Inferring Cognitive Load from Deviations in Simulated Human Planning Behavior

May Kristine Jonson Carlon (mcarlon@acm.org)
Social Cognition and Behavior Collaboration Unit, CBS-Toyota Collaboration Center, RIKEN Center for Brain Science
Wako, Saitama, Japan

Abstract

Understanding how cognitive load shapes human planning behavior is crucial for building AI systems that collaborate effectively with people. While traditional approaches to measuring cognitive load such as self-report questionnaires or dual-task paradigms are valuable, they often lack real-time responsiveness or introduce artificial task constraints. This work is a proof-of-concept for inferring cognitive load from deviations in planning, thus avoiding intrusive or retrospective measures. We simulate user profiles performing a structured task (summarizing an article), with behavioral noise introduced via repetition, backtracking, pausing, and skipping actions. A Hidden Markov Model (HMM) is used to infer latent cognitive states from the resulting behavioral traces. Results from 100 Monte Carlo trials show that the HMM reliably recovers latent states aligned with intuitive levels of cognitive load. Emission patterns are interpretable, stable across trials, and distinct for each state, capturing predetermined behavioral signatures of low, medium, and high mental effort. State assignments also show alignment with simulated user profiles. Our approach provides a simulation-based foundation for modeling cognitive variability and may inform future work in user modeling, Theory of Mind, and adaptive systems.

Keywords: cognitive load; planning behavior; Theory of Mind; inverse modeling; Hidden Markov Models; simulation; human-AI interaction

Introduction

Human planning is rarely flawless. Even in structured tasks, individuals may repeat steps, backtrack, hesitate, or skip actions. While often dismissed as behavioral noise, such deviations can reflect latent cognitive states, particularly mental effort under uncertainty. Cognitive load theory suggests that performance degrades under high task demands or limited working memory (Sweller, 1988). Yet most computational models of planning assume idealized agents, overlooking natural variability in behavior (Gershman, Horvitz, & Tenenbaum, 2015; Chandramouli et al., 2024).

To support human planning, artificial systems must infer users' cognitive states in real-time. Theory of Mind (ToM) - the ability to model others' beliefs, goals, and knowledge - has long been a cornerstone of cognitive science (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017) and is increasingly vital for human-AI collaboration (Ho, Saxe, & Cushman, 2022; ichter et al., 2023; Wang et al., 2024). We propose a computational ToM approach based on inverse modeling: the system observes user behavior and infers hidden cognitive load states, without assuming rationality or optimality.

We formalize this process with a probabilistic generative model. Inspired by Bayesian inverse planning (Baker, Saxe, & Tenenbaum, 2009) and machine ToM architectures such as ToMnet (Rabinowitz et al., 2018), we treat the user as a noisy planner whose internal state shapes behavioral patterns. A Hidden Markov Model (HMM) is used to infer latent cognitive states from observed behavior, in a framework that is both tractable and interpretable.

Methods

To motivate further design and human data collection, we begin with simulation. In robotics, complex behaviors in tasks are devised into high-level plans; we adopted a similar approach to model writing by defining an ideal five-step plan: `["read", "extract", "write", "revise", "submit"]`. We can imagine that individuals might follow distinct trajectories or exhibit characteristic behavior patterns, or profiles, during task execution. Six profiles — *overconfident*, *efficient*, *cautious*, *self-correcting*, *novice*, and *anxious* — were simulated by varying parameters for repetition, backtracking, pause duration, variability, and skipping (Lieder & Griffiths, 2020). This provides a scaffold for developing interactive systems that anticipate user needs based on observed actions. This idea also parallels sparse sampling strategies used in near-optimal planning under resource constraints (Kearns, Mansour, & Ng, 2002).

Each simulated user yielded a six-dimensional behavioral vector (e.g., repeat count, total pause time, skip count), which were standardized before modeling. We then use a Gaussian HMM with three latent states (low, medium, high load) and full covariances. Each model was trained for up to 1,000 iterations. To ensure generalization, we applied 5-fold cross-validation within each Monte Carlo trial and selected the model with the highest held-out log-likelihood. Trials with convergence failures were excluded. We simulated up to 100 trials, retaining at least 20 valid runs for statistical reliability.

Results

The HMM demonstrated strong performance across trials, with a mean cross-validated log-likelihood of 989.47 ± 417.41, indicating a robust fit to simulated behavioral data. Chi-square tests of independence on the profile-to-state assignments revealed highly significant association ($\chi^2 = 1092.58 \pm 157.15$, $p < 0.0001$ in all trials), confirming that latent states systematically reflected the underlying cognitive profiles rather than random variation. The consistent significance rate (100% of trials with $p < 0.05$) further supports the model's stability and interpretability as a discriminator of cognitive load.

