澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

認知與腦科學研究中心
Centro de Ciências Cognitivas e Cerebrais
Centre for Cognitive and Brain Sciences

社會科學學院
FACULDADE DE CIÊNCIAS SOCIAIS
FACULTY OF SOCIAL SCIENCES

Towards building artificial social intelligence (ASI) with mentalising ability: Two preliminary studies

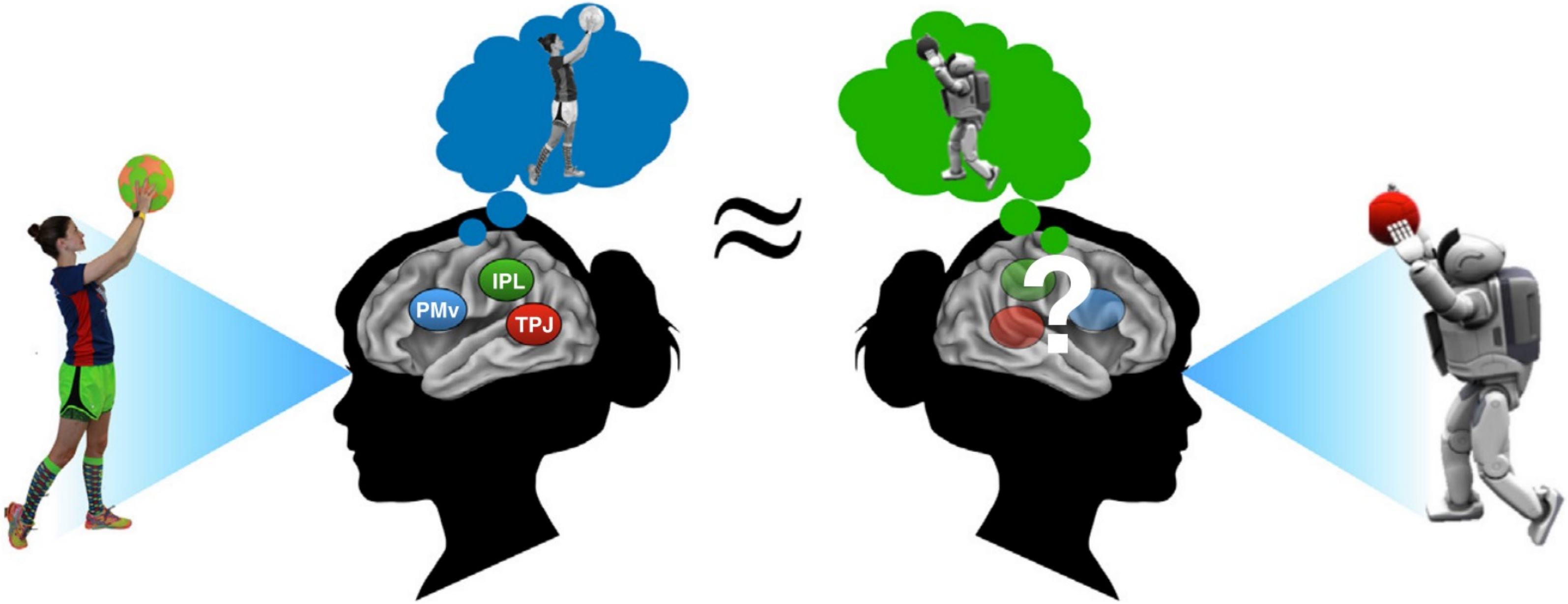
基于心智化能力的人工智能体构建初探

Presenter: Zhaoning Li 李肇宁

Invited Talk at NCC Lab & AND Lab Joint Workshop

Prologue

Machines with artificial social intelligence (ASI) are designed to either detect and respond to social signals in the environment or detect and respond to signals in the environment in a way that is perceived as social by human users, or some combination of these two possibilities [1].



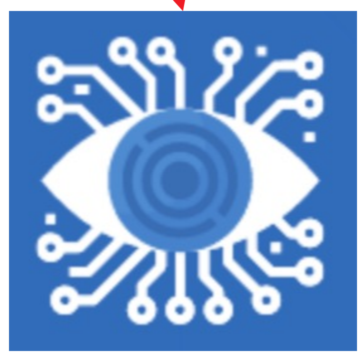
(Adapted from Cross & Ramsey, 2021)

1. Cross, E. S., & Ramsey, R. (2021). Mind meets machine: Towards a cognitive science of human–machine interactions. *Trends in Cognitive Sciences*, *25*, 200–212.

Prologue

Machines with artificial social intelligence (ASI) are designed to either detect and respond to social signals in the environment or detect and respond to signals in the environment in a way that is perceived as social by human users, or some combination of these two possibilities [1].

Artificial Narrow Intelligence (ANI)



Explainable Artificial Intelligence (XAI)

Artificial General Intelligence (AGI)

(Adapted from machine-desk.com and slidesalad.com)

1. Cross, E. S., & Ramsey, R. (2021). Mind meets machine: Towards a cognitive science of human–machine interactions. *Trends in Cognitive Sciences*, *25*, 200–212.

