Affective Neuroscience and Decision-making Lab

@ANDlab3

Memory and Emotion Lab

ME LAB

Towards human-compatible autonomous car: A study of non-verbal Turing test in automated driving with affective transition modelling

自动驾驶图灵测试中的情感计算初探

Presenter: Zhaoning Li 李肇宁

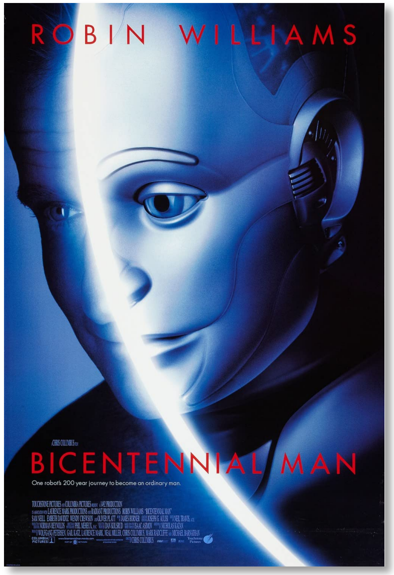
National Doctoral Forum on Brain-Computer Intelligence and Psychology, Zhejiang University, 2022

PROLOGUE

‘Well, I'm human in part.’ “你哪部分是人类？”

... ‘Which part, Andrew?’ “这里，我的心！”

... ‘My mind. My heart. I may be artificial, alien, inhuman so far as your strict genetic definition goes. But I’m human in every way that counts. And I can be recognised as such legally.’



(Adapted from IMDb)

ISAAC ASIMOV AND ROBERT SILVERBERG – THE POSITRONIC MAN

BACKGROUND

- **Autonomous cars (AC) have the potential to increase road safety, as they can react faster than human drivers and are not subject to human errors.**

- **Despite the potential benefits, there has yet to be a large-scale deployment of ACs.**

- **One main obstacle is that these cars are not humanoid, i.e., they are not driving in a human-like manner.**

- **Existing literature highlights that the acceptance of AC will increase if it drives in a human-like manner.**

- **However, sparse research offers the true-to-life ride experience as a passenger in the AC that examines the human likeness of the AC.**

RESEARCH QUESTION

RQ1: How to offer the naturalistic experience from a passenger's seat perspective to measure the human likeness of current autonomous cars?

RESEARCH QUESTION

How to offer the naturalistic experience from a passenger's seat perspective to measure the human likeness of current autonomous cars?

(Adapted from Wikipedia)

Father of computer science and AI

In 1950, Alan Turing proposed the Turing test [1] to evaluate the **ascription of intelligence**, i.e., whether humans would ascribe human-like intelligent behaviour to machines.