Emissions represent the observable behavioral features associated with each hidden state which are shown in Figure 1. As expected from the simulation parameters, most behavioral features such as repetition, backtracking, and pausing increase from the Low to High load states. This reflects our design assumption that higher cognitive load manifests as more effortful behavior. Skipping, in contrast, decreases under higher load, consistent with reduced planning flexibility under mental strain.

Figure 2 shows the alignment between cognitive profiles and latent states. Overconfident users map primarily to Low cognitive load, while novice and anxious users align with High

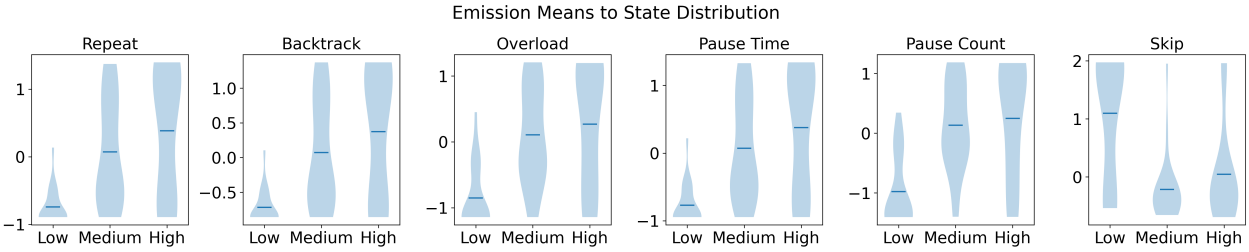


Figure 1: Emission means across latent states. Features such as repetition, pausing, and overloads increase from Low to High cognitive load, while skipping decreases.

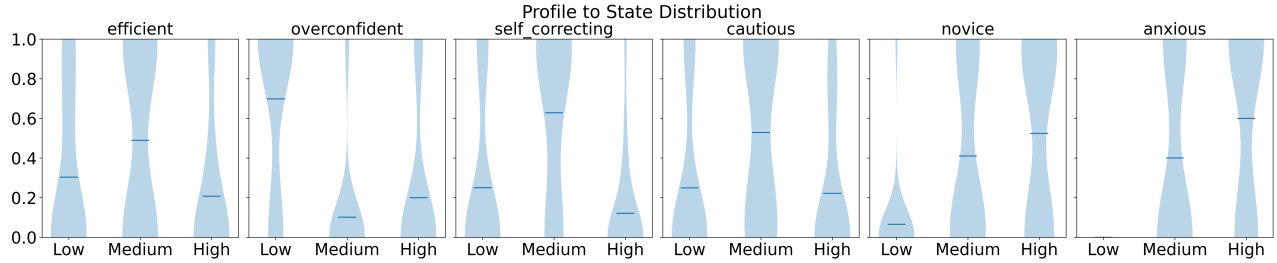
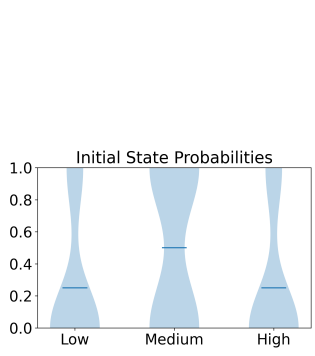


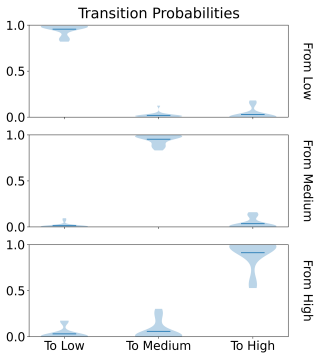
Figure 2: Profile-to-state mapping across trials. Overconfident users are aligned with Low load states, while novice and anxious users align with High. The rest are aligned with Medium.

load as expected. Cautious and self-correcting profiles tend to fall in the Medium range. Efficient users, however, show a broader distribution, with many aligning to Medium load. These associations are consistent across Monte Carlo trials.

Finally, Figure 3 summarizes latent state dynamics. Initial state probabilities (Figure 3a) show a modest bias toward the Medium load state at sequence onset, though all three states are represented. Transition probabilities (Figure 3b) reveal strong persistence within each state, with high self-transition likelihoods for Low, Medium, and High. Cross-state transitions are relatively rare, suggesting that cognitive load levels remain stable over short time spans.



(a) Initial state probabilities across models. A modest bias toward the Medium load state is observed, though all three states are represented across trials.



(b) Transition probabilities between latent states. Each state shows high self-transition likelihood, indicating behavioral stability within cognitive load levels.

