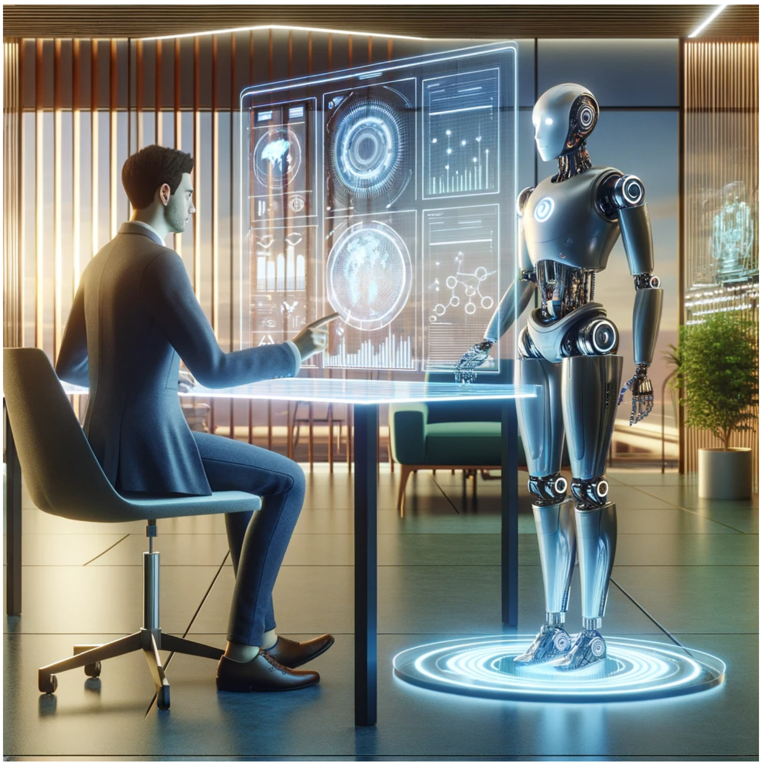
Report on ‘Bringing Algorithms to Practice - Experiments to Aid Valorization’ (2018-4-143TKI)



Willem van Jaarsveld[1], Eirini Spiliotopoulou[2], Lijia Tan[3]

[1] Eindhoven University of Technology, w.l.v.jaarsveld@tue.nl

[2] Tilburg University, e.spiliotopoulou@tilburguniversity.edu

[3] Eindhoven University of Technology, l.tan1@tue.nl

February 1, 2024

Contents

List of Figures

List of Tables

1 Introduction

In the contemporary landscape of decision-making, the integration of advanced artificial intelligence (AI) technologies, specifically neural network algorithms, presents a frontier of immense potential and complexity. This project focuses on the pivotal interactions between human planners and AI recommendations generated by sophisticated neural network algorithms. The impetus for this inquiry is anchored in two critical considerations.

Firstly, the project aims to dissect the dynamics of decision-making in scenarios where human experience and intuition intersect with the data-driven insights provided by neural networks. Neural networks, with their unparalleled ability to analyze vast datasets and identify patterns beyond human capability, represent a pinnacle of AI development. However, their integration into human decision-making processes is a nuanced and intricate phenomenon. This study endeavors to unravel these complexities, offering insights into how human planners interpret, trust, and utilize the advanced recommendations generated by neural networks.

Secondly, this exploration is driven by the necessity to understand how the use of such cutting-edge AI tools influences the behavior and choices of human decision-makers. The interaction between human cognition and machine intelligence is not merely additive; it is transformative. The potential of neural network algorithms to enhance decision-making processes in terms of efficiency, precision, and innovation is immense. However, the role of the human planner in interpreting, contextualizing, and ultimately making decisions based on these AI recommendations is equally crucial.

This project, therefore, seeks to address a gap in the current understanding of human-AI collaboration: how do advanced neural network recommendations influence human decision-makers, and what are the implications of this interplay for decision quality and the future of AI in decision-making? By delving into these questions, the project aims to offer a comprehensive analysis of the synergies and frictions in human-neural network collaborations, highlighting the importance of human oversight in an increasingly AI-driven world.