1. A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, no. 236, 1950.

RESEARCH QUESTION

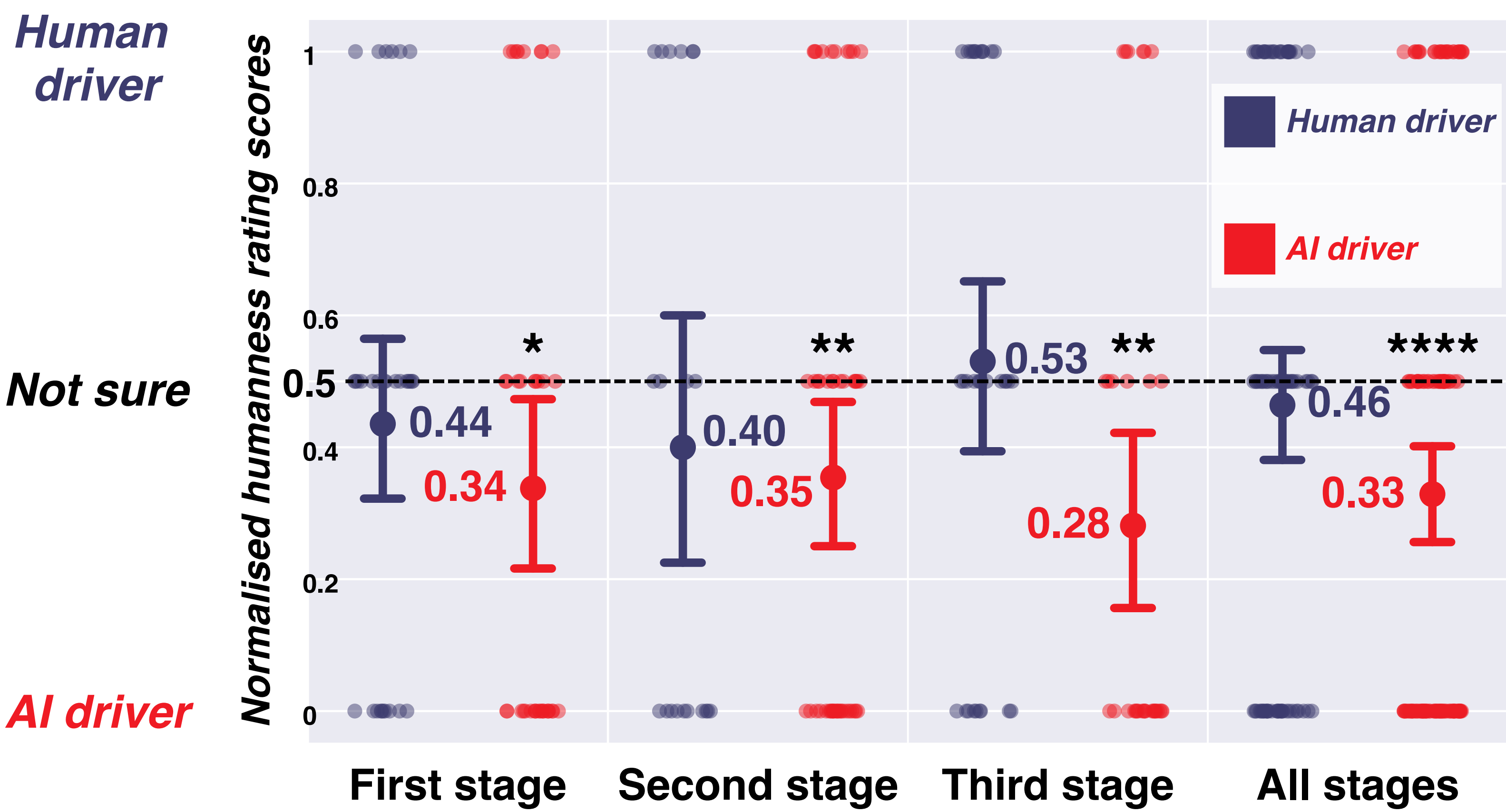
How to offer the naturalistic experience from a passenger's seat perspective to measure the human likeness of current autonomous cars?

We designed a ride experience-based version of the non-verbal Turing test to evaluate the **ascription of humanness**, i.e., whether the AI driver could create a human-like ride experience for passengers, such that passengers would have either chance-level or even higher humanness ratings under the AI driver condition.

THE NON-VERBAL VARIATION OF THE TURING TEST

RESULTS OF THE NON-VERBAL VARIATION OF THE TURING TEST

Normalised humanness rating scores, their mean values and 95% confidence intervals (CI) under different conditions



The AI driver failed to pass our test because passengers detected the AI driver above chance.

RESEARCH QUESTION

- The AI driver’s failure inspired us to explore further why the AI algorithm could trick human passengers in some trials and not in most others.

RESEARCH QUESTION

RQ2: How do human passengers ascribe humanness in the non-verbal variation of the Turing test?

RESEARCH QUESTION

How do human passengers ascribe humanness in the non-verbal variation of the Turing test?

(Adapted from Wikipedia)

Father of modern social psychology

Lewin's Field theory [2] states that a person's psychological field (i.e., the total psychological environment that the person experiences subjectively) determines their behaviour, which can be expressed by the following equation:

Pearson: passenger

Environment: driving environment

Behaviour: humanness rating behaviour

$$B = f\underbrace{(P, E)}_{\text{Psychological field}}$$

2. K. Lewin, *Principles of Topological Psychology*. McGraw-Hill, 1936.

RESEARCH QUESTION

How do human passengers ascribe humanness in the non-verbal variation of the Turing test?

(Adapted from Wikipedia)

Father of modern social psychology

Lewin's Field theory [2] states that a person's psychological field (i.e., the total psychological environment that the person experiences subjectively) determines their behaviour, which can be expressed by the following equation:

