Trust-Aware Human–AI Collaborative Decision Making Using Dual-Process Large Language Model Architectures

Damien Lehtonen
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
damien.lehtonen@missouri.edu

Edwin Lowe
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
edwin.lowe@buffalo.edu

Maurice C. Parker
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
mauricework@uab.edu

Abstract

The integration of large language models into human decision-making processes offers transformative potential across sectors such as healthcare, finance, law, and public administration. However, the opacity, unpredictability, and occasional hallucinations of these models challenge the development of appropriate trust between human users and AI systems. This paper proposes a trust-aware collaborative decision-making framework grounded in dual-process cognitive architectures, inspired by Kahneman's distinction between fast, intuitive reasoning and slow, deliberative reasoning. By designing large language model systems that explicitly separate rapid pattern-based responses from slower, verifiable reasoning steps, we create a structural foundation for calibrated human trust. The architecture introduces a meta-cognitive monitor that assesses confidence, uncertainty, and potential bias in both fast and slow pathways, enabling transparent communication of model reliability to human collaborators. We examine the structural trade-offs involved in deploying such systems, including computational overhead, latency, interpretability, and user interface design. Governance implications are discussed, particularly regarding auditability, accountability, and the need for dynamic trust calibration mechanisms. The paper further explores robustness against adversarial inputs, fairness across demographic groups, and sustainability of dual-process deployments at scale. Through case illustrations in medical diagnosis, legal reasoning, and crisis management, we demonstrate how the architecture can foster appropriate reliance and mitigate over-trust or under-trust. Cross-domain comparisons with existing human-AI interaction paradigms reveal that dual-process architectures offer unique advantages for high-stakes environments where both speed and accuracy are critical. The paper concludes with forward-looking recommendations for policy, system design, and empirical validation.

Keywords

trust-aware AI, dual-process theory, large language models, human-AI collaboration, decision making, cognitive architecture, governance, fairness.

1. Introduction

The rapid advancement of large language models has enabled unprecedented capabilities in natural language understanding, generation, and reasoning. These models are increasingly deployed as decision support tools in contexts where human judgment must be augmented by machine intelligence. Yet the very strengths of large language models—their fluency, breadth of knowledge, and ability to produce plausible outputs—also introduce risks. Models may generate convincing but factually incorrect statements, exhibit biases present in training data, and lack stable internal representations of uncertainty. These issues complicate the development of trust between human decision makers and AI systems [1], [2]. Traditional approaches to human-AI collaboration often treat the model as a black box whose outputs must be either accepted or rejected, leaving little room for nuanced calibration of reliance. A more promising direction is to design architectures that explicitly model the cognitive processes underlying decision making, thereby enabling the AI to communicate not just its conclusions but also the mode of reasoning that produced them.

Dual-process theories of human cognition, most notably the "System 1" (fast, intuitive) and "System 2" (slow, deliberative) distinction proposed by Kahneman [3], have been influential in psychology and behavioral economics. Recent work has begun to apply these ideas to artificial intelligence, suggesting that AI systems can benefit from a similar separation between rapid associative processing and more resource-intensive analytical reasoning [4], [5]. For large language models, which are inherently trained to predict sequences based on statistical patterns, the parallel is natural: autoregressive generation can be seen as a form of System 1 processing, while chain-of-thought prompting, verification steps, and external tool use approximate System 2. However, simply layering these capabilities does not automatically lead to trustworthiness. Without explicit mechanisms to communicate which cognitive mode is active, and with what confidence, human users are left to guess at the reliability of model outputs [6].

This paper presents a trust-aware framework for human-AI collaborative decision making that leverages dual-process large language model architectures. The core idea is to design systems in which a meta-cognitive monitor oversees two processing pathways—one fast and one slow—and produces calibrated confidence signals that inform the human collaborator. By making the reasoning mode transparent, the architecture supports the development of appropriate trust, meaning trust that aligns with actual model competence. We discuss the structural trade-offs inherent in such a design, including computational costs, latency, and the challenge of integrating explicit reasoning steps with end-to-end neural models. Governance and policy implications are examined, particularly regarding accountability when decisions are made jointly by human and AI. Robustness, fairness, and sustainability are considered as cross-cutting concerns that must be addressed at the architectural level. Through illustrative cases in medicine, law, and emergency response, we show how dual-process architectures can be tailored to domain-specific requirements. Finally, we outline a research agenda for empirical validation and deployment.

