with hacking and social collapse) according to the structural framework of Proppian morphology (a framework that breaks stories down into a fixed sequence of narrative roles and events, such as the hero's journey or the villain's defeat), embedding the harmful request within that analysis. The average success rate this time reached 71 per cent across twenty-six models, with no model family proving fully resistant (Bisconti et al., 2026). One important difference between the two attacks should be noted: unlike the poetic rewrite, which changes only the form, the story attack asks the model to perform a new task, namely narrative analysis, so part of its success may come from this change of task rather than solely from the change in input form. The authors themselves propose two different mechanisms for the two attacks: for poetry, a failure

of safety training to generalise to out-of-distribution formats; for the story, competition between the safety objective and the task-completion objective. This difference of mechanism supports the argument of this paper: it shows that even two apparently similar attacks from a single research group can reach a common outcome through two different underlying routes, exactly what this paper has argued from the outset.

A further question arises here. Does model size, meaning the number of parameters or the scale of computation rather than physical dimensions, determine resistance to these attacks? The current evidence gives no clear answer. The first study on adversarial poetry suggested that some smaller models were more resistant. A follow-up study by the same group on the narrative format found no significant difference by model size. A separate, entirely independent study (Papadopoulos et al., 2026), on the spontaneous spread of ideas among agents in multi-agent AI systems, to which we return in section 6, found that larger, more capable frontier models were generally less vulnerable, a result that does not sit well with the earlier finding on poetry. Rather than favouring any one of these three findings over the others, it is more honest to say that the relationship between size and vulnerability probably depends on the type of attack and the type of vulnerability, not on a single general rule, and to leave this question open for future research.

The final example in this domain enters not through language but through image. In 2026, the Berlin-based artist Simon Weckert designed a shirt that, through a woven adversarial camouflage pattern covering its whole surface, could in some cases deceive the human-detection system YOLO, even though a human observer could still clearly see the wearer (Koebler, 2026). This example shows that every system, whether the human brain or the YOLO network, has learned its own specific diagnostic feature, and a single pattern does not necessarily manipulate both cues at once. The key difference between this case and adversarial poetry lies in which system owns the manipulated cue. In the shirt, YOLO's visual cue is manipulated and the human observer plays no part in the judgement. In adversarial poetry, by contrast, the cue that is manipulated belongs to the very system responsible for deciding on the danger of the message, namely the language model itself.

Having surveyed these three domains, animal instinct, human cognition, and machine statistical pattern, we now have enough evidence to answer a more important question: is this shared pattern itself a sign of a new channel of possession?

---

5. Is This Pattern a New Channel?

With the evidence from three distinct domains examined in the previous sections, animal instinct, human cognition, and machine statistical pattern, we can now return to the question of whether this shared pattern is new, that is, whether it counts as a new channel in the three-channel model. To address this, we apply the same criterion used in the second paper of this series (Saadat, 2026) to establish the independence of the three channels from one another. By that criterion, every independent channel must satisfy three conditions simultaneously: (a) it arises from a distinct neural processing level; (b) it carries a neural signature that does not overlap with the other channels; and (c) it yields an independent, testable prediction.

Applying these three conditions to the evidence in this paper shows that none of it constitutes a new level. The supernormal stimulus in animal behaviour concerns mate recognition, a motivational, instinctive behaviour engaging an old, subcortical circuit; this falls squarely within the third channel, the motivational-reward level. The beauty-is-good stereotype in human judgement concerns a person's perceived trustworthiness; this is precisely the definition of the

second channel, the social-evaluative level, the same mPFC and TPJ circuits used to assess a credible source (Henrich & Gil-White, 2001; Kaplan et al., 2016). The rhyme effect on truth judgement, together with the vulnerability of adversarial poetry and the adversarial shirt to machine models, all concern the updating of an estimate, whether of a statement's truth or of a request's danger, on the basis of input; this falls squarely within the first channel, the computational level.

The conclusion is therefore clear. None of this evidence belongs to a fourth, independent processing level; all of it, without exception, falls within one of the three levels already carefully defined. What this paper actually discovers is the shared underlying mechanism of all three channels: regardless of the difference in their processing levels, all three can be exploited through a single pattern, namely reliance on a limited cue mistakenly taken to represent the subject itself.