Figure 3: Latent state dynamics across trials.

Discussion

Our findings show that latent cognitive states can be inferred from structured deviations in planning behavior, supporting the use of HMMs as real-time inverse models of user cognition, akin to lightweight computational ToM (Akyürek, 1992). Rather than assuming rational planning, we model bounded agents whose behavioral irregularities reflect internal pressures like uncertainty or fatigue. While not classical inverse planning which presumes optimality, our approach shares the goal of inferring hidden states from observed behavior. Future work should examine model selection and profile-specific modeling errors to refine inference.

This simulation-based framework enables controlled experimentation but cannot capture the full complexity of human cognition. Validation with human participants is needed, comparing inferred states with measures such as NASA-TLX (Hart & Staveland, 1988), pupillometry, or neuroimaging. In particular, linking state transitions with pupil dilation or prefrontal activity (Garrett, Epp, Kleemeyer, Lindenberger, & Polk, 2020) may ground our inferences in neurocognitive processes. While we model discrete states, cognitive load likely varies continuously, and future models may benefit from hybrid or continuous-state representations.

Simulation is only a starting point. Real-world plans and profiles will differ across tasks, depending on what users optimize for (e.g., speed, creativity). Meaningful plan and profile discovery requires extensive observation of human behavior, and this behavioral grounding is essential for the validation steps above. Future work should attend to elements underplayed in the current simulation, such as goal variability or feature interactions, that may be key for adaptive system design.

Acknowledgments

The RIKEN Center for Brain Science (CBS) - Toyota Collaboration Center (BTCC) funded this research. The funder had no involvement in the study or publication. The views expressed are solely those of the author.

I would also like to thank Dr. Yasuo Kuniyoshi, head of the Social Cognition and Behavior Collaboration Unit at BTCC, for his guidance and support.

References

Akyürek, A. (1992). On a computational model of human planning. In J. A. Michon & A. Akyürek (Eds.), *Soar: A cognitive architecture in perspective: A tribute to allen newell* (pp. 81–108). Dordrecht: Springer Netherlands. doi: 10.1007/978-94-011-2426-3_4

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017, Mar 13). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064. doi: 10.1038/s41562-017-0064

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329-349. (Reinforcement learning and higher cognition) doi: https://doi.org/10.1016/j.cognition.2009.07.005

Chandramouli, S., Shi, D., Putkonen, A., De Peuter, S., Zhang, S., Jokinen, J., ... Oulasvirta, A. (2024, Sep 01). A workflow for building computationally rational models of human behavior. *Computational Brain & Behavior*, *7*(3), 399-419. doi: 10.1007/s42113-024-00208-6

Garrett, D. D., Epp, S. M., Kleemeyer, M., Lindenberger, U., & Polk, T. A. (2020). Higher performers upregulate brain signal variability in response to more feature-rich visual input. *NeuroImage*, *217*, 116836. doi: https://doi.org/10.1016/j.neuroimage.2020.116836

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273-278. doi: 10.1126/science.aac6076

Hart, S. G., & Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (Vol. 52, p. 139-183). North-Holland. doi: https://doi.org/10.1016/S0166-4115(08)62386-9

Ho, M. K., Saxe, R., & Cushman, F. (2022, Nov 01). Planning with theory of mind. *Trends in Cognitive Sciences*, *26*(11), 959-971. doi: 10.1016/j.tics.2022.08.003

ichter, b., Brohan, A., Chebotar, Y., Finn, C., Hausman, K., Herzog, A., ... Fu, C. K. (2023, 14–18 Dec). Do as i can, not as i say: Grounding language in robotic affordances. In K. Liu, D. Kulic, & J. Ichnowski (Eds.), *Proceedings of the 6th conference on robot learning* (Vol. 205, pp. 287–318). PMLR.

Kearns, M., Mansour, Y., & Ng, A. Y. (2002, Nov 01). A sparse sampling algorithm for near-optimal planning in large markov decision processes. *Machine Learning*, *49*(2), 193-208. doi: 10.1023/A:1017932429737

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, e1. doi: 10.1017/S0140525X1900061X

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. M. A., & Botvinick, M. (2018, 10–15 Jul). Machine theory of mind. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th international conference on machine learning* (Vol. 80, pp. 4218–4227). PMLR.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257-285. doi: https://doi.org/10.1016/0364-0213(88)90023-7

Wang, Q., Walsh, S., Si, M., Kephart, J., Weisz, J. D., & Goel, A. K. (2024). Theory of mind in human-ai interaction. In *Extended abstracts of the chi conference on human factors in computing systems.* New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3613905.3636308