Quantifying Uncertainty in Machine Theory of Mind Across Time

Shanshan Zhang[1,*,†], Chuyang Wu[1,2,†] and Jussi P. P. Jokinen[2]

[1]University of Helsinki, Pietari Kalmin katu 5, 00560 Helsinki Finland
[2]University of Jyväskylä, Seminaarinkatu 15, PL 35, 40014 Jyväskylä Finland

Abstract

As intelligent interactive technologies advance, ensuring alignment with user preferences is critical. Machine theory of mind enables systems to infer latent mental states from observed behaviors, similarly to humans. Currently, there is no formal mechanism for integrating multiple observations over time and quantifying the uncertainty of inferences as the function of accumulated evidence in a provably human-like way. This paper addresses the issue through Bayesian inference, proposing a model that maintains a posterior belief about mental states as a probability distribution, updated with observational data. The advantage of Bayesian statistics lies in the possibility of evaluating the certainty of these inferences. We validate the model's human-like mental inference capabilities through an experiment.

Keywords

Human-Computer Interaction, Machine Theory of Mind, Mentalizing, Uncertainty Quantification

1. Introduction

Theory of mind, the innate human capacity to deduce others' latent mental states from observable behavior [1, 2], underpins social collaboration [3, 4]. As artificial intelligence (AI) advances, aligning intelligent machines with users' preferences becomes imperative [5]. Achieving alignment between human and machine objectives is facilitated when machines adopt reasoning processes that can be understood by humans [6], suggesting the importance of machines emulating human mental inference. A *machine theory of mind* seeks to provide machines with the ability to infer mental states in a human-like manner.

Mental inference facilitates collaboration by informing the agent and impacting its actions. The idea is that if an intelligent machine has knowledge of the user's goals, it can better make decisions to help the user. However, there is also an inherent risk in making decisions based on inferences: because all inferences contain uncertainty [7, 8], the intelligent agent should have a way of considering the amount of uncertainty when taking actions. There needs to be a way to quantify the amount of uncertainty, so that the agent can robustly consider this when choosing what actions to take. In this paper, we formalize a computational model that infers preferences of observed agents. Observations from multiple time steps are integrated, and the uncertainty associated with inferences is quantified in a posterior distribution.

The problem that our paper tackles is illustrated in Figure 1. The three panels depict an evolving inference by an observer of Janice's drink preference under varying conditions in three consecutive days. Initially, Janice selects tea, but the positioning of coffee on a high shelf introduces ambiguity regarding her preference – does she favor tea, or does she simply wish to avoid climbing the kitchen ladder? This uncertainty prevents a clear inference of her preference. In the second panel, Janice uses a stool to reach the now higher-placed tea jar, while the coffee remains even further out of reach, potentially accessible with taller kitchen stairs. The scenario hints at a preference for tea, yet the possibility that Janice may have an aversion to heights carries a degree of uncertainty, nudging the likelihood slightly in favor of tea.

The final panel of Figure 1 offers a decisive moment: both coffee and tea jars are easily accessible, and Janice opts for coffee. Given the equal effort required to reach both, her choice of coffee indicates a genuine preference for coffee, revealing that her earlier decisions were influenced by a reluctance to climb too high rather than a preference for tea. Consequently, our inference shifts significantly towards coffee with increased certainty. In this paper, we hypothesize that humans are able to carry out these sorts of inferences and meta-cognitively assess how certain they are in inferred preferences. Moreover, we formalize a computational model of this process.

2. Background Review

Theory of mind, or mentalizing, enables humans to infer others' mental states [9, 10, 11]. It facilitates social interaction [3, 4] such as communication [12, 13] and collaboration [1, 2]. Likewise, a machine that is able to carry out mentalization can better account user variability, improving the quality of interaction [14, 15, 16]. Experiments have demonstrated that machines capable of mentalization achieve superior performance in communication [17, 18] and team cooperation tasks [19].

Models of mentalizing target the inference of mental states such as preferences, costs [20], knowledge [21], and beliefs [9]. These models incorporate psychological hypotheses concerning of observed actors as computational frameworks, enabling the simulation of predicted behavior. Parameters within the model reflect various mental states, including goals, guiding the behavior prediction for actors under specific objectives in a given context [22]. Assuming the psychological underpinnings are accurate, these models can predict an actor's behavior based on their goals. Inverse modeling techniques are then employed to deduce the parameters most likely to account for the observed behavior [23, 24].