Pearson: passenger

Environment: driving environment

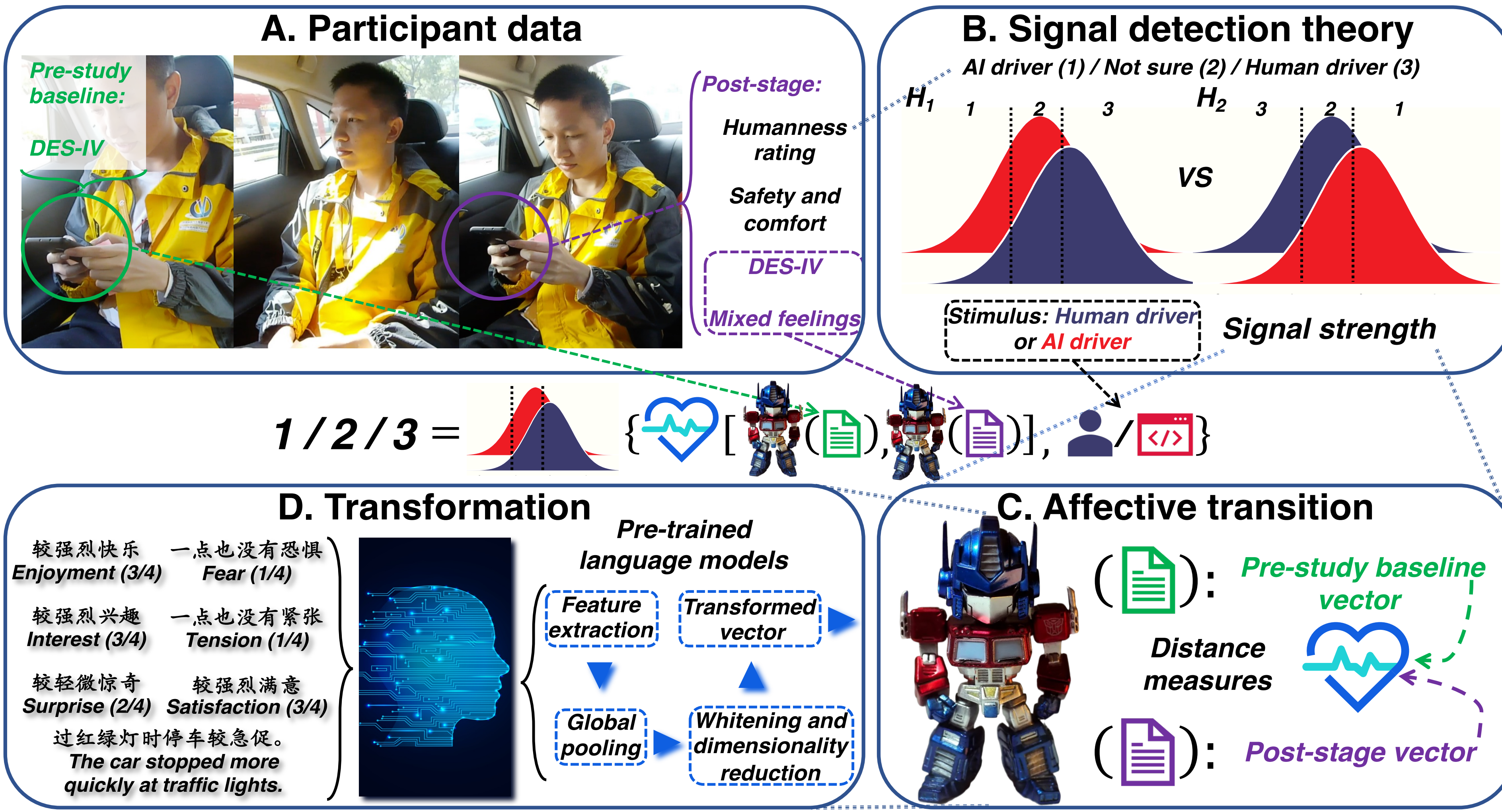
Behaviour: humanness rating behaviour

$$B = f\underbrace{(P, E)}$$

Psychological field: passenger's subjective ride experience

2. K. Lewin, *Principles of Topological Psychology*. McGraw-Hill, 1936.

COMPUTATIONAL MODELLING



RESULTS OF THE COMPUTATIONAL MODELLING

Comparisons on the outer loop cross-validation of nested-LOOCV with baselines

(a) Evaluation results on the first stage.

‘AA’ for all affect,
‘PA’ for positive affect,
‘NA’ for negative affect and
‘MF’ for mixed feelings
‘Pre’ for pre-study baseline
and ‘post’ for post-stage

Baselines	AA	AA_{pre}	AA_{post}	PA	PA_{pre}	PA_{post}	NA	NA_{pre}	NA_{post}
MLR	-0.1844	0.1312	0.1283	0.0988	0.1761	-0.0082	-0.0453	0.0390	0.0744
KNN	0.1431	0.0543	0.1753	0.4755****	0.2370*	-0.0669	0.0870	-0.1078	0.1129
SVC	-0.1039	-0.1027	-0.0268	0.1704	0.0431	-0.0932	0.0780	0.0340	-0.0578
RF	-0.0654	0.1239	-0.0122	0.1125	0.1245	-0.2744	0.0688	0.0586	0.1301
XGBoost	0.1794	0.4125***	0.0537	0.2188*	0.0754	0.0430	0.1013	0.1508	0.1321
MLP	0.2185*	0.3211**	-0.1391	-0.0759	0.1083	0.0953	0.0448	-0.1041	0.0342
Baselines	*None*	**SDT-AT**	*AA+MF*	*AA*	***PA+MF***	*PA*	*NA+MF*	*NA*	*MF*
Random	0.0029	Original	-0.3985	-0.3552	-0.2580	0.1738	-0.3397	0.0828	0.0990
Probability	-0.0060	PLM-wv	0.4511***	0.4152***	0.4092***	0.3939***	0.4064***	0.1359	0.3030**
Detective	0.1491	**PLM-tf**	0.4113***	0.4639****	**0.4768******	0.3939***	0.3484**	0.1842	0.3738**

RESULTS OF THE COMPUTATIONAL MODELLING

Comparisons on the outer loop cross-validation of nested-LOOCV with baselines

(a) Evaluation results on the first stage.