In essence, this project is motivated by the urgent need to unravel the complexities of human decision-making in the age of advanced neural network algorithms. It aspires to contribute significantly to the discourse on optimizing human-AI collaboration, recognizing the transformative impact such collaboration has on various sectors and the critical importance of maintaining a balance between machine efficiency and human judgment.

Building upon the motivation laid out, this report unfolds through a series of meticulously structured sections, each focusing on a distinct aspect of the interplay between human decision-making and AI recommendations in the context of supply chain management.

Section 2 introduces the practical application of our study - a supply chain game designed as a testbed for human-AI interaction. This section details the AI decision support tool, developed using advanced neural network algorithms, and explains how it integrates into the supply chain game, providing recommendations to human participants.

Section 3 outlines the design of the experiment, including participant selection, data collection methods, and the specific metrics used to evaluate the effectiveness of AI recommendations in influencing human decision-making within the game.

Section 4 summarizes the behavioral patterns of human participants when interacting with the AI tool, highlighting how their decisions were influenced by AI recommendations and identifying key trends and insights from the data.

Section 5 discusses the implications of the findings for future applications of AI in supply chain management and decision-making. This section also suggests avenues for further research, considering the limitations of the current study, and concludes with a summary of the overarching insights and their significance in the broader context of human-AI collaboration.

Each section is designed to build upon the previous ones, leading to a comprehensive understanding of how neural network-based AI recommendations can shape human decision-making in complex, real-world scenarios.

2 Supply chain game and decision support embedded algorithms

2.1 Supply chain game

We develop a digital supply chain planning game as Figure 1 depicts. The planner makes decisions regarding manufacturing products to fulfill the demands from three markets: cellphone market in the NL, cellphone market in DE, and game computer market in the NL.

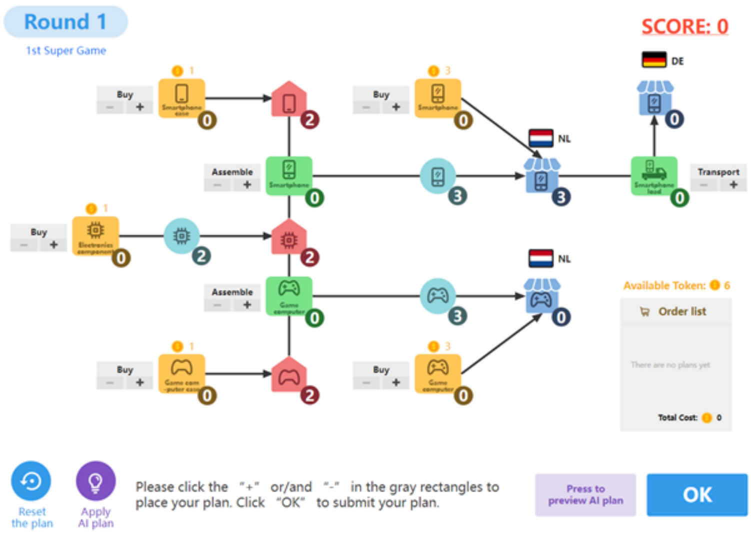


Figure 1: Decision interface in the supply chain game

The planner makes three types of decisions in each round – ordering, assembling, and shipping.

- Ordering - the planner has 6 tokens to order either product components (needed for assembling a product) or finished products.
- Assembling - once the components arrive in the warehouse, the planner can assemble them into a final product (a smartphone or a game computer).
- Shipping - When there are smartphones available on the NL market), the planner can ship some (<= the inventory in NL) to the DE market.

While the planner is making these decisions, he does not know the demand of the current round. The demand is realized in the end of the round, which is about which market requests how many products (randomly drawn from 1-6). After the planner submits his decisions, Figure 2 appears informing the planner of the realized demand. The game lasts 10 rounds.