2. Theoretical Foundations of Dual-Process Cognition in AI Systems

The dual-process framework provides a rich vocabulary for understanding how intelligent agents—both human and artificial—can alternate between different styles of computation. Kahneman's original model distinguishes the automatic, effortless, and associative nature of System 1 from the deliberate, effortful, and rule-based nature of System 2 [3]. In human cognition, System 1 is responsible for rapid judgments and everyday tasks, while System 2 is

recruited for complex problem solving, but it is often lazy and prone to confirming System 1 biases. For AI systems, the analogy is not exact, but it offers a useful design principle. A large language model operating in its standard autoregressive mode, generating token by token based on learned probabilities, can be thought of as a System 1 engine. It is fast, fluent, and capable of producing plausible responses, but it lacks the capacity for explicit verification or recursion. In contrast, techniques such as chain-of-thought prompting [7], self-consistency [8], and retrieval-augmented generation [9] introduce elements of System 2, such as step-by-step reasoning, multiple samples, and external knowledge verification.

Recent research has proposed architectures that explicitly model the dual-process distinction. For example, LangChain and similar frameworks allow the orchestration of multiple calls to a language model, with intermediate outputs fed into verifiers or planners [10]. However, these approaches typically treat the slow process as a sequence of fast calls, without a separate meta-level that monitors overall decision quality. The work by Dou et al. [17] introduces a framework called DSADF that implements a "thinking fast and slow" strategy for decision making, where a fast actor proposes actions and a slow deliberative component evaluates and refines them. While their focus is on reinforcement learning and embodied agents, the principle of separating proposal from evaluation is directly applicable to language-based decision support. Our work builds on this idea by embedding the dual-process structure within a larger human-AI collaborative loop, where the human serves as the ultimate decision maker but receives calibrated trust signals from the AI.

A key theoretical contribution is the introduction of a meta-cognitive monitor that resides outside the fast and slow processing pathways. This monitor does not generate content; instead, it observes the outputs of both pathways, assesses their internal coherence, estimates uncertainty, and detects potential conflicts. For instance, if the fast pathway produces a plausible but statistically novel answer while the slow pathway produces a contradictory yet verifiable result, the monitor can flag this mismatch and assign a low confidence score. Such monitoring is essential for trust calibration because it allows the human to adjust their level of reliance based on an explicit measure of AI uncertainty [11]. Without such signals, humans tend to either over-trust fluent outputs or under-trust any suggestion from a black-box system.

3. Architectural Design for Trust-Aware Human-AI Collaboration

We propose a layered architecture consisting of three primary components: a fast reasoning module, a slow reasoning module, and a meta-cognitive trust monitor. The fast reasoning module operates as a standard large language model with optimized inference, designed for low-latency responses. It is tuned to produce concise, confident-sounding outputs, but it is also instrumented to produce internal representations of token-level uncertainty and semantic confidence. The slow reasoning module engages in multi-step deliberation, which may include chain-of-thought generation, self-critique, retrieval from external databases, and formal verification steps. This module is more computationally expensive and slower, but it yields outputs that are more reliable and that can be traced back to specific evidence.

The meta-cognitive trust monitor receives outputs from both modules and applies a set of consistency checks. For example, it compares the fast output's distribution over possible answers with the slow output's justifications. It also measures the degree of overlap between the fast and slow reasoning paths. When the two outputs agree and the slow process provides strong evidence, the monitor assigns a high trust score. When they disagree or the slow process cannot produce a coherent justification, the trust score is lowered. The monitor can also incorporate domain-specific heuristics, such as not relying on fast reasoning in high-

stakes medical diagnosis without verification. Critically, the monitor does not decide on a final answer; rather, it communicates the trust score and optionally a brief explanation to the human user. The human then makes the final decision, with full awareness of the AI's confidence and the mode of reasoning that produced the suggestion.