How to create a psychologically plausible model that can be parametrized with mental states and that then simulates behavior? One emerging popular approach is called *computational rationality* [25, 26]. It posits that intelligent agents,

TKTP 2024: Annual Doctoral Symposium of Computer Science, 10.-11.6.2024 Vaasa, Finland

[*] Corresponding author.

[†] These authors contributed equally.

shanshan.zhang@helsinki.fi (S. Zhang); chuyang.wu@helsinki.fi (C. Wu); jussi.p.p.jokinen@jyu.fi (J. P. P. Jokinen)

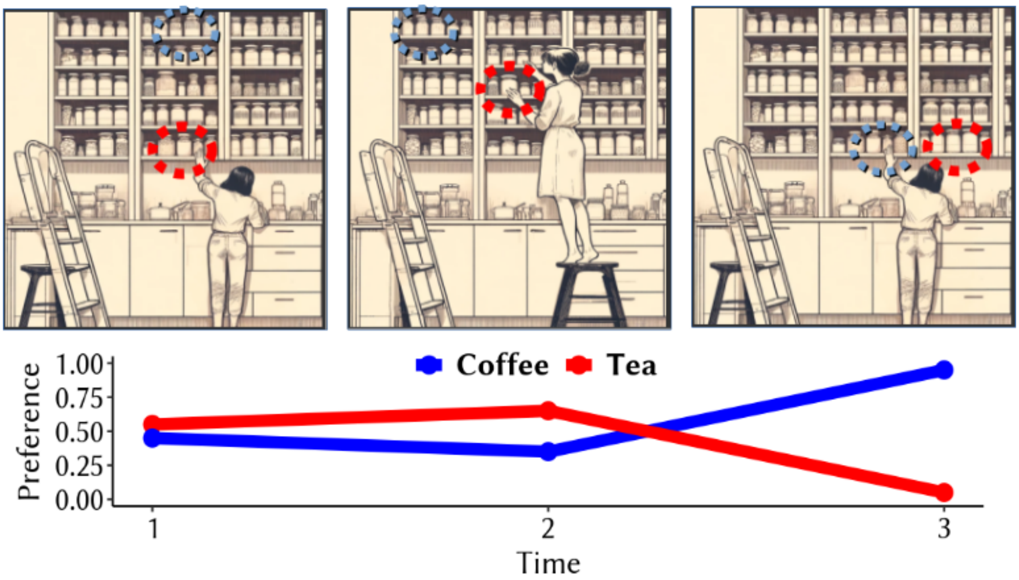

Figure 1: Inferences of preferences based on observed behavior contain uncertainty, especially when there are confounding factors such as effort. As more evidence accumulates, certainty increases.

such as humans, choose actions that maximize expected utility. The agent must optimize its behavior with respect to the constraints environment. In addition, the approach is sensitive to the fact that intelligent agents have internal cognitive bounds as well, such as limited knowledge and information processing capacity. The approach is suitable for computational modeling of theory of mind, because it helps to prune the space of possible explanations by assuming that the observed behavior is produced by a computational rational agent. When the bounds of the environment and the cognition are known and modeled correctly, the model can then be applied for reliable parameter inference [27].

Inferences, including those related to mentalizing, are often made under conditions of limited data, inherently involving uncertainty [28, 7]. The similarity in actions among individuals with diverse preferences in specific contexts implies that observations alone may not suffice for conclusive inferences. The complexity of social settings further amplifies this uncertainty, highlighting the importance of incorporating it into models of social collaboration [29]. Thus, agents capable of mentalizing should not only emulate human-like inference of mental states but also assess the uncertainty of these inferences.