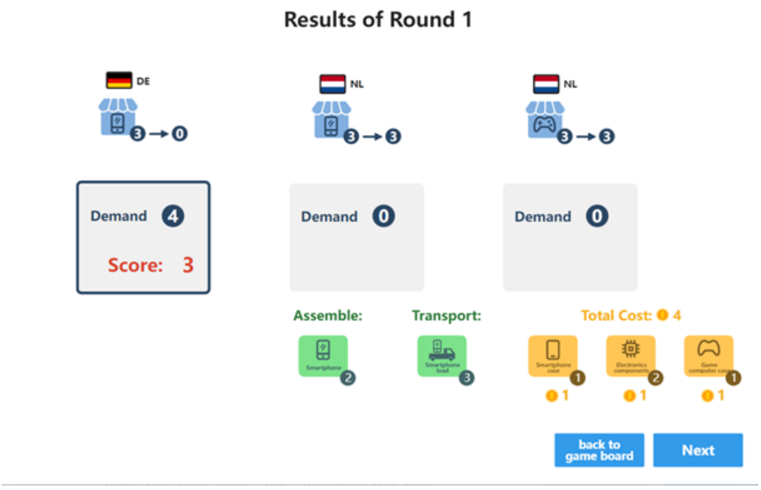


Figure 2: Market demand after planning decision

2.2 Algorithm-embedded decision support tool

While the planner is making decisions, the game provides him a suggestion, so called AI plan. He can click the "apply AI plan" button to apply the AI plan in the current round, see the left bottom corner in Figure 1. He can also press the button on the right bottom to preview AI plan. The AI plan is based on a neural network and reinforcement learning algorithm. We trained a model via trial and error methods/reinforcement learning. and employed DynaPlex/DCL. Figure 3 depicts the three steps.

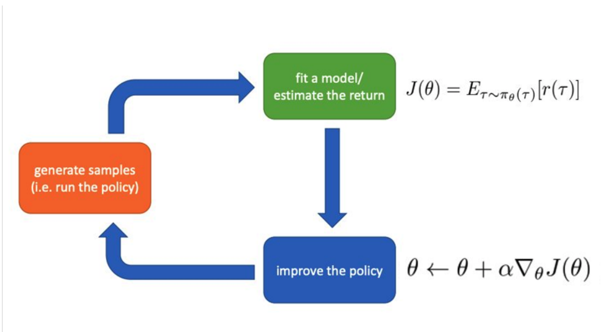


Figure 3: Policy gradient

The first step is to generate samples by using the policy (a neural network) in the environment. The second step is to keep track of the rewards obtained. And the third step is to perform a gradient descent step on the policy. The trained algorithm performs quite well in the context: it clearly outperformed PPO (a well-known benchmark), it outperformed about 50 human subjects in semi-controlled trial, where the algorithm was fed the same demand sequence as the humans, and it outperforms several well-known heuristics.

3 Experiment design and implementation

We conducted controlled incentivized lab experiments with humans in the role of planners in the supply chain game outlined in the previous section. We ran 8 sessions in total for 4 treatments at Tilburg University in Spring 2023. The treatments were between-subjects which did not allow a subject to attend more than one treatment. In total 150 subjects participated in the experiment. They received cash payment and the amount depended on their score in the experiment. Each session lasted about 1.5 hours.

3.1 Treatments

In the experiment, each subject played three supply chain games. Each supply game consists of 10 rounds. The first game is referred to as Stage I, the second game is Stage II, and the third is Stage III in Table 1. In Stage I, subjects play the supply chain game without the support of the AI plan. In Stages II and III, the subjects in relevant treatments get the support of the AI plan, which depends on the treatment.

Table 1: Treatments completed in the spring of 2023

Treatment	Stage I	Stage II	Stage III	# of subjects
Baseline	No AI	No AI	No AI	34
Learning	No AI	AI	No AI	39
Experienced	No AI	No AI	AI	38
AI-able	No AI	AI	AI	39

As we shown in Table 1, there are four treatments related to when the subjects can get the support of the AI plan.

- **Baseline** treatment consists of three supply chain games in which AI plan is not present.
- **Learning** treatment is about how subjects learn from the AI plan. The AI plan is only present in Stage II.
- **Experienced** treatment is to observe how the AI plan can help the experienced subjects. The AI plan is only present in Stage III.
- **AI-able** treatment provides subjects the AI plan in Stage II and Stage III.