Baselines	AA	AA_{pre}	AA_{post}	PA	PA_{pre}	PA_{post}	NA	NA_{pre}	NA_{post}
MLI									
KNI									
SVC									
RF									
XGBo									
MLI									
Baseli									
Rand									
Probab									
Detect									

‘AA’ for all affect, ‘PA’ for positive affect, ‘NA’ for negative affect and ‘MF’ for mixed feelings

‘Pre’ for pre-study baseline and ‘post’ for post-stage

(b) Evaluation results on the second stage.

Baselines	AA	AA_{pre}	AA_{post}	PA	PA_{pre}	PA_{post}	NA	NA_{pre}	NA_{post}
MLR	0.2752*	0.1524	-0.2298	0.1539	0.2095*	-0.1659	0.0205	0.1947	-0.1728
KNN	0.2046*	0.3069**	-0.3189	0.1436	0.1297	-0.3123	-0.2696	-0.1486	-0.1639
SVC	0.1061	0.0945	-0.1743	0.1270	-0.0558	-0.0776	0.0161	0.0541	0.0997
RF	0.0416	0.3126**	-0.1799	0.2379*	0.2588*	-0.2196	0.0573	0.2087*	-0.3861
XGBoost	0.0835	0.2839**	-0.2254	0.1895	0.3613**	-0.1368	-0.0965	-0.2473	-0.1788
MLP	0.1986	0.1981	-0.3661	0.1302	0.3687**	-0.1213	-0.0608	-0.3048	-0.3838

Baselines	*None*	**SDT-AT**	*AA+MF*	*AA*	*PA+MF*	***PA***	*NA+MF*	*NA*	*MF*
Random	0.0010	Original	0.1750	0.2409*	0.1539	0.1912	0.1865	-0.0105	0.1824
Probability	-0.0017	PLM-wv	0.4569****	0.4195***	0.4402***	0.4635****	0.3167**	0.1703	0.4276***
Detective	0.0394	**PLM-tf**	0.4375***	0.4173***	0.4545****	**0.4739********	0.3528**	0.2636*	0.3578**

RESULTS OF THE COMPUTATIONAL MODELLING

Comparisons on the outer loop cross-validation of nested-LOOCV with baselines

‘AA’ for all affect,
‘PA’ for positive affect,
‘NA’ for negative affect and
‘MF’ for mixed feelings
‘Pre’ for pre-study baseline
and ‘post’ for post-stage

(a) Evaluation results on the first stage.

Baselines	AA	AA_{pre}	AA_{post}	PA	PA_{pre}	PA_{post}	NA	NA_{pre}	NA_{post}
MLI									
KNI									
SVC									
RF									
XGBo									
MLI									
Baseli									
Rand									
Probab									
Detect									

(b) Evaluation results on the second stage.

Baselines	AA	AA_{pre}	AA_{post}	PA	PA_{pre}	PA_{post}	NA	NA_{pre}	NA_{post}
ML									
KN									
SVC									
RF									
XGBc									
ML									
Baseli									
Rand									
Probal									
Detec									

(c) Evaluation results on the third stage.

Baselines	AA	AA_{pre}	AA_{post}	PA	PA_{pre}	PA_{post}	NA	NA_{pre}	NA_{post}
MLR	0.2154*	0.3482**	0.2852*	0.0593	-0.0535	0.0076	0.3994***	0.3294**	0.3954***
KNN	0.1782	0.4317***	0.2630*	0.0885	0.1510	0.1899	0.3998***	0.4161***	0.3301**
SVC	0.1425	0.3438**	0.2218*	-0.0157	-0.0608	0.1165	0.1932	0.1456	0.3215**
RF	0.1180	0.3615**	0.0360	0.0654	0.1642	0.0294	0.3397**	0.2815*	0.3244**
XGBoost	0.2186*	0.3625**	0.1942	0.0674	0.1525	0.1175	0.3339**	0.4016***	0.2987**
MLP	0.1302	0.2144*	0.2740*	0.0347	0.0722	0.2187*	0.3674**	0.3126**	0.2512*
Baselines	*None*	**SDT-AT**	*AA+MF*	*AA*	*PA+MF*	*PA*	*NA+MF*	*NA*	***MF***
Random	0.0001	Original	0.1490	0.2019	0.1978	-0.0258	0.4037***	0.4245***	0.1104
Probability	-0.0021	**PLM-wv**	0.4861****	0.4556***	0.4624***	0.4322***	0.4419***	0.4256***	**0.5615******
Detective	0.3168**	PLM-tf	0.4807****	0.4974****	0.4654****	0.4570***	0.4769****	0.4429***	0.5422****