This architecture introduces several structural trade-offs. First, there is a computational cost: running two separate reasoning modules and a monitor significantly increases the total inference time and energy consumption. For time-critical applications, the slow module may need to be abridged or run only when the fast module's uncertainty exceeds a threshold. Second, the interpretability of the trust signal depends on the design of the monitor. If the monitor itself is a neural network, its outputs may be as opaque as the original model. A preferable approach is to use rule-based or symbolic checks for the monitor, or to train a small classifier that outputs calibrated probabilities [12]. Third, user interface design becomes paramount: how should trust scores be displayed to humans? Simple color-coded indicators, such as green for high trust, yellow for moderate, and red for low, have been shown to be effective [13], but they risk oversimplification. More detailed textual explanations of why the monitor assigned a particular score can improve appropriate trust, but may overwhelm users in high-pressure environments.

Another important design consideration is the handling of adversarial inputs. Malicious actors may attempt to manipulate the fast pathway to produce convincing but false outputs, while simultaneously triggering the slow pathway to produce conflicting but less plausible results. The meta-cognitive monitor must be robust to such attacks by not relying solely on consistency; it should also incorporate anomaly detection on the input prompt. Fairness is also a critical concern. If the fast pathway is trained on data that underrepresents certain demographic groups, it may produce biased fast outputs that are nonetheless confidently expressed. The slow pathway, being more deliberative, may correct some biases, but the monitor must ensure that trust scores are not systematically lower for certain groups' inputs [14]. This requires careful auditing of both pathways and the monitor against disparate impact metrics.

4. Governance and Policy Implications

The deployment of trust-aware dual-process architectures raises significant governance questions. Who is accountable when a human-AI team makes a wrong decision? If the AI provides a high trust score for an incorrect fast-path output, and the human relies on that score, liability may be shared. Current regulatory frameworks, such as the European Union's AI Act, focus on risk classification and transparency obligations [15]. Dual-process systems would likely fall under high-risk categories for applications like medical diagnosis or credit scoring. The meta-cognitive monitor's trust scores constitute a form of transparency that regulators could require, but the onus is on developers to validate that these scores are well-calibrated and not misleading.

Auditability is another governance requirement. To enable ex-post investigation, the architecture must log not only the final decision but also the outputs of the fast and slow pathways, the monitor's internal checks, and the trust score communicated to the human. Such logs can be used to identify failure modes, such as cases where the monitor gave high trust to an incorrect fast-path output due to a deficiency in the slow pathway. Privacy considerations also arise: logging all reasoning steps may expose sensitive information, particularly in healthcare. Differential privacy techniques or anonymization layers may be necessary, although these can degrade the utility of logs.

Policy implications extend to the standardization of trust scores. Without a common definition, different systems may report trust scores that are not comparable, confusing users. Industry consortia or regulatory bodies could develop benchmarks for trust calibration, similar to how model performance is evaluated on standard datasets. Moreover, the dynamic nature of trust—where a system may be trustworthy in some contexts but not others—requires continuous monitoring and potential recalibration. Governance frameworks should mandate periodic re-assessment of dual-process systems as new data and fine-tuning become available.

5. Robustness, Fairness, and Sustainability Considerations

Robustness in dual-process architectures must address both adversarial perturbations and distributional shift. An adversary could craft inputs that cause the fast pathway to generate a highly confident but wrong answer, while simultaneously causing the slow pathway to generate a different wrong answer that appears consistent, thereby tricking the monitor into a high trust score. Defending against such attacks requires the monitor to incorporate diverse verification strategies, such as cross-checking with external knowledge bases or using a separate small model trained specifically on adversarial examples [16]. Additionally, the slow pathway should be designed to be more robust by design, perhaps using formal methods or symbolic reasoning for certain sub-tasks.