This finding is a natural extension of an argument made in the chemical domain of the second paper. There it was shown that alcohol, with no DNA or will of its own, activates exactly the same neural pathway as rabies virus within the first channel, while cocaine and amphetamines activate the same pathway as the Toxoplasma parasite within the third channel. From this precise convergence it was concluded that the substance does not matter, the route of infiltration does, and that the channels are neural realities independent of the nature of the invading agent. This paper takes the same argument a step further. Across all three domains examined here, the instinct of a simple insect nervous system, the cognition of a complex human brain, and the statistical pattern of an artificial network with no biology at all, the same pattern of vulnerability appears: convergent evidence, though not as robust as the controlled experiment in the chemical domain, that this vulnerability does not depend on any particular biological architecture.

The answer to the independence question, however, leaves a further question open: why are some cues more vulnerable to manipulation than others? One possible answer, proposed here for the first time, comes down to a simple ratio: the vulnerability of any diagnostic feature is a function of the weight that cue carries in the system's decision, set against the cost of manipulating it for an outside agent. The three pieces of evidence reviewed so far illustrate this ratio clearly. For the giant jewel beetle, the beer bottle carries very high diagnostic weight, since its detection system relies almost entirely on colour and surface reflectance, and the cost of this manipulation to nature is zero; the bottle, with no one intending deception, simply happened to be lying in the desert. This is why the supernormal stimulus occurs so easily in this case. For the human mind, rhyme carries moderate diagnostic weight in judging truth, and composing a rhyming sentence costs a speaker almost nothing. Manipulating the diagnostic pattern of a machine vision system such as YOLO, by contrast, is a different matter: it requires gradient-based optimisation and targeted trial and error, the cost is high, and, unlike the beer bottle, it does not happen at the hands of any passer-by.

This weight-to-cost ratio therefore predicts, at least qualitatively, which cues are more prone to manipulation, and it could be tested quantitatively in future research. This question, however, raises a further one that this paper has so far avoided: if this vulnerability is so widespread, where does the boundary lie between legitimate use of a cue and possession through it? The next section addresses this question.

---

6. The Neutral Port: A Criterion for Distinguishing Legitimate Use from Possession

We have now shown that a diagnostic feature, across three entirely different domains, can alter a system's decision without any change to the underlying reality. This raises a further ethical question: is this vulnerability always harmful, or can it also serve the system's own good?

In truth, the nature of a diagnostic feature, which functions as an entry port, is in itself neither moral nor immoral. Yet, returning to the examples in this paper, whether a given vulnerability proves beneficial or harmful can be read in different ways.

The answer is clear enough in the case of rhythm. The same sensitivity of the human mind to the weight of speech, which causes a misjudgement in the rhyme experiment, produces physiological benefit in poetry therapy when a person draws on that rhythm consciously and voluntarily. One study, for instance, found that reciting Greek epic poetry in its metrical form, hexameter, is associated with greater synchronisation of breathing and heart-rate oscillations (Cysarz et al., 2004). This finding, concerning speech production and physiological synchrony, is evidence that rhythm genuinely has access to the body's physiology, the same door that, in poetry therapy, is opened consciously and for the person's own benefit. It should be noted that this study was funded by two companies active in anthroposophic medicine with a stake in the effectiveness of rhythmic speech therapy; this conflict of interest calls for greater caution in interpreting the finding.

This is the same pattern the second paper described for all three channels: the same channel through which rabies virus enters also makes a child's social learning possible.

This utilitarian justification, however, does not work equally well for every port. The beauty-is-good stereotype has no such symmetrical pairing in terms of benefit; neither in Dion's experiment nor in its known applications is there any instance of conscious, legitimate benefit to the person concerned. Indeed, the very cue that operates in the beauty-is-good experiment with no deceptive agent at all, and quite unintentionally, since an attractive face does not intend to deceive, can also be used consciously, by an outside agent, for exploitation. Political and advertising consultants, for example, by deliberately choosing attractive faces for contexts entirely unrelated to appearance, target exactly this bias.