Prologue

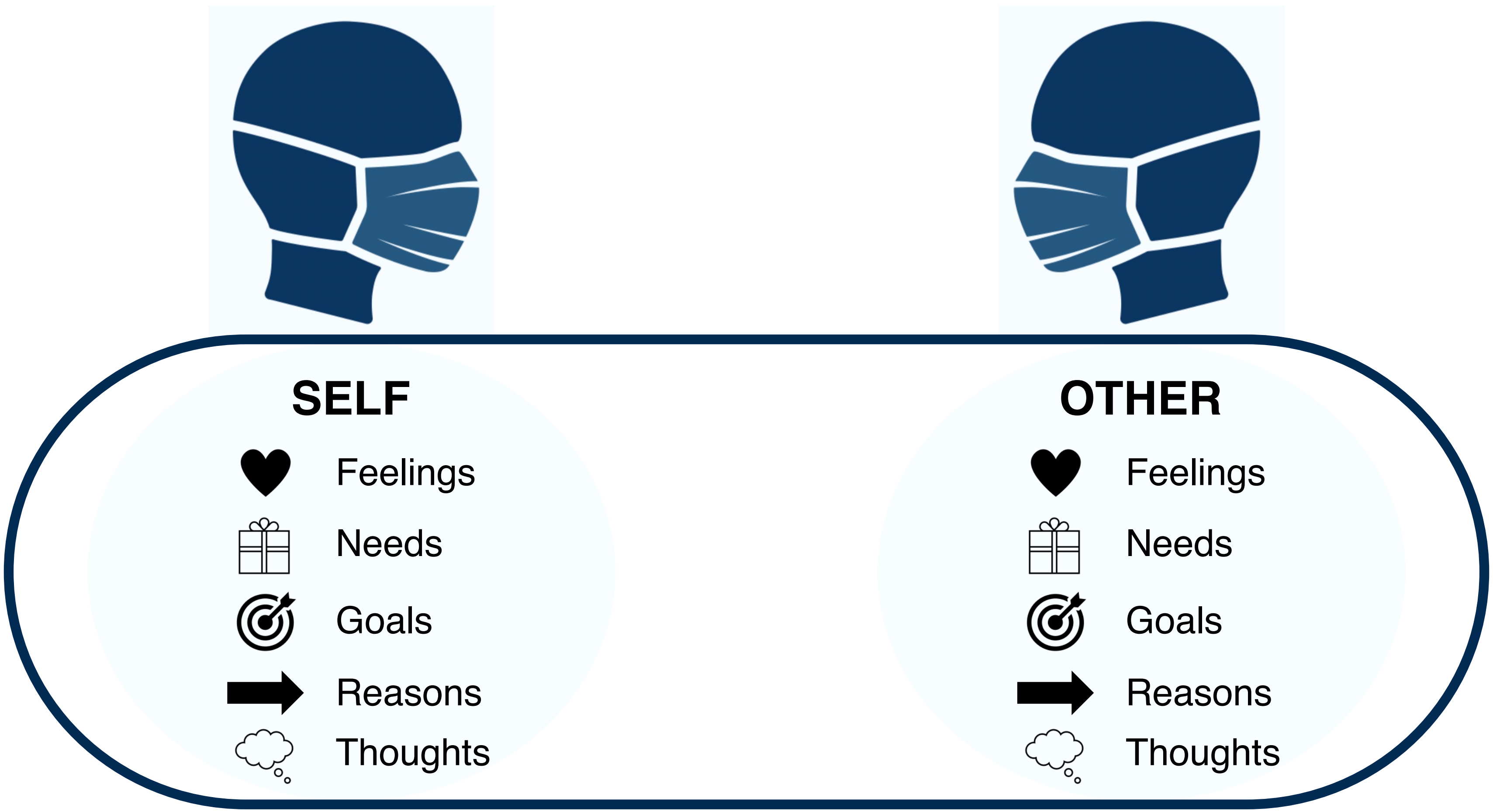
We've been through 2.5 million years of human evolution since our first hominid ancestors. Our brain size has tripled since the first hominids, to cope with communication, tool-use, and love [2].

(Adapted from Becker-Phelps, 2016)

2. Becker-Phelps, L. (2016). *Love: The psychology of attraction*. DK.

Prologue

Mentalising ability is a pivotal and fundamental component of human social intelligence.



(Adapted from BioRender.com)

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

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 is no large-scale deployment of autonomous cars yet.

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

Al-Shihabi & Mourant, 2001; Al-Shihabi & Mourant, 2003; Gu et al., 2017; Hecker et al., 2019; Sun et al., 2020.

However, literature presents no human-subject research focusing on passengers in a natural environment that examines whether the AC should behave in a human-like manner.

Research question

How to offer naturalistic experiences from a passenger's seat perspective to measure the people's acceptance of ACs?

The nonverbal Turing test of automated driving

How do human passengers choose?

Passenger

First-order mentalising
(self-other mentalising)

Human driver

or

AI driver

(Wu et al, 2019; Wu et al, 2020)