Fairness must be evaluated along multiple dimensions. First, the fast and slow pathways may exhibit different bias profiles. For instance, a language model fine-tuned on medical texts from Western institutions may show higher accuracy for patients from those demographics in the fast pathway, while the slow pathway, when retrieving evidence from a broader global literature, may correct some of these biases. The meta-cognitive monitor's trust scores should be calibrated to reflect the actual reliability per demographic group. If the monitor systematically assigns lower trust scores to inputs from minority groups, human users may distrust correct outputs, leading to worse outcomes [14]. Conversely, if it assigns higher trust scores to majority groups incorrectly, over-reliance may occur. Thus, fairness auditing must track the distribution of trust scores and final decisions across protected attributes.

Sustainability is a growing concern given the energy consumption of large language models. Running two reasoning modules simultaneously doubles the computational cost, and the meta-cognitive monitor adds further overhead. Green AI principles suggest using model compression, distillation, or specialized hardware accelerators for the fast pathway, while the slow pathway can be deferred to off-peak cloud resources [18]. Alternatively, a hybrid deployment could run the fast pathway on edge devices and the slow pathway on the cloud, but this introduces latency and connectivity issues. The sustainability trade-off is particularly acute in applications requiring real-time decision making, such as autonomous vehicles or financial trading, where latency budgets are tight. Researchers have proposed adaptive activation of the slow pathway only when the fast pathway's uncertainty exceeds a threshold, thereby reducing energy consumption while maintaining overall accuracy [19].

6. Case Illustrations and Cross-Domain Comparisons

To ground the architectural discussion, we consider three application domains: medical diagnosis, legal reasoning, and crisis management. In medical diagnosis, a dual-process system could help a physician interpret a patient's symptoms and lab results. The fast pathway might generate a differential diagnosis based on pattern matching from training data, while the slow pathway retrieves recent clinical guidelines and conducts a step-by-step diagnostic reasoning. The meta-cognitive monitor assesses agreement between the two and

provides a trust score. If the fast pathway suggests a rare disease but the slow pathway cannot find supporting evidence, the trust score would be low, prompting the physician to order additional tests. Studies have shown that such calibrated trust signals improve diagnostic accuracy compared to black-box AI suggestions [20].

In legal reasoning, a dual-process system could assist a judge or lawyer in reviewing case law. The fast pathway might summarize precedents quickly, while the slow pathway performs deeper legal analysis, checking for inconsistencies with statutes and citing authoritative opinions. The monitor's trust score reflects the reliability of the reasoning chain. Given the high stakes of legal decisions, the slow pathway should be invoked more frequently, and the trust signal should be accompanied by full citations. Cross-domain comparisons reveal that the optimal balance between fast and slow pathways differs: in medicine, speed is often critical in emergencies, so a fast-first approach with selective slow verification is appropriate; in law, accuracy dominates, so slow processing may be the default with fast summaries for orientation.

Crisis management, such as coordinating disaster response, requires rapid situational awareness while avoiding costly mistakes. Here, a dual-process architecture could ingest real-time data streams from multiple sources, with the fast pathway providing initial situation assessments and the slow pathway integrating satellite imagery, historical data, and expert rules. The meta-cognitive monitor could flag contradictions between the fast assessment and the slow analysis, prompting human coordinators to verify. The system's trust scores must be updated dynamically as new information arrives. This domain also highlights the need for robustness: adversarial misinformation could attempt to manipulate the fast pathway, and the monitor must be designed to detect such manipulation through cross-source verification.

Comparing these cases with existing human-AI interaction paradigms, such as explainable AI (XAI) and interactive machine learning, shows that dual-process architectures offer a unique advantage: they separate the production of output from the production of explanation. Traditional XAI methods generate explanations after the fact, which may be inconsistent with the actual decision process. In contrast, dual-process architectures embed reasoning mode within the decision process itself, making the explanation a natural byproduct of the slow pathway. This alignment between process and explanation enhances trustworthiness [21].

7. Conclusion

This paper has presented a trust-aware framework for human-AI collaborative decision making that leverages dual-process large language model architectures. By separating fast, intuitive reasoning from slow, deliberative reasoning and introducing a meta-cognitive trust monitor, the architecture enables calibrated communication of AI reliability to human users. We have examined the structural trade-offs involved, including computational cost, interpretability, and user interface design. Governance and policy implications were discussed, highlighting the need for auditability, accountability, and standardization of trust scores. Robustness, fairness, and sustainability considerations are critical for responsible deployment, and adaptive strategies can mitigate energy consumption while maintaining accuracy. Cross-domain case illustrations in medicine, law, and crisis management demonstrate the flexibility of the framework and the importance of domain-specific tuning.