3. Method

Following the standard modeling pipeline in computational rationality [26], we formalize the task environment as a Markov Decision Process (MDP). It is represented as a tuple $< S, A, T, R >$, consisting state space S, action space A, transition probabilities T and reward function R. A state $s \in S$ encoding current information of the environment, transfers to next state $s' \in S$ by performing an action $a \in A$ according to transition probability $T(s, a, s') = P(s'|s, a)$, and gains the reward $r = R(s, a)$. Reinforcement learning (RL) solves the optimization problem of how to choosing the action a through policy $\pi(a) = P(a|s)$ that maximizes the expected reward by interacting with the environment and learning from experience. The learning process can be expressed as the function

$$V_{\pi^*}(s) = \max_a [R(s, a) + \gamma \sum_{s' \in S} T(s, a, s')V_{\pi^*}(s')],$$

where $V_{\pi^*}(s)$ is the value of a state $s \in S$ under an optimal policy π^*, discounting future rewards using $\gamma \in [0, 1]$. This optimality assumption ties in with computational rationality. Importantly, it is possible to implement bounds in the MDP formalism, forcing a bounded optimal behavior to emerge.

The bounded optimal agent described via an MDP can be parametrized. For instance, a parameter can govern its preferences, that is, the state rewards. This permits mentalizing: given observed data, what parameters best produce predicted data that fits the observations? To this end, we utilize Bayesian inference, described by Bayes' rule:

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)},$$

where θ represents the latent factors to be inferred, and x represents observed data. The inference uses a prior $P(\theta)$ and a likelihood $P(x|\theta)$ to calculate posterior probability $P(\theta|x)$, normalized with marginal likelihood $P(x)$. However, the intractability of the likelihood $P(x|\theta)$ prevents us from deriving the posterior directly. This can be overcome with approximation and likelihood free inference methods [30], such as Bayesian Optimization for Likelihood-Free Inference (BOLFI) [31].

Figure 2 illustrates the information flow in our model. Prior knowledge and observation data serve as inputs of an inference module, which parameterizes a RL agent. The agent then learns a bounded optimal policy within a simulator modeling the observed real-world task. Through multiple samplings, plausibility for various parameter values is evaluated, forming a posterior distribution that serves as the prior for subsequent inference with new observation data. This framework facilitates the temporal integration of inferences and allows for uncertainty analysis within the posterior probability distribution. All model details are available at the model's code repository (https://version.helsinki.fi/shanz/quantifying-uncertainty-in-mtom.git).

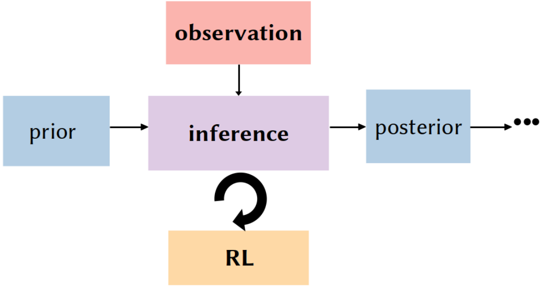

Figure 2: The overall structure of the model. It consists of simulation of external world and inference module, which can be repeated as new observation comes.

4. Evaluation

4.1. Participants

We recruited $N = 10$ participants via the Prolific online platform. The number of participants was small, but because our experiment setup was well defined, we expected them to have a high agreement with each other. This was the case, meaning that a larger number of participants would likely not have changed the results. Their mean age was 35.6, and age range 23-56. They were required to be fluent in English, and be on a PC (no mobile devices were allowed).

4.2. Materials

The experiment consisted of eight distinct tasks, each including five stimulus images. One image shows a trajectory of a robot on a grid from a birds-eye perspective. The robot is moving from its starting position to either a blue or red circle, representing charging stations. There may also be walls, and the robot must navigate around them. Each picture is different, and there were a total of $8 \cdot 5 = 40$ stimuli. An example task is shown in Figure 3.

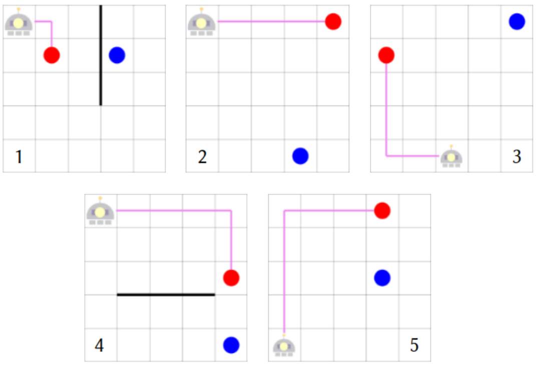

Figure 3: The five stimuli shown sequentially to the participants, Task 1. Stimulus numbers are added here, and were not present in the experiment.