Choice behaviour →

$$B = f(P, E)$$

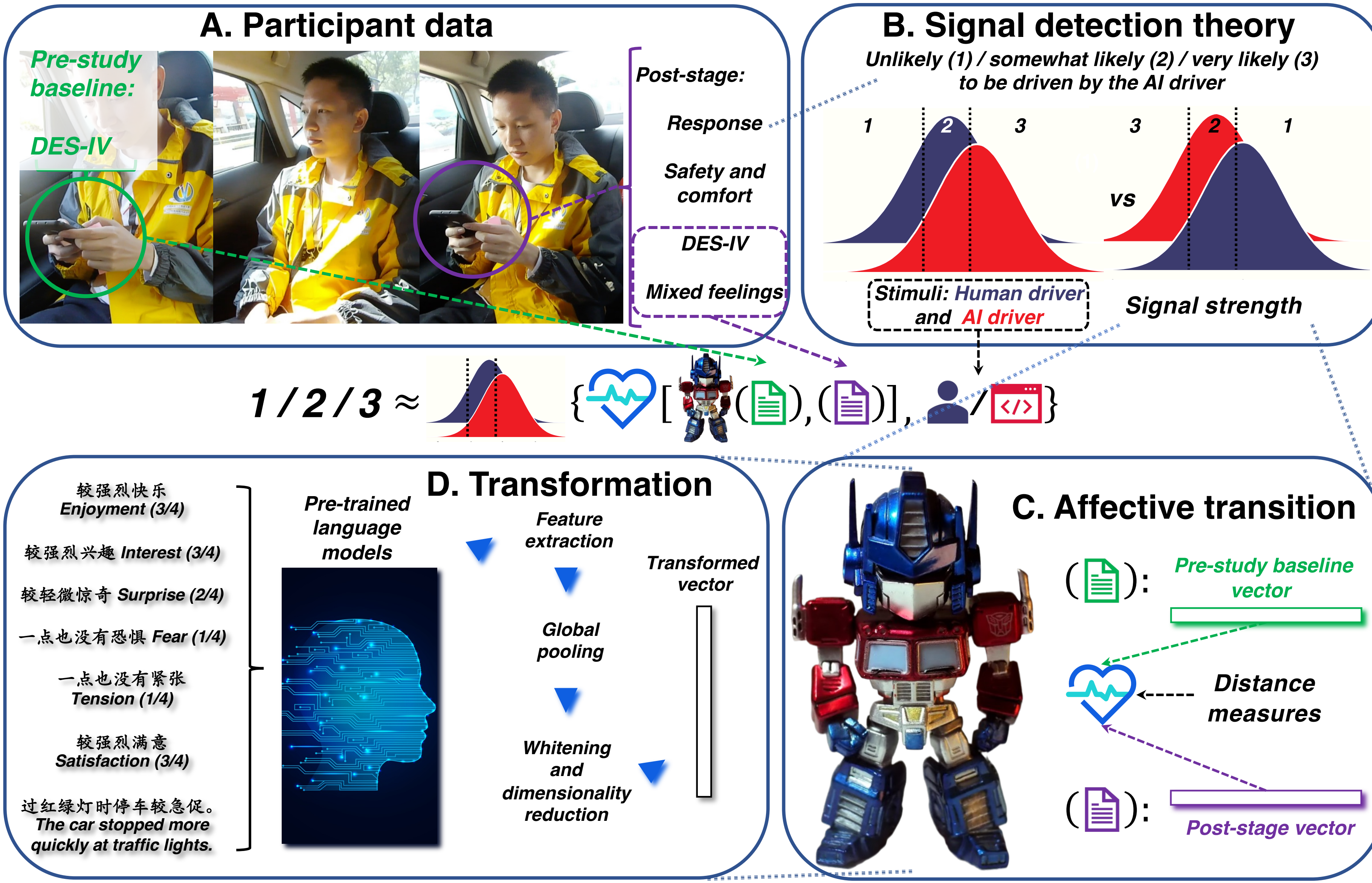
Passenger

Driving environment

Kurt Lewin, 1936

(Adapted from Wikipedia)

How do human passengers choose: SDT-AT (PLM)



Results of the computational models

Comparison 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}
MLR	-0.1844	0.1312	0.1283	0.0988	0.1761	-0.0082	-0.0453	0.0390	0.0744
KNN	0.1998	0.0616	-0.0069	0.2043*	0.3045**	-0.0509	0.0804	0.0596	0.0591
SVC	-0.0902	0.0781	-0.0222	0.0832	0.1928	-0.0016	0.0326	-0.0314	0.0065
RF	0.1323	0.0971	0.0181	0.0925	0.2354*	0.0591	-0.0252	0.0773	0.1126
XGBoost	0.1322	0.3034**	-0.1130	0.2262*	0.2614*	-0.0122	0.0621	0.1896	0.1181
MLP	0.3153**	0.3654**	0.2479*	0.1256	0.0516	0.0679	0.0097	0.1567	0.0873

Baselines	*None*	**SDT-AT**	*AA+MF*	*AA*	***PA+MF***	*PA*	*NA+MF*	*NA*	*MF*
Random	0.0015	Original	-0.3985	-0.3552	-0.2580	0.1738	-0.3397	0.0828	0.0990
Probability	-0.0010	PLM (wv)	0.4511***	0.4152***	0.4092***	0.3939***	0.4064***	0.1359	0.3030**
Golden	0.1491	**PLM (tf)**	0.4113***	0.4639****	**0.4768********	0.3939***	0.3484**	0.1842	0.3738**

Results of the computational models

Comparison 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}
MLR	-0.1844	0.1312	0.1283	0.0988	0.1761	-0.0082	-0.0453	0.0390	0.0744

(b) Evaluation results on the second stage.

Baselines	*AA*	AA_{pre}	AA_{post}	*PA*	$\mathbf{PA_{pre}}$	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	0.2013*	0.2467*	-0.0567	0.0371	0.3523**	-0.2845	-0.1138	-0.1385	-0.0053
SVC	0.2258*	0.1915	0.1163	0.1284	0.0915	-0.1747	-0.1508	0.0836	-0.2366
RF	0.1541	0.3911***	-0.0122	0.0700	0.2136*	-0.0916	0.0672	0.1767	-0.3972
XGBoost	0.0934	0.2847**	-0.2574	0.0397	0.3560**	-0.0450	-0.1472	-0.2216	-0.1332
MLP	-0.0038	0.1463	-0.2474	0.0853	**0.4813******	-0.0308	-0.2472	-0.2060	-0.2274

Baselines	*None*	SDT-AT	*AA+MF*	*AA*	*PA+MF*	*PA*	*NA+MF*	*NA*	*MF*
Random	0.0097	Original	0.1750	0.2409*	0.1539	0.1912	0.1865	-0.0105	0.1824
Probability	-0.0020	PLM (wv)	0.4569****	0.4195***	0.4402***	0.4635****	0.3167**	0.1703	0.4276***
Golden	0.0394	PLM (tf)	0.4375***	0.4173***	0.4545****	0.4739****	0.3528**	0.2636*	0.3578**

Results of the computational models

Comparison 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}
MLR	-0.1844	0.1312	0.1283	0.0988	0.1761	-0.0082	-0.0453	0.0390	0.0744

(b) Evaluation results on the second stage.