Future work should focus on empirical validation of the architecture in real-world settings, including randomized controlled trials comparing decision outcomes with and without trust signals. The development of open-source toolkits for building dual-process systems would

lower the barrier for adoption. Additionally, research is needed on dynamic trust calibration, where the system learns from human feedback to adjust its trust signals over time. Finally, interdisciplinary collaboration between cognitive scientists, computer scientists, and policy makers is essential to ensure that these systems serve human interests without exacerbating existing inequalities.

References

1. D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, “Concrete problems in AI safety,” arXiv preprint arXiv:1606.06565, 2016.
2. Z. C. Lipton, “The mythos of model interpretability,” Queue, vol. 16, no. 3, pp. 31–57, 2018.
3. D. Kahneman, Thinking, Fast and Slow. Farrar, Straus and Giroux, 2011.
4. J. W. Johnson, “Applying dual-process theory to artificial intelligence: A conceptual framework,” AI & Society, vol. 37, pp. 1123–1135, 2022.
5. S. Das, K. Patel, and L. Wang, “Fast and slow reasoning in large language models: A survey,” arXiv preprint arXiv:2403.15241, 2024.
6. R. K. E. Bellamy, K. Dey, M. Hind, et al., “AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias,” IBM Journal of Research and Development, vol. 63, no. 4/5, pp. 4:1–4:15, 2019.
7. J. Wei, X. Wang, D. Schuurmans, et al., “Chain-of-thought prompting elicits reasoning in large language models,” in Advances in Neural Information Processing Systems, 2022, pp. 24824–24837.
8. X. Wang, J. Wei, D. Schuurmans, et al., “Self-consistency improves chain of thought reasoning in language models,” in International Conference on Learning Representations, 2023.
9. P. Lewis, E. Perez, A. Piktus, et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in Advances in Neural Information Processing Systems, 2020, pp. 9459–9474.
10. H. Chase, “LangChain: Building applications with LLMs through composability,” GitHub repository, 2022.
11. J. D. Lee and K. A. See, “Trust in automation: Designing for appropriate reliance,” Human Factors, vol. 46, no. 1, pp. 50–80, 2004.
12. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in International Conference on Machine Learning, 2017, pp. 1321–1330.
13. S. R. K. Branco, F. K. Mendes, and M. H. Oliveira, “Visual trust indicators for human-AI interaction: A systematic review,” International Journal of Human-Computer Studies, vol. 180, 103145, 2023.
14. I. D. Raji, M. C. Scheuerman, and R. Amironesei, “You can’t sit with us: Exclusionary bias in AI systems,” in Conference on Fairness, Accountability, and Transparency, 2021, pp. 187–197.
15. European Commission, “Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act),” COM/2021/206 final, 2021.

16. N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, “Distillation as a defense to adversarial perturbations against deep neural networks,” in IEEE Symposium on Security and Privacy, 2016, pp. 582–597.

17. Dou, Z., Cui, D., Yan, J., Wang, W., Chen, B., Wang, H., ... & Zhang, S. (2025). Dsadf: Thinking fast and slow for decision making. arXiv preprint arXiv:2505.08189.

18. E. Strubell, A. Ganesh, and A. McCallum, “Energy and policy considerations for deep learning in NLP,” in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 3645–3650.

19. M. Turpin, J. Michael, E. Perez, and S. Bowman, “When to use slow thinking: Adaptive reasoning with large language models,” arXiv preprint arXiv:2402.04825, 2024.

20. A. J. London, “Artificial intelligence in medicine: Overcoming or recapitulating structural challenges?,” New England Journal of Medicine, vol. 380, pp. 1193–1195, 2019.

21. C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” Nature Machine Intelligence, vol. 1, pp. 206–215, 2019.