RESULTS OF THE COMPUTATIONAL MODELLING

Comparisons on the outer loop cross-validation of nested-LOOCV with baselines

‘AA’ for all affect,
‘PA’ for positive affect,
‘NA’ for negative affect and
‘MF’ for mixed feelings
‘Pre’ for pre-study baseline
and ‘post’ for post-stage

(a) Evaluation results on the first stage.

Baselines	AA	AA_{pre}	AA_{post}	PA	PA_{pre}	PA_{post}	NA	NA_{pre}	NA_{post}

(b) Evaluation results on the second stage.

Baselines	AA	AA_{pre}	AA_{post}	PA	PA_{pre}	PA_{post}	NA	NA_{pre}	NA_{post}

(c) Evaluation results on the third stage.

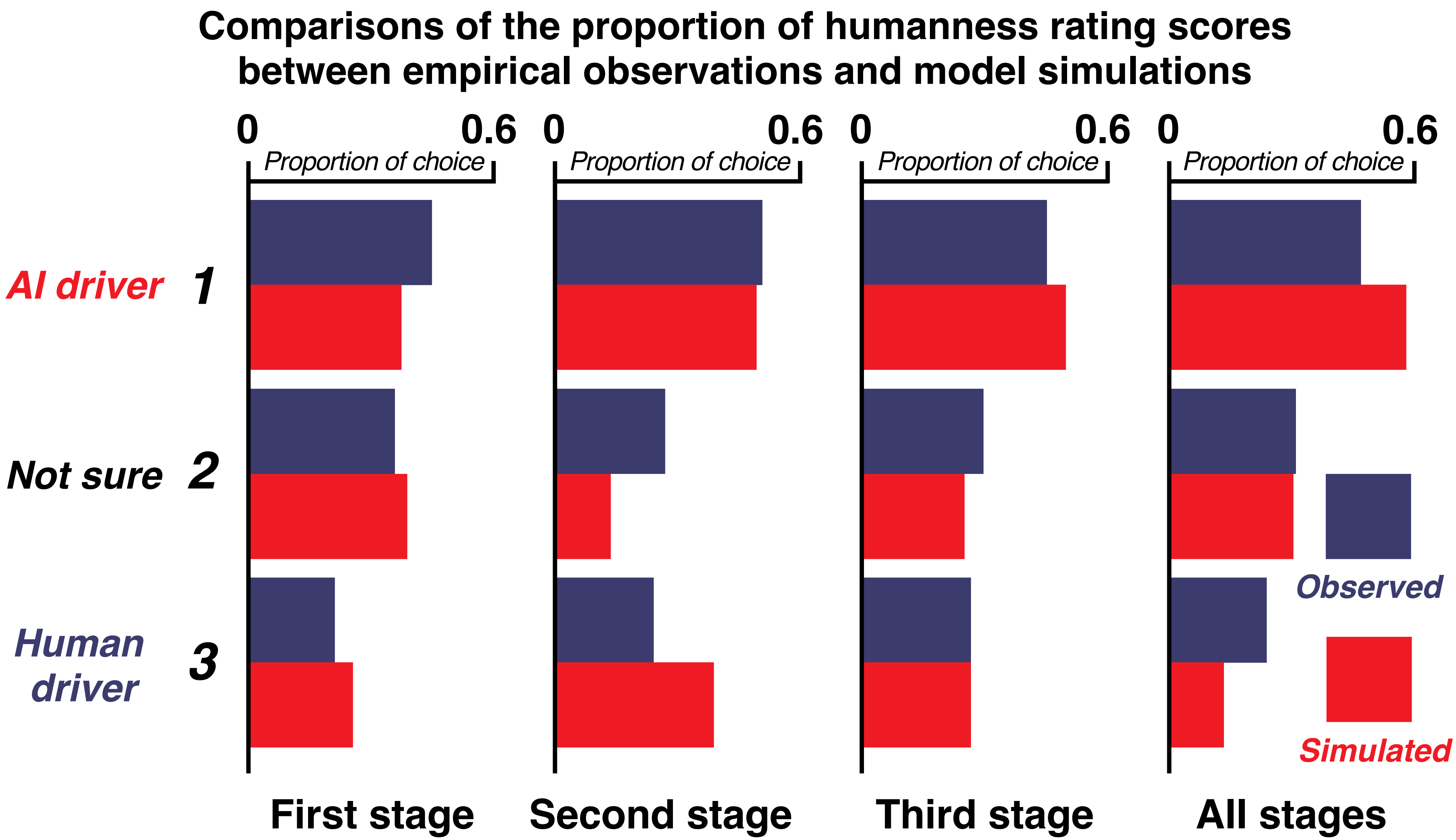
Baselines	AA	AA_{pre}	AA_{post}	PA	PA_{pre}	PA_{post}	NA	NA_{pre}	NA_{post}

(d) Evaluation results on all stages.