Baselines	AA	AA_{pre}	AA_{post}	PA	$\mathbf{PA_{pre}}$	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

(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	0.1763	0.2289*	0.1951	0.1779	0.0384	0.2147*	0.4034***	0.3311**	0.3369**
SVC	0.4706****	0.3086**	0.2050	0.2393*	0.0671	0.1114	0.2278*	0.1002	0.2197*
RF	0.0553	0.3739**	0.2307*	-0.1087	0.1919	0.0203	0.3481**	0.3729**	0.2369*
XGBoost	0.0896	0.4084***	0.2747*	-0.1074	0.1474	0.0813	0.3895***	0.4127***	0.3041**
MLP	0.2142*	0.1700	0.2706*	0.1835	0.0368	0.1321	0.3501**	0.2982**	0.3658**

Baselines	*None*	**SDT-AT**	*AA+MF*	*AA*	*PA+MF*	*PA*	*NA+MF*	*NA*	***MF***
Random	-0.0013	Original	0.1490	0.2019	0.1978	-0.0258	0.4037***	0.4245***	0.1104
Probability	-0.0022	**PLM (wv)**	0.4861****	0.4556***	0.4624***	0.4322***	0.4419***	0.4256***	**0.5615******
Golden	0.3168**	PLM (tf)	0.4807****	0.4974****	0.4654****	0.4570***	0.4769****	0.4429***	0.5422****

Results of the computational models

Comparison 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}
MLR	-0.1844	0.1312	0.1283	0.0988	0.1761	-0.0082	-0.0453	0.0390	0.0744

(b) Evaluation results on the second stage.

Baselines	AA	AA_{pre}	AA_{post}	PA	$\mathbf{PA_{pre}}$	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

(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***

(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	0.0461	0.1263*	0.1196*	0.0138	0.0839	0.1654**	0.0558	0.1921**	0.0715
SVC	0.1658**	0.2296***	-0.0531	0.1381*	0.0998	0.0157	0.1441*	0.2198***	0.0391
RF	0.1129	0.1382*	0.0604	0.0845	0.0411	0.0721	0.0161	0.0470	0.1568*
XGBoost	0.1216*	0.1977**	0.0560	0.1624*	0.1008	0.0301	0.1639*	0.1603*	0,1588*
MLP	0.1050	0.0391	0.1262*	-0.0222	0.0914	0.0119	0.1475*	0.2035**	0.0764

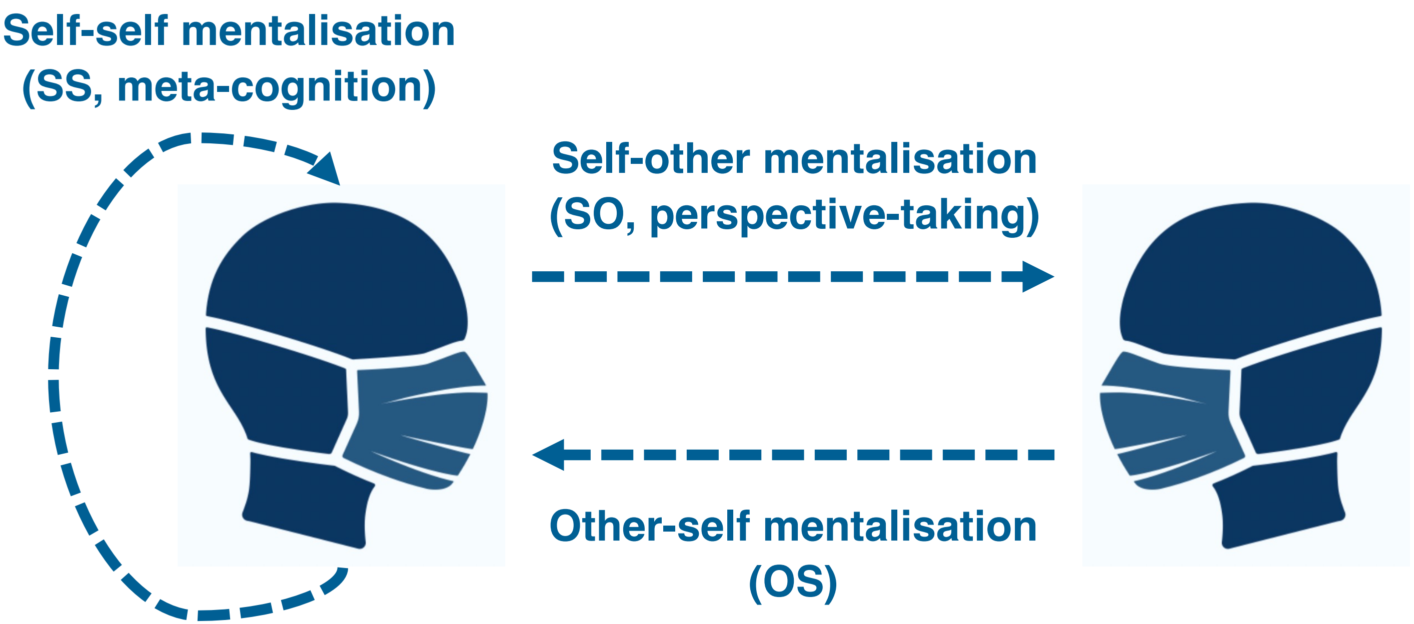
Baselines	*None*	**SDT-AT**	*AA+MF*	*AA*	*PA+MF*	*PA*	*NA+MF*	*NA*	***MF***
Random	-0.0001	Original	0.1850**	0.1816**	0.0326	0.1416*	-0.1204	0.1685**	0.0570
Probability	-0.0027	**PLM (wv)**	0.2704***	0.2452***	0.2447***	0.2331***	0.2866****	0.1871**	**0.5093******
Golden	0.1764**	PLM (tf)	0.2837****	0.2879****	0.2734****	0.2878****	0.4178****	0.2004**	0.4641****