3.2 Prior experiment survey

Before the experiment starts, we conduct a survey which consists of risk attitude measures, cognitive measures, and attitudes toward AI/algorithmic tool measures to capture individual characteristics of the participating subjects.

We use the lottery choice test from Holt and Laury (2002) to measure risk attitude. We also measure subjects' cognitive reflection by administering a cognitive reflection test (CRT): a comprehensive measure including reasoning, perception, memory, verbal, mathematical ability, and problem solving. The CRT consists of three questions (Frederick, 2005). Last, we measure subjects' attitude toward AI/algorithmic tools following the technology acceptance model (TAM). Its questionnaire intends to capture subject's attitude and use of technology which we adapted to focus on AI/algorithmic tools. We focused on the following determinants of TAM, that are relevant in our context: (1) perceived usefulness, (2) perceived ease of use, (3) attitude towards use, (4) intention to use and (5) actual use. In total, there are 11 questions. Each question has a scale from 1(strongly disagree) to 5 (strongly agree).

4 Results and behavioral insights

4.1 Do people use/trust the AI tool?

First we focus on whether and how human decision-makers use the AI tool in planning, when available. Do people just apply the recommendations or do they modify them? When looking at the plan implemented in a round, in around half of the cases the AI plan is implemented "as-is". The implies that decision makers largely use the AI tool (even if it is "black-box"), but they also very frequently modify it. Looking at each decision separately, Table 2 presents how often decision makers decision makers deviate from AI recommendations per decision and Game. We observe more deviations for buying (component) decisions and, more interestingly, a drop in the frequency of deviations in Game 3.

	Deviations from AI recommendations			
Decision	Avg	Frequency	Game 2	Game 3
Buy smartphone case	0,26	18,5%	20,8%	16,2%
Buy electronics	0,55	25,3%	28,5%	22,1%
Buy computerccase	0,18	14,2%	17,2%	11,2%
Buy smartphone	0,18	15,4%	15,5%	15,3%
Buy computer	0,11	9,2%	9,4%	9,1%
Assemble smartphone	0,11	8,8%	11,3%	6,4%
Assemble computer	0,10	8,8%	10,6%	6,9%
Total		*14,3%*	*16,2%*	*12,4%*

Table 2: Deviations from AI plan per decision and game

To shed light on what drives these deviations, or alternatively (mis-)trust in AI recommendations, we estimate the impact of individual level characteristics (i.e., attitude towards AI technology, risk preferences, cognitive reflection, demographics) and task experience (i.e., round and game number) on subject's decision to algorithmic modifications. Table 4 summarizes the results of the GLS panel data regression with errors clustered at the subject level. We find that the general Attitude towards AI/algorithmic tools (as measured by our questionnaire) decreases deviations from the AI plan, i.e., it impacts positively decision makers' adherence to algorithmic recommendations. Similarly, both the round and the game number also have a negative impact on deviations. This implies that overall experience with the task (within a game and across games) positively impacts trust in AI recommendations and use of the AI tool. While we control for demographic characteristics (age, gender, level of education, field of study, work experience), none of these factors seem to have a significant impact on decisions in this context.

Deviation AI	Coefficient	St.err.	p-value
TAM score	-0.61	0.0232	0.008
CRT score	-0.16	0.1129	0.169
Risk	0.01	0.0734	0.883
Round	-0.16	0.0192	0.000
Game	-0.38	0.1395	0.006
Constant	4.963	1.9026	0.009
Demographics	Yes		
sigma u	1.2048		
sigma e	2.1563		

Table 3: Drivers of deviations from AI plan

4.2 Does the AI tool help improve performance?

Next we shift our attention to whether the AI tool improves performance. We first establish that, in line with our expectations, larger adherence to algorithmic recommendations, i.e., AI plan, improves performance. As shown in Table **??**, participants' score in a round decreases in the size of deviations from AI recommendations, controlling for the round and demand (similar results hold for the total game score). This implies that lower trust in AI has a negative effect on performance: subjects who use the AI plan more make better decisions on average.