Baselines	AA	AA_{pre}	AA_{post}	PA	PA_{pre}	PA_{post}	NA	NA_{pre}	NA_{post}
MLR	0.0573	0.1516*	0.0749	0.0543	0.1264*	0.0988	0.0931	0.1160	0.0520
KNN	0.0992	0.1521*	0.1198*	0.0419	0.0144	0.1216*	0.1116	0.1422*	-0.0497
SVC	0.0854	0.0755	0.1457*	0.0414	0.0991	0.0688	0.0467	0.0676	0.0038
RF	0.0505	0.1308*	0.0292	0.1491*	0.0457	-0.0001	0.0117	0.0500	0.1426*
XGBoost	0.1411*	0.2586***	0.0198	0.1254*	0.1157	0.0044	0.2176**	0.1969**	0.1357*
MLP	0.0952	0.1949**	0.0701	0.1349*	0.0540	0.0830	0.2037**	0.2078**	0.0842
Baselines	*None*	**SDT-AT**	*AA+MF*	*AA*	*PA+MF*	*PA*	*NA+MF*	*NA*	***MF***
Random	0.0013	Original	0.1850**	0.1816**	0.0326	0.1416*	-0.1204	0.1685**	0.0570
Probability	-0.0006	**PLM-wv**	0.2704***	0.2452***	0.2447***	0.2331***	0.2866****	0.1871**	**0.5093******
Detective	0.1764**	PLM-tf	0.2837****	0.2879****	0.2734****	0.2878****	0.4178****	0.2004**	0.4641****

Based on Lewin’s equation, our proposed SDT-AT models provided superior within- (Table a-c) and cross-stage performance (Table d) than all other baselines, demonstrating the overall effectiveness of these models.

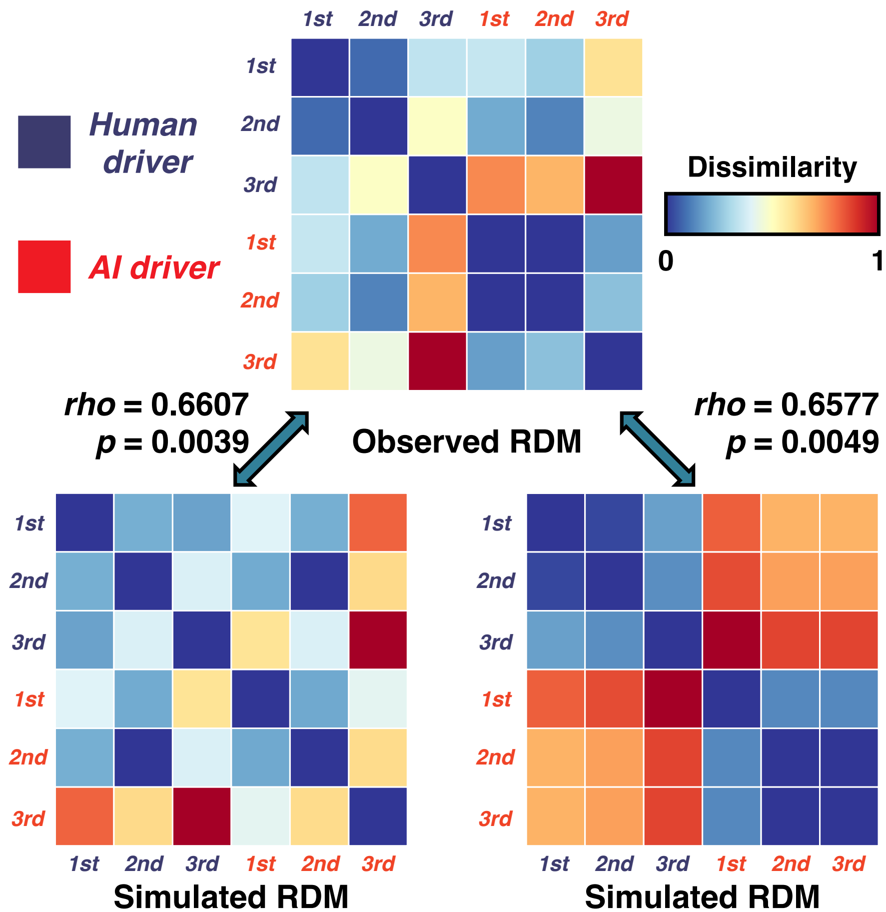
RESULTS OF THE COMPUTATIONAL MODELLING



Our computational model accurately captured the passenger's humanness rating behaviour patterns.