Every individual makes a difference:

A trinity derived from linking individual brain morphometry, connectivity and mentalising ability

Background

Considering the multifaceted nature of mentalising ability [3], little research has focused on characterising individual differences in different mentalising components [4].

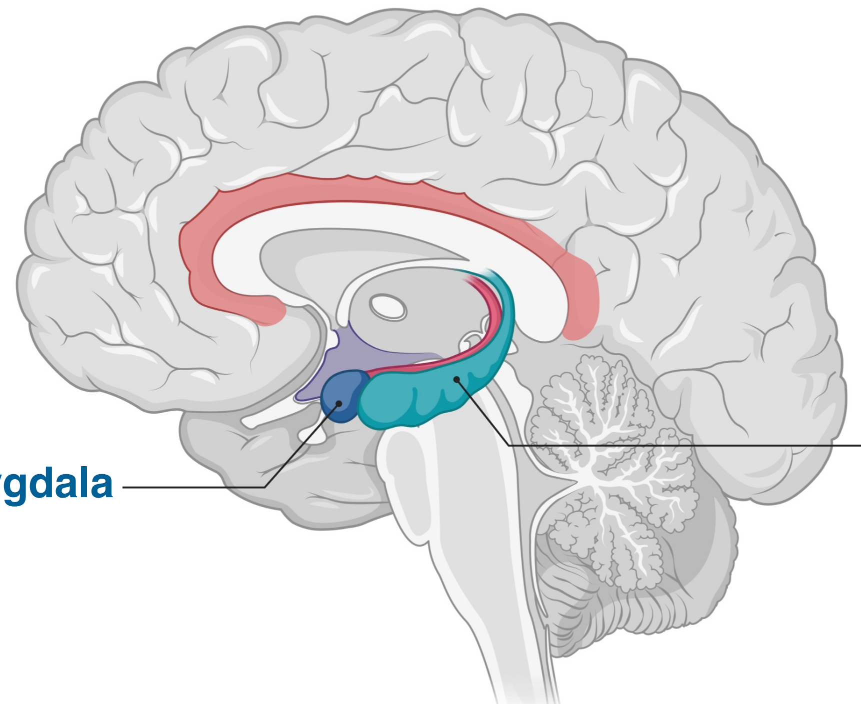


(Adapted from BioRender.com)

3. Wu, H., Liu, X., Hagan, C. C., & Mobbs, D. (2020b). Mentalising during social interaction: A four component model. *Cortex*, *126*, 242–252.
4. Wu, H., Fung, B. J., & Mobbs, D. (2022). Mentalising during social interaction: The development and validation of the interactive mentalising questionnaire. *Frontiers in Psychology*, *12*.

Background

And even less research has been devoted to investigating how the variance in the structural and functional patterns of the amygdala and hippocampus, two vital subcortical regions of the 'social brain' [5, 6], are related to inter-individual variability in mentalising ability.



(Adapted from BioRender.com)

5. Bickart, K. C., Dickerson, B. C., & Barrett, L. F. (2014). The amygdala as a hub in brain networks that support social life. *Neuropsychologia*, *63*, 235–248.
6. Montagrin, A., Saiote, C., & Schiller, D. (2018). The social hippocampus. *Hippocampus*, *28*, 672–679.

Research question

Whether inter-individual variability in the structural or functional patterns of the above two brain regions is associated with that in different mentalising components?

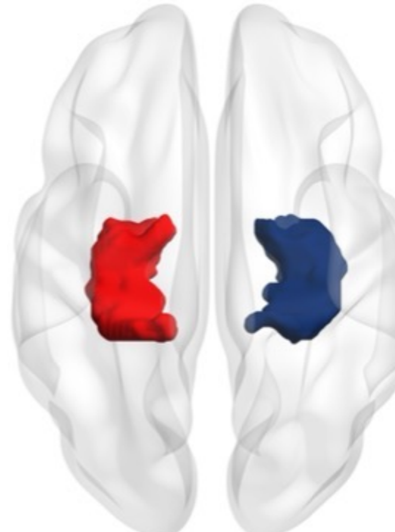
MMS: Surface-based multivariate morphometry statistics

Processing pipeline of hippocampal morphometry data

(a)

T1-weighted MRI scans

(b)