Per round score	Coefficient	St.err.	p-value
Cum. dev. AI	-0.005	0.0017	0.003
Round	0.091	0.0076	0.000
Demand	0.682	0.0123	0.000
Constant	0.104	0.0536	0.053

Table 4: Drivers of per round score

Next we focus on whether the availability of the AI tool increases performance on average. We established that while subjects largely use the AI plan, when available, they often modify it, and this behavior has negative implications for performance. Therefore, does making the AI tool available significantly improve performance on average?

Table 5 shows the average total score per Game and treatment. (The light blue colour indicates the AI plan availability.) In Game 1, there is no AI plan available in any of the treatmens and averahe performance is not different (Wilxocon rank-sum tests at the subject-level, $p > 0.1$). However, in Game 2, total score is significantly higher under the AI-able treatment and the baseline ($p = 0.0310$) but that is not the case between Learning and Baseline ($p > 0.1$). In Game 3, we observe no statistically significant differences among treatments. These results suggest that having the AI tool available does not necessarily increase performance and that an AI tool seems to be important only for less experienced users.

	Total Score								
	Game 1			Game 2			Game 3		
	Avg	Median	St.dev	Avg	Median	St.dev	Avg	Median	St.dev
Baseline	22,6	23,0	3,8	24,8	25,0	2,7	25,5	26,5	3,1
Learning	22,7	23,0	3,3	24,3	25,0	3,1	25,3	25,0	3,1
Experienced	23,2	24,0	3,2	24,0	24,0	3,1	26,2	26,5	2,9
AI-Able	23,1	23,0	3,2	26,2	27,0	2,8	25,6	26,0	3,4

Table 5: Total score per game and treatment

Since there are different demand realizations across games and treatments, we also calculate the fill rate achieved in a game as an alternative measure of performance. We define aggregate fill rate as the ratio of total points achieved in the game over total demand. Table 6 summarizes the results.

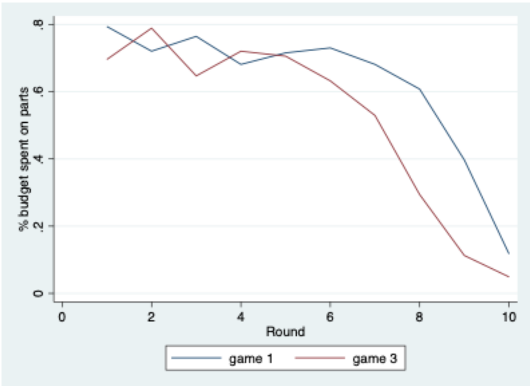
	Aggregate fill rate								
	Game 1			Game 2			Game 3		
	Avg	Median	St.dev	Avg	Median	St.dev	Avg	Median	St.dev
Baseline	79,87%	79,66%	13,00%	80,96%	81,06%	8,70%	83,16%	84,12%	8,39%
Learning	77,18%	76,19%	10,67%	85,55%	85,71%	7,99%	85,44%	85,71%	8,88%
Experienced	74,93%	74,24%	9,64%	84,08%	85,02%	10,88%	85,81%	84,78%	7,99%
AI-Able	79,40%	79,17%	8,71%	85,96%	86,21%	9,26%	87,18%	85,71%	6,23%

Table 6: Fill rate per game and treatment

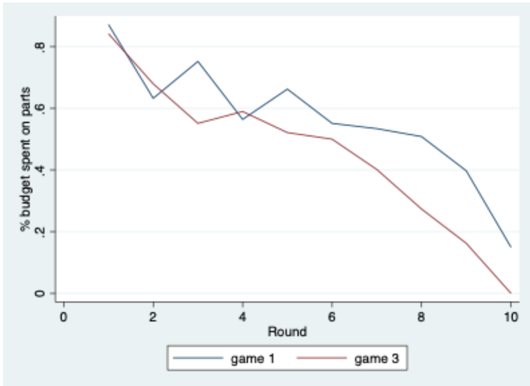
Across all games, fill rates are higher when the AI tool is available ($p = 0.0000$). If we zoom in at the game number, fill rates are higher when AI tool is present for Game 2 ($p = 0.0589$) but this is not the case for Game 3's (0.2544). This is in line with the notion that the AI tool is beneficial for less experienced users.