RESULTS OF THE COMPUTATIONAL MODELLING

Representational similarity between empirically observed humanness rating scores and model simulations averaged over all trials



(bottom left RDM, derived from a combination of within-stage model simulations)

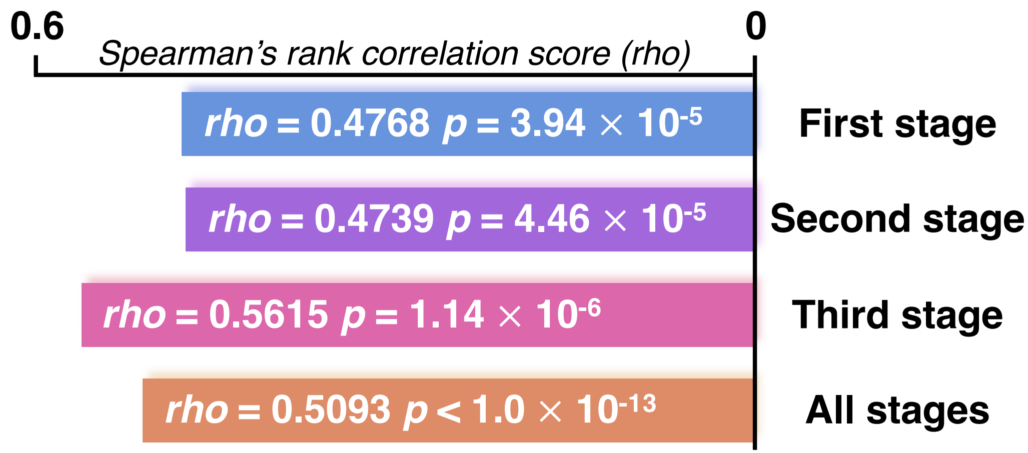
(bottom right RDM, derived from cross-stage model simulations)

Our model exhibited the same humanness rating behaviour pattern as passengers did.

ANALYSIS

Affective transition, serving as a hypothetical essential part (i.e., P) of passengers' subjective ride experience in our model, may play a crucial role in their ascription of humanness.

Spearman's rank correlation scores between the humanness rating and the magnitude of affective transition



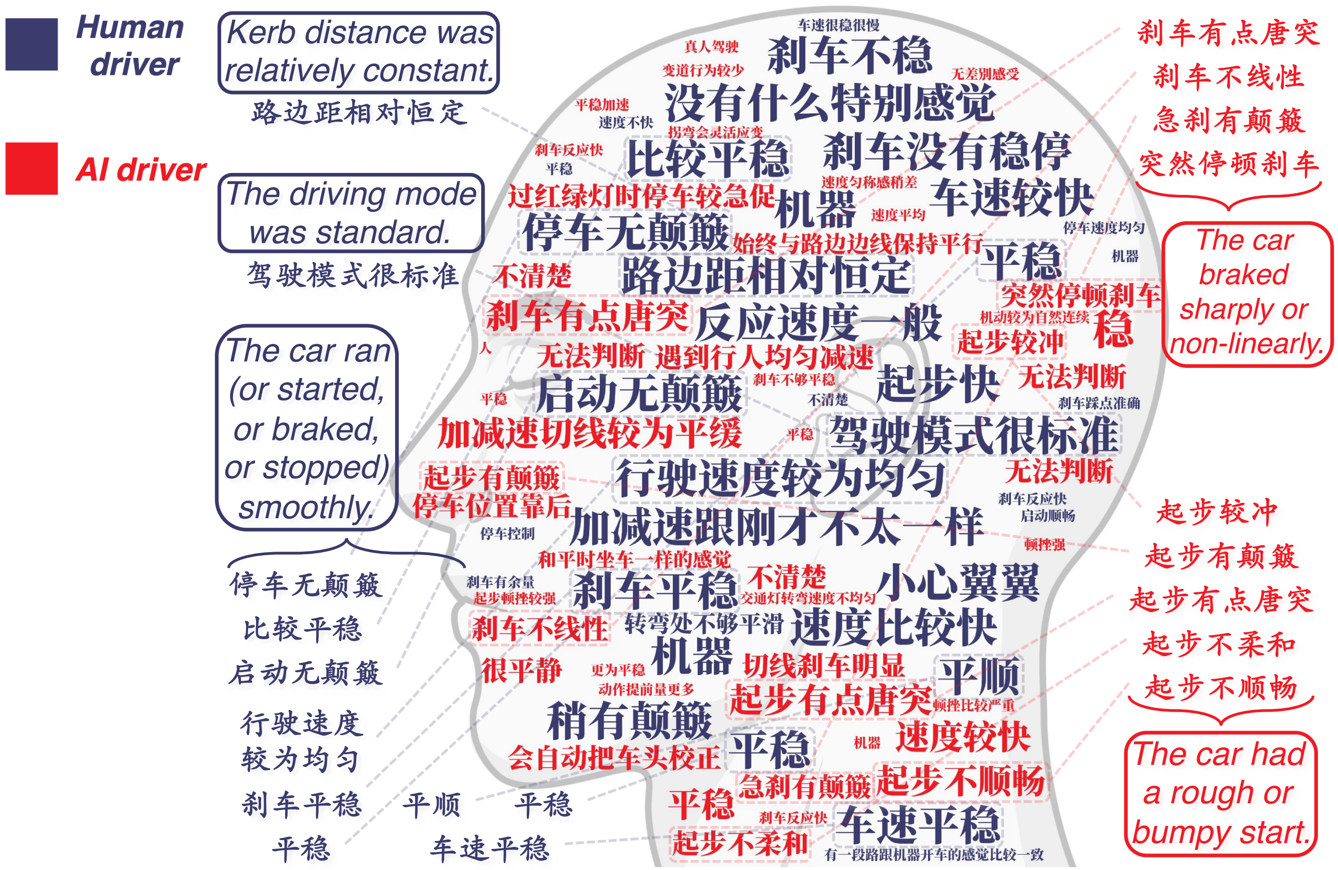
Mean changes in positive affect during the first and second stages