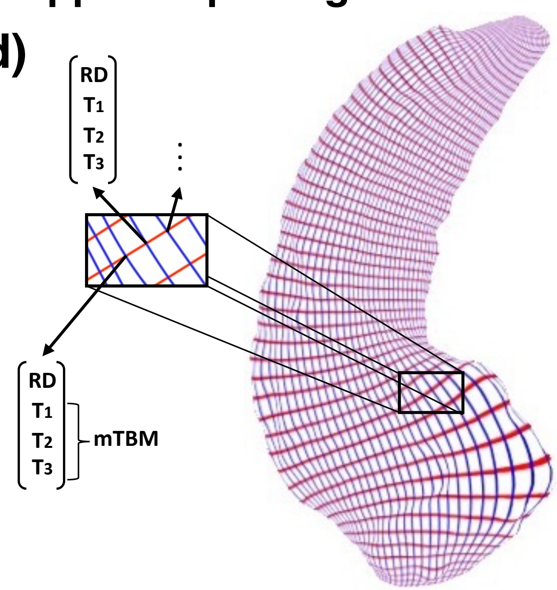


Hippocampal segmentation

(c)

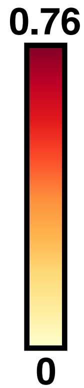
Smoothed surface

(d)

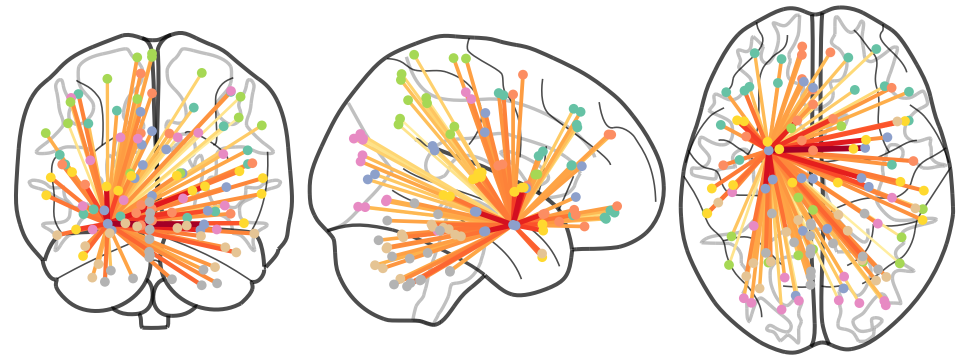


Multivariate morphometry statistics

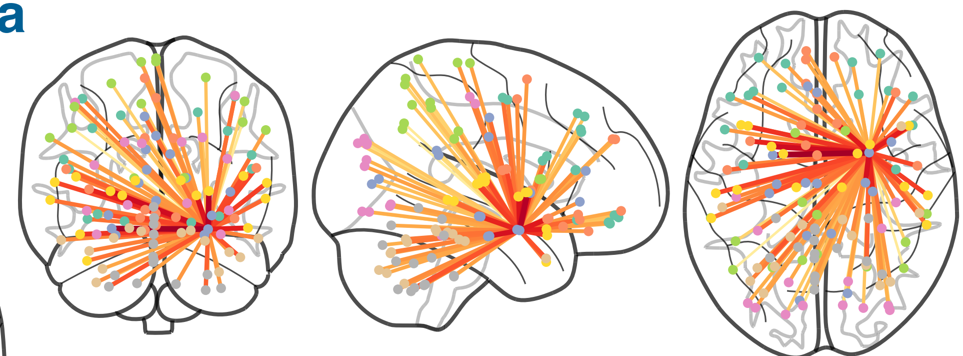
Rs-FC: Resting-state functional connectivity