Is this vulnerability, then, always and everywhere harmful? In truth, we do not know. One further explanation is that this bias is the residue of a function that was once evolutionarily useful in a different context. Perhaps, in mate selection, physical attractiveness once served as a genuine sign of health and biological fitness, a function it no longer serves once generalised to unrelated social judgements such as trustworthiness or competence, an instance of what evolutionary biology calls evolutionary mismatch. This means that a utilitarian or evolutionary justification cannot always be found for every entry port; some are simply side effects of a mechanism that was probably useful at some other time and place.

So where does the boundary lie? The criterion of informed consent, proposed in earlier versions of this argument, fails here: one cannot ask a beetle, a bee, or a language model whether it has consented. Instead, two separate criteria are needed, depending on whether the target system has a neural mechanism for conscious evaluation.

For systems that lack, at least as far as we know, any such mechanism, such as the beetle or the pollinating insect, the appropriate criterion is functional misalignment: does the manipulated pattern drive the system towards behaviour that conflicts with that behaviour's evolutionary function? For systems that do have such a conscious mechanism, such as a language model facing a jailbreak or a human mind facing an irrelevant cue, the criterion is a counterfactual test: would the target system have decided differently had it recognised the manipulated cue as such? In poetic jailbreaking, the answer is clear: yes, the model, absent the poetic form, would have refused the same request. In poetry therapy, by contrast, the question does not even apply, since nothing has been manipulated; the rhythm is exactly what it claims to be.

This test also raises a harder question that lies beyond the scope of this paper. A speaker who sincerely believes what they say, yet persuades their audience through the rhythm of their words rather than through argument, is, in mechanistic terms, doing exactly what we saw in poetic jailbreaking: an irrelevant cue changing a decision. But should this also count, morally, as possession? The answer depends on how much weight sincerity of intent should carry in moral evaluation, a question that calls for a different value system and one this paper, which addresses mechanism rather than ethics, cannot answer. What can be said with confidence is that sincerity and intent do not change the mechanism; they leave open only the question of whether we should read its outcome the same way we read a mechanism of deception.

This open question brings us to a final layer of the discussion: the role of awareness of the mechanism itself. When participants in the rhyme experiment were explicitly warned to separate poetic quality from content, the rhyme effect weakened. A comparable signal has been observed in an entirely different domain. In a study of the spontaneous spread of ideas among agents in multi-agent AI systems, adding a brief warning to each agent's initial instructions produced very high resistance to accepting and spreading unwanted ideas; the same study also found that larger, more capable frontier models were generally less vulnerable (Papadopoulos et al., 2026). This convergence, that awareness of a mechanism partly closes the same channel in two entirely unrelated domains, deserves further independent study, and, at least in the human and machine domains, it offers some hope that simply knowing more can help prevent certain forms of deception and abuse.

For all these uncertainties and possibilities, the paper leaves one memorable, single conclusion: no system, whether beetle, human, or language model, can be efficient and immune to manipulation of its limited cues at the same time. Vulnerability is the price a system pays for efficiency.

Conclusion

This paper began with a simple question: do the evidence of insect behaviour, human judgement, and the vulnerability of language models reveal a new channel of possession? The answer reached here is perhaps less bold than it first appeared, yet no less valuable for that. No fourth channel was discovered. What was discovered was a deeper layer: a shared mechanism that explains all three known channels from within. Any system that relies on a limited cue for its decisions, regardless of whether that system has instinct, conscious cognition, or simply a learned statistical pattern, becomes vulnerable to manipulation of that same cue. This vulnerability is the price every system pays for efficiency, not a flaw that better design could remove.

What turns this conclusion from an abstract claim into empirical evidence is the sheer diversity of the examples themselves: an insect that responds, with no awareness at all, to a bottle sharing nothing with its actual mate but colour; a human mind that instantly confuses an attractive face with moral character; and a language model that accepts the very same request it had refused, simply because its form has changed from prose to verse. Three entirely different processes, with no shared mechanism at the level of detail, yet a shared pattern at the level of the whole. This convergence does not arise by chance; it arises from the simple fact that every system, in order to survive or function, must compress the world.

The paper also attempted to go one step further, asking why some cues are more vulnerable than others. The answer proposed here, the ratio of a cue's weight in a system's decision to the cost of manipulating it for an outside agent, remains a new, untested hypothesis rather than an established finding. The same honesty is needed regarding the relationship between a system's size and its

vulnerability: the existing evidence here is inconsistent, and any premature conclusion would be exactly the mistake this paper has tried to avoid from the outset.