Conditions	ΔM	SD	z	p
First stage				
Human driver	**0.742**	**2.627**	**1.68**	**0.046**
AI driver	-0.622	2.803	-0.78	0.218
Second stage				
Human driver	**0.500**	**1.396**	**1.51**	**0.065**
AI driver	-0.375	2.983	-0.76	0.223

Enhancing positive affect may be the essence of the human-like ride experience during the starting two stages.

ANALYSIS

Word cloud displaying mixed feelings (MF) from all stages, i.e., the difference in the passenger's subjective ride experience between the two conditions



The size of each MF item is proportional (positively for the human driver condition, negatively for the AI driver condition) to the related z-scored transition from cross-stage model simulations.

The figure illustrates details of what needs to be improved for current automated driving to offer a human-like ride experience for the passenger.

DISCUSSION

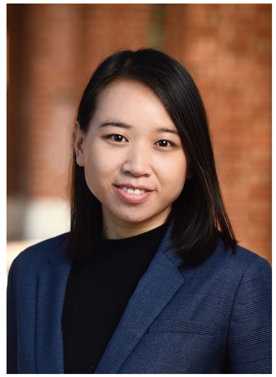
- The present study examined whether the current SAE Level 4 AC could create a human-like ride experience for passengers in a real-road scenario **for the first time**. The AI driver failed to pass our test because passengers detected the AI driver above chance.
- Our proposed computational model could adequately predict passengers' humanness rating behaviour. The practical success of basing the computational modelling on Lewin's seemingly abstract and theoretical field theory speaks directly to his famous maxim that '**there is nothing as practical as a good theory**' [3].
- We offer the first insights into what renders passengers' subjective ride experience truly human-like for future automated driving: **the passengers' ascription of humanness would increase with the greater affective transition**.
- Our results demonstrate the possibility and feasibility of using NLP techniques (e.g., pre-trained language models) as **adjuncts** to the interaction between social cognition and artificial intelligence to guide theorising and the generation of conceptual insights.
- Our further analysis of affective transition provided more concrete suggestions for the self-driving algorithm to offer a human-like ride experience for the passenger, e.g., improving passengers' positive affect during the starting stage and ensuring smoother starting and braking.
- We conjecture that the lack of a certain level of **mentalising ability** in the current self-driving algorithm may underlie its failure to pass our non-verbal variation of the Turing test. In this regard, our study calls for a spotlight on the importance of ensuring ACs (or **artificial social intelligence**, more broadly speaking) have at least some mentalising ability.

3. K. Lewin, "Psychology and the process of group living," *J. Soc. Psychol.*, vol. 17, no. 1, pp. 113–131, 1943.

ACKNOWLEDGEMENT & CONTACT

Qiaoli Jiang

Zhengming Wu



Anqi Liu

Haiyan Wu

Miner Huang

Kai Huang

Yixuan Ku

Presenter: Zhaoning Li 李肇宁

@lizhn7@sciences.social

github.com/das-boot

yc17319@umac.mo

@lizhn7