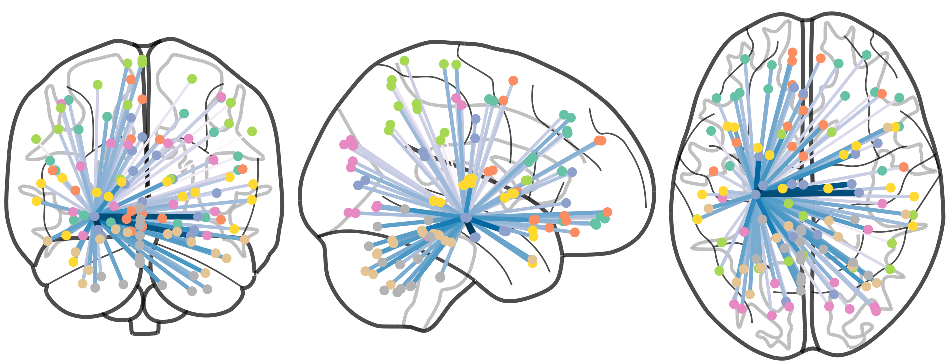


Left amygdala

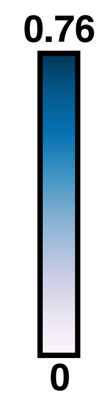


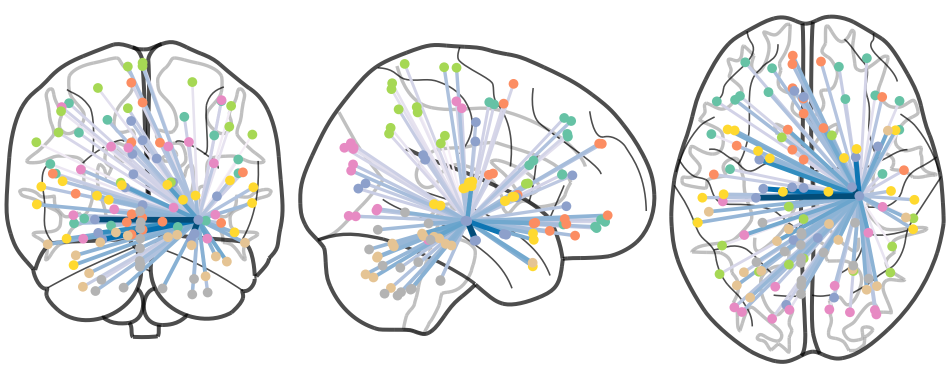
Right amygdala





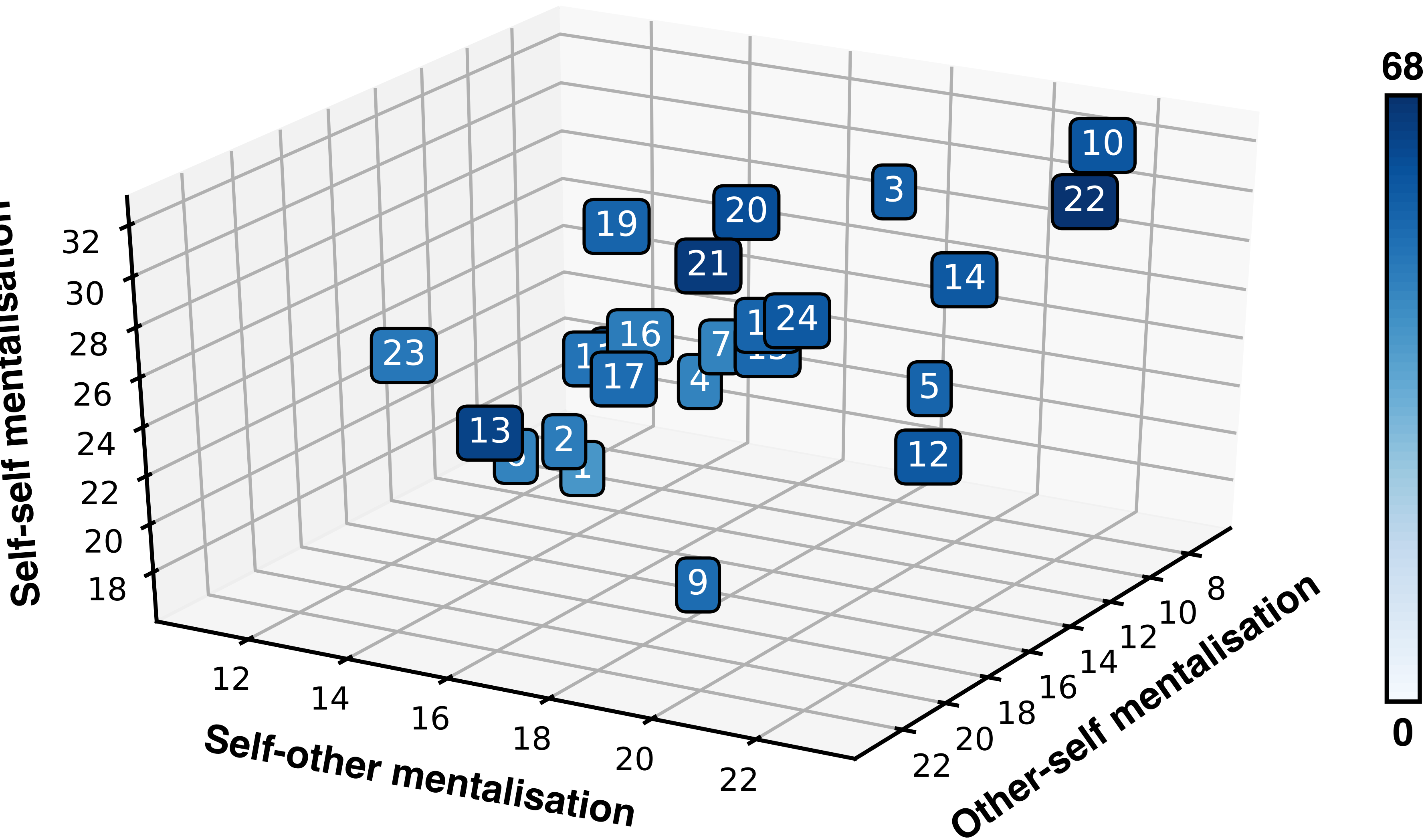
Left hippocampus





Right hippocampus

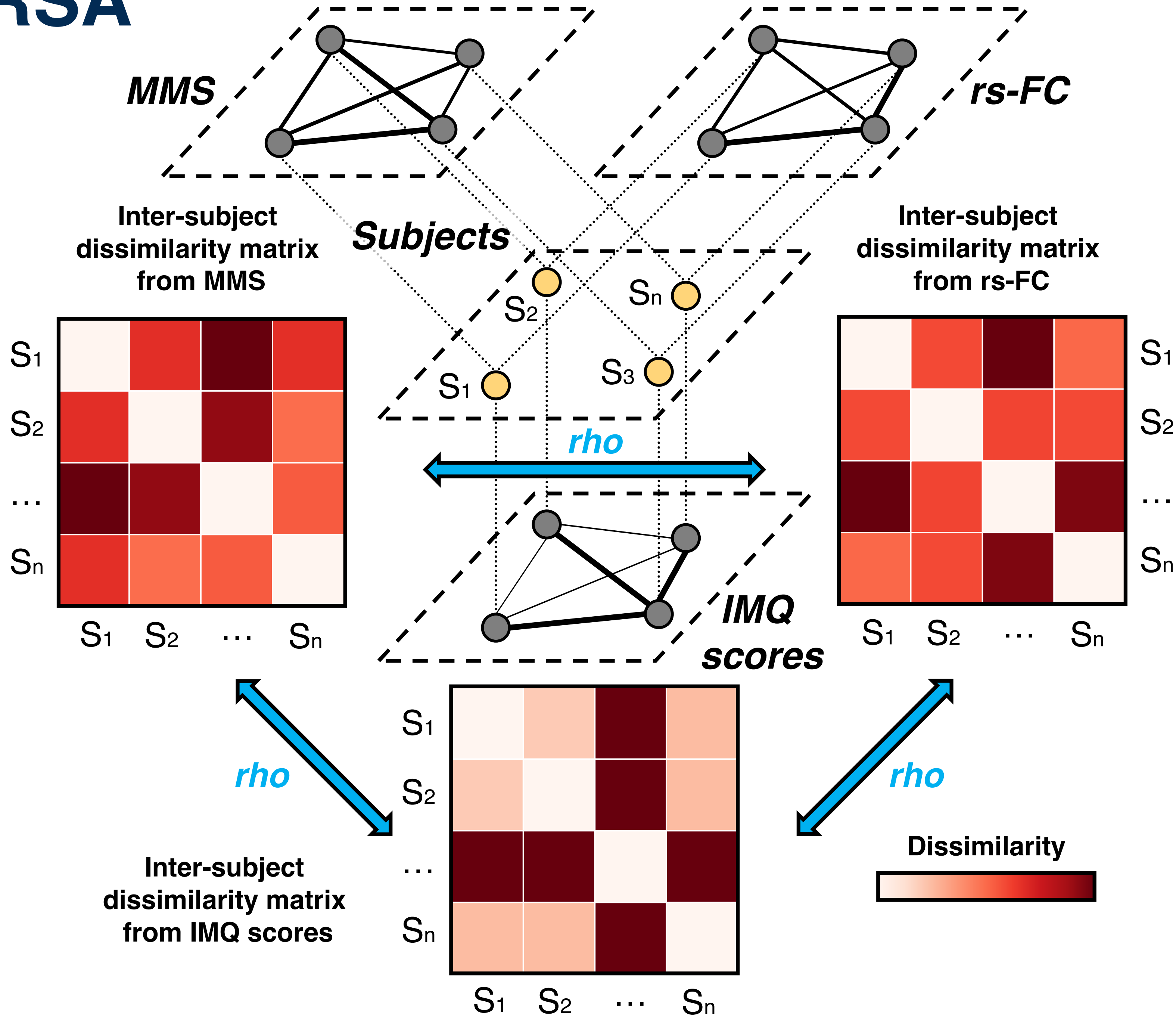
IMQ: Interactive mentalisation questionnaire [3, 4]



3. Wu, H., Liu, X., Hagan, C. C., & Mobbs, D. (2020b). Mentalising during social interaction: A four component model. *Cortex*, *126*, 242–252.
4. Wu, H., Fung, B. J., & Mobbs, D. (2022). Mentalising during social interaction: The development and validation of the interactive mentalising questionnaire. *Frontiers in Psychology*, *12*.

IS-RSA: Inter-subject representational similarity analysis

IS-RSA


MMS
rs-FC
Inter-subject dissimilarity matrix from MMS
Subjects
Inter-subject dissimilarity matrix from rs-FC
S2
Sn
S1
S3
rho