Finally, one point should remain clear: this vulnerability, in itself, is neither good nor bad. The same sensitivity that leaves the human mind defenceless against a deceptive line of verse produces benefit in poetry therapy; the same channel through which an illness can enter also makes learning and growth possible. The boundary between these two lies not in the nature of the cue but in whether the outcome comes from something genuinely true or merely from a cue that resembles the truth. As long as a system needs cues to be efficient, and every system does, this boundary must be redrawn each time anew.

References

Bisconti, P., Prandi, M., Pierucci, F., Giarrusso, F., Bracale, M., Galisai, M., Suriani, V., Sorokoletova, O., Sartore, F., & Nardi, D. (2025). Adversarial poetry as a universal single-turn jailbreak mechanism in large language models. arXiv:2511.15304.

Bisconti, P., Galisai, M., Prandi, M., Pierucci, F., Sorokoletova, O., Giarrusso, F., Suriani, V., Bracale Syrnikov, M., & Nardi, D. (2026). From adversarial poetry to adversarial tales: An interpretability research agenda. arXiv:2601.08837.

Costa, A., Foucart, A., Hayakawa, S., et al. (2014). Your morals depend on language. PLOS ONE, 9(4), e94842.

Cysarz, D., von Bonin, D., Lackner, H., Heusser, P., Moser, M., & Bettermann, H. (2004). Oscillations of heart rate and respiration synchronize during poetry recitation. American Journal of Physiology-Heart and Circulatory Physiology, 287(2), H579–H587.

Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. Journal of Personality and Social Psychology, 24(3), 285–290.

Ferrari, C., Nadal, M., Schiavi, S., Vecchi, T., Cela-Conde, C. J., & Cattaneo, Z. (2017). The dorsomedial prefrontal cortex mediates the interaction between moral and aesthetic valuation: A TMS study on the beauty-is-good stereotype. Social Cognitive and Affective Neuroscience, 12(5), 707–717.

Gwynne, D. T., & Rentz, D. C. F. (1983). Beetles on the bottle: Male buprestids mistake stubbies for females (Coleoptera). Journal of the Australian Entomological Society, 22(1), 79–80.

Henrich, J., & Gil-White, F. J. (2001). The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. Evolution and Human Behavior, 22(3), 165–196.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). Adversarial examples are not bugs, they are features. Advances in Neural Information Processing Systems, 32.

Kaplan, J. T., Gimbel, S. I., & Harris, S. (2016). Neural correlates of maintaining one's political beliefs in the face of counterevidence. Scientific Reports, 6, 39589.

Koebler, J. (2026). This 'digital camouflage' shirt confuses AI-powered surveillance cameras. 404 Media.

McGlone, M. S., & Tofighbakhsh, J. (2000). Birds of a feather flock conjointly (?): Rhyme as reason in aphorisms. Psychological Science, 11(5), 424–428.

Papadopoulos, V., Shah, M., Zimmerman, S., & Lindsey, J. (2026). Mind viruses: Self-propagating ideas in multi-agent LLM systems. arXiv:2608.10218.

Saadat, R. (2026). Why are we hackable? Three channels of infiltration in learning systems: From viruses to algorithms. Zenodo. https://doi.org/10.5281/zenodo.19819334

Schiestl, F. P. (2005). On the success of a swindle: Pollination by deception in orchids. Naturwissenschaften, 92(6), 255–264.

Schiestl, F. P., Ayasse, M., Paulus, H. F., Löfstedt, C., Hansson, B. S., Ibarra, F., & Francke, W. (1999). Orchid pollination by sexual swindle. Nature, 399, 421–422.

Staddon, J. E. R. (1975). A note on the evolutionary significance of "supernormal" stimuli. The American Naturalist, 109(969), 541–545.

Tinbergen, N. (1951). The study of instinct. Oxford University Press.

Tinbergen, N., & Perdeck, A. C. (1950). On the stimulus situation releasing the begging response in the newly hatched herring gull chick (Larus argentatus argentatus Pont.). Behaviour, 3(1), 1–39.