4.3 Does the AI tool help humans learn?

Last, we turn our attention on whether humans perform better (without an AI tool), after they had access to an AI tool in a previous game (i.e., does the AI tool facilitate learning?). We look at planners' performance without AI when in the prior game they had access to such a tool and the possibility to observe AI recommendations. To account for differences in skills, we look at subject's performance increase between Game 1 and Game 3 when in Game 2 AI was available and when it was not (Baseline versus Learning treatments). We are not able to detect a significant overall effect of AI on learning at the aggregate level ($p > 0.1$), considering either total score or fill rate as a measure of performance. Next, we look closer at specific strategies to assess whether the exposure to AI plans helps human decision makers learn specific strategies that are effective in this complex setting but (maybe) non-intuitive. Specifically, we first look at whether planners learn to buy more finished products towards the end of the game. In figure 4 we plot the proportion of tokens spent on components per round in Game 3 versus Game 1. The difference between games does not seem to depend on the treatment. Last, we look at whether planners learn to take into account the



(a) Baseline treatment



(b) Learning treatment

Figure 4: The proportion of tokens spent on components

delay in the sourcing of electronic components (EC). We calculate the average % of budget spent for buying EC in the first 3 rounds and observe no significant differences between the Baseline and Learning treatments (Figure 5).

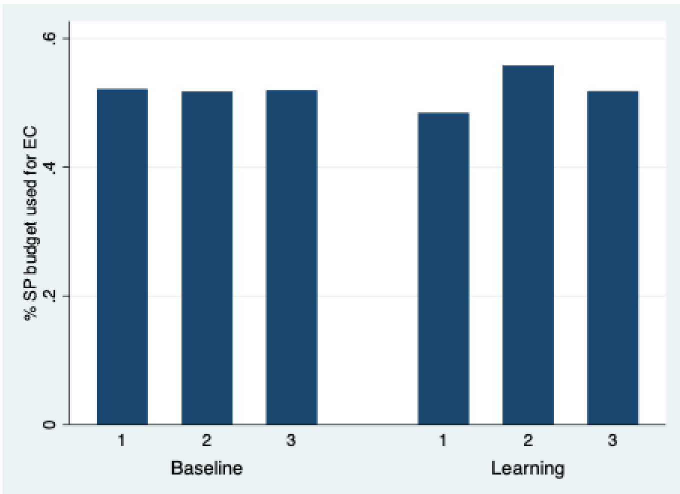


Figure 5: The proportion of tokens spent on EC in baseline and learning treatments

5 Conclusion and the next research plan

To conclude, we observe that in this complex setting human decision makers largely use the AI tool (even if the information provided about the algorithm is limited) but the also very frequently modify it, which decreases performance. Participants' general attitude towards AI and algorithms, as well as their experience with the task affect trust in AI recommendations, or alternatively, deviations from the suggested AI plan. Overall, the availability of an AI tool, taking into account how human planers use it/modify it, improves the quality of human decisions (measured either by total score or achieved fill rate) but only for less experienced users (Game 2). In Game 3, performance is not significantly different for planners who have an AI plan available versus not. Having more experience with the task also helps improve performance, suggesting that experience and the AI tool can be substitutes in this context.

While the task is complex, humans seem to learn simple and effective strategies with task experience. However, the experimental results raised a new question.

Are subjects able to learn and perform well in a more complex setting?

For example, one kind of complexity can be decision-makers under time pressure to finish the planning decisions. This presents an exciting opportunity for future research. Furthermore, adherence to AI recommendations increases performance in this context. Can we nudge decision makers to make better use of the AI plan by increasing transparency about its past performance (outcome transparency), providing more information about the underlying algorithmic tool (operational transparency) or training participants (task knowledge) to better understand AI recommendations?

References

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic perspectives*, 19(4):25–42.

Holt, C. A. and Laury, S. K. (2002). Risk aversion and incentive effects. *American economic review*, 92(5):1644–1655.