4.3. Experiment Procedure

Participants were tasked with discerning the preferred charging station of a specific task's robot, understanding that while the robot could charge at either, it had a latent preference for one. Instructed that the robot also aimed to conserve energy, possibly choosing a less favored station if it were closer, participants rated the likelihood of the robot's preference for each station on a scale from 1 (very unlikely) to 5 (very likely). After making their likelihood assessment for the stations, they were presented with the next stimulus, with instructions to refine their inferences based on all previously shown images of the present task. Only one image was shown at any single time. Upon the task changing after five stimuli, participants were reminded that a new robot with different preferences was introduced.

For our model, we represented the tasks within a grid world that the RL agent needed to navigate. It incurred a minor negative penalty for movement and obtained positive rewards from both charging stations, determined by two specific parameters. The objective was to infer these parameters based on the observed data. We measured the discrepancy between observed and generated trajectories using Jaccard similarity. Essentially, our inference engine recreated the world as depicted in the stimulus, then ran the RL agent across varying parameters, comparing the generated trajectory against the observed one to form a posterior distribution for the two preferences. Preference likelihood ratings for the model were derived by computing the mean of the posterior distribution for preferences associated with both the blue and red charging stations.

4.4. Results

The preference ratings of each response were first standardized so that they sum up to 1. Then, a mean rating for each stimulus in each task was computed. The model's ratings were likewise standardized to sum up to 1, allowing comparison between human and model inferences. This comparison is shown in Figure 4. For calculating model fit, we selected only the inferences of the other color, because their values are inversions of each other after standadization. The model achieves a good fit, $R^2 = 0.78$, $RMSE = 0.1$. The most salient discrepancy between the model and human inferences is that the model is more careful in its estimates. Importantly, these results were obtained without any parameter tuning, meaning the model was not fit to the human data, but emerged similar data due to strong psychological assumptions about theory of mind.

The results exhibit the expected patterns of inference. Initially, participants faced uncertainty due to the limited evidence available. As they were exposed to additional stimuli, their inferences regarding the robot's preferences became more definite: one station's likelihood ratings increased, while the other's decreased. Task 1 serves as an example of this (Figure 3): the participants' inference that the robot prefers the red station gets stronger with each stimulus image shown. However, in tasks 3, 4, 6, 7, and 8, early stimuli suggested a certain preference, but subsequent stimuli revealed a stronger preference for the alternate station. This is similar to our motivating example in Figure 1. In these instances, the inferred preference for the more favored station shifted as the task progressed. Task 6 is an example of this (Figure 5): the participants are shown that the robot selects the red station, but it is always closer than the blue one, so there is uncertainty. Finally, in stimulus 5, it is revealed that the robot in fact prefers the blue station.

4.5. Discussion

Human-AI alignment necessitates that both humans and intelligent machines accurately interpret each other's inten-

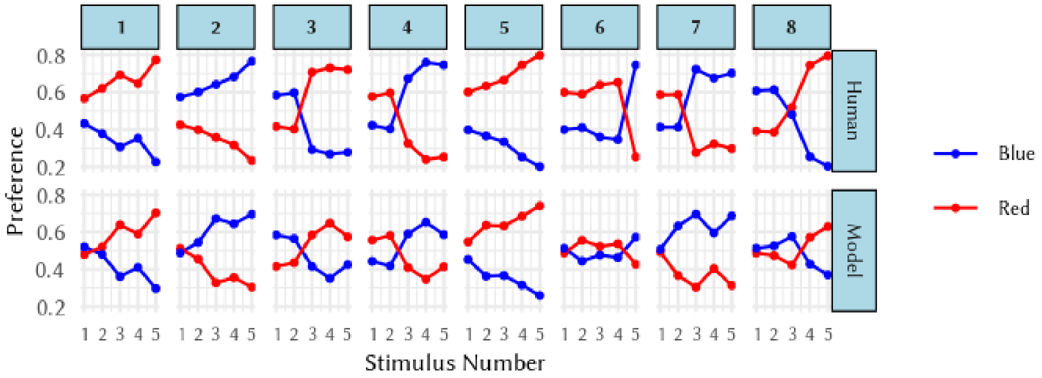

Figure 4: Comparison of model and human inferences across eight tasks. As more evidence accumulates, the inferences become more certain. Values close to 0.5 indicate high uncertainty, and values close to either 0 or 1 high certainty.

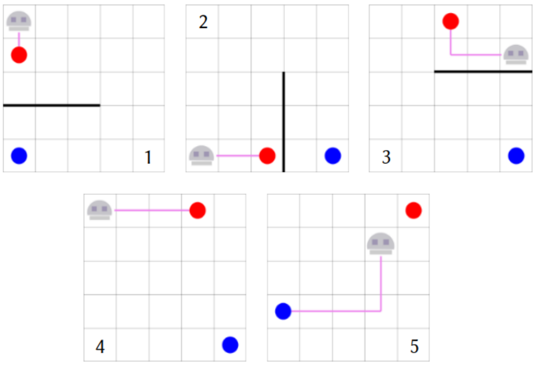

Figure 5: In Task 6, the participants only learned the true preference in the final image.

tions and actions [5]. This paper introduces a human-like theory of mind model capable of temporal observation integration, while being sensitive to uncertainty inherent in mentalizing. We validated the model's human-like inference capabilities through a grid world task focused on preference determination between two goals. The work carried here is theoretical in nature, and future studies should focus on more complex scenarios. While computational rationality has effectively modeled complex behaviors, such as multitasking while driving [32] and touchscreen typing [33], the exploration of long-term parameter inference in such contexts remains to be done.

Exploring decision-making under uncertainty is a large research topic. In our experiments, both humans and the model engaged in inferences and explicitly evaluated uncertainty, but they were not required to act on these inferences. A scenario where the model assists the observed actor will introduce the question of how to integrate uncertainty into decision-making. Taking the example of Janice from Figure 1, if adjusting the positions of the coffee and tea jars could aid her, the decision to do so necessitates careful consideration of potential consequences, ensuring the action truly benefits rather than hinders her. The manner in which a decision-making algorithm accounts for uncertainty during collaborative efforts is impacts the helpfulness of interventions and carries a risk of unintended obstruction.

All code, materials, and data are published online (https://version.helsinki.fi/shanz/quantifying-uncertainty-in-mtom.git) to facilitate open science.

Acknowledgments

This research has been supported by the Academy of Finland (grant 330347).

References

[1] E. Etel, V. Slaughter, Theory of mind and peer cooperation in two play contexts, Journal of Applied Developmental Psychology 60 (2019) 87–95.

[2] T. Paal, T. Bereczkei, Adult theory of mind, cooperation, machiavellianism: The effect of mindreading on social relations, Personality and individual differences 43 (2007) 541–551.

[3] M. I. Brown, A. Ratajska, S. L. Hughes, J. B. Fishman, E. Huerta, C. F. Chabris, The social shapes test: A new measure of social intelligence, mentalizing, and theory of mind, Personality and Individual Differences 143 (2019) 107–117.

[4] J. F. Kihlstrom, N. Cantor, Social intelligence. (2000).

[5] S. Russell, Human compatible: Artificial intelligence and the problem of control, Penguin, 2019.

[6] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, S. J. Gershman, Building machines that learn and think like people, Behavioral and brain sciences 40 (2017).

[7] I. Cho, N. Kamkar, N. Hosseini-Kamkar, Reasoning about mental states under uncertainty, PloS one 17 (2022) e0277356.

[8] O. FeldmanHall, A. Shenhav, Resolving uncertainty in a social world, Nature human behaviour 3 (2019) 426–435.

[9] C. L. Baker, J. Jara-Ettinger, R. Saxe, J. B. Tenenbaum, Rational quantitative attribution of beliefs, desires and percepts in human mentalizing, Nature Human Behaviour 1 (2017) 1–10.

[10] S. Liu, T. D. Ullman, J. B. Tenenbaum, E. S. Spelke, Ten-month-old infants infer the value of goals from the costs of actions, Science 358 (2017) 1038–1041.

[11] H. Richardson, G. Lisandrelli, A. Riobueno-Naylor, R. Saxe, Development of the social brain from age three to twelve years, Nature communications 9 (2018) 1–12.

[12] I. Dziobek, S. Fleck, E. Kalbe, K. Rogers, J. Hassenstab, M. Brand, J. Kessler, J. K. Woike, O. T. Wolf, A. Convit, Introducing masc: a movie for the assessment of so-

cial cognition, Journal of autism and developmental disorders 36 (2006) 623–636.

[13] R. Markiewicz, F. Rahman, I. Apperly, A. Mazaheri, K. Segaert, It is not all about you: Communicative cooperation is determined by your partner's theory of mind abilities as well as your own., Journal of Experimental Psychology: Learning, Memory, and Cognition (2023).

[14] M. Harbers, K. Van Den Bosch, J.-J. Meyer, Modeling agents with a theory of mind, in: 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, volume 2, IEEE, 2009, pp. 217–224.

[15] S. Devin, R. Alami, An implemented theory of mind to improve human-robot shared plans execution, in: 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), IEEE, 2016, pp. 319–326.

[16] K.-J. Kim, H. Lipson, Towards a simple robotic theory of mind, in: Proceedings of the 9th workshop on performance metrics for intelligent systems, 2009, pp. 131–138.

[17] S. Lin, B. Keysar, N. Epley, Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention, Journal of Experimental Social Psychology 46 (2010) 551–556.

[18] Q. Wang, K. Saha, E. Gregori, D. Joyner, A. Goel, Towards mutual theory of mind in human-ai interaction: How language reflects what students perceive about a virtual teaching assistant, in: Proceedings of the 2021 CHI conference on human factors in computing systems, 2021, pp. 1–14.

[19] L. M. Hiatt, A. M. Harrison, J. G. Trafton, Accommodating human variability in human-robot teams through theory of mind, in: Twenty-second international joint conference on artificial intelligence, 2011.

[20] J. Jara-Ettinger, L. E. Schulz, J. B. Tenenbaum, The naive utility calculus as a unified, quantitative framework for action understanding, Cognitive Psychology 123 (2020) 101334.

[21] P. Shafto, N. D. Goodman, M. C. Frank, Learning from others: The consequences of psychological reasoning for human learning, Perspectives on Psychological Science 7 (2012) 341–351.

[22] Jokinen, Remes, Kujala, Corander, Bayesian parameter inference for cognitive simulators, in: J. Williamson, A. Oulasvirta, P. Kristensson, N. Banovic (Eds.), Bayesian methods for interaction design, Cambridge University Press, 2022.

[23] C. L. Baker, R. Saxe, J. B. Tenenbaum, Action understanding as inverse planning, Cognition 113 (2009) 329–349.

[24] A. Kangasrääsiö, J. P. Jokinen, A. Oulasvirta, A. Howes, S. Kaski, Parameter inference for computational cognitive models with approximate bayesian computation, Cognitive science 43 (2019) e12738.

[25] R. L. Lewis, A. Howes, S. Singh, Computational rationality: Linking mechanism and behavior through bounded utility maximization, Topics in cognitive science 6 (2014) 279–311.

[26] A. Oulasvirta, J. P. Jokinen, A. Howes, Computational rationality as a theory of interaction, in: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 2022, pp. 1–14.

[27] A. Howes, J. P. Jokinen, A. Oulasvirta, Towards machines that understand people, AI Magazine 44 (2023) 312–327.

[28] J. X. O'reilly, Making predictions in a changing world—inference, uncertainty, and learning, Frontiers in neuroscience 7 (2013) 105.

[29] O. FeldmanHall, M. R. Nassar, The computational challenge of social learning, Trends in Cognitive Sciences 25 (2021) 1045–1057.

[30] M. U. Gutmann, J. Corander, et al., Bayesian optimization for likelihood-free inference of simulator-based statistical models, Journal of Machine Learning Research (2016).

[31] J. Lintusaari, H. Vuollekoski, A. Kangasrääsiö, K. Skytén, M. Järvenpää, P. Marttinen, M. U. Gutmann, A. Vehtari, J. Corander, S. Kaski, Elfi: Engine for likelihood-free inference, Journal of Machine Learning Research 19 (2018) 1–7.

[32] J. P. Jokinen, T. Kujala, A. Oulasvirta, Multitasking in driving as optimal adaptation under uncertainty, Human factors 63 (2021) 1324–1341.

[33] J. Jokinen, A. Acharya, M. Uzair, X. Jiang, A. Oulasvirta, Touchscreen typing as optimal supervisory control, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–14.