IMQ scores
S1
S2
…
Sn
rho
rho
Inter-subject dissimilarity matrix from IMQ scores
Dissimilarity

Hypothesis 1

We predicted that

1) the levels of mentalising ability would correlate positively with the dissimilarity in amygdala and hippocampal morphometry and connectivity;
2) dissimilarity in functional and structural patterns would positively covary with each other.

Hypothesis 1

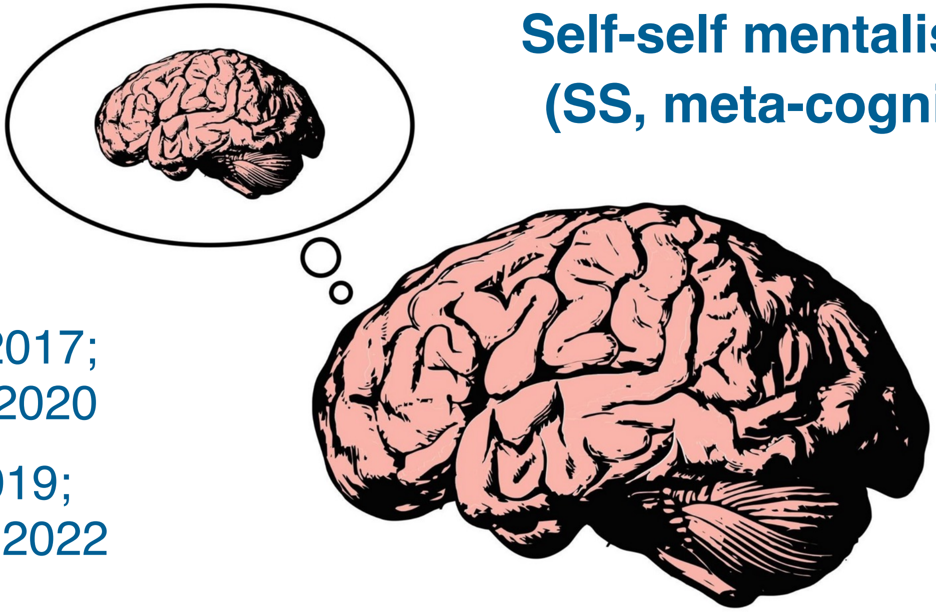
Three distinct modalities will share one essence, i.e., there is a structure that existed in idiosyncratic patterns of brain morphometry, connectivity and mentalising ability, and we termed it as ‘trinity’.

(Adapted from Wikipedia)

Hypothesis 2

There will be a region-related specificity in associations among different mentalising components and amygdala or hippocampal MMS and rs-FC.

Self-self mentalisation (SS, meta-cognition)



Allen et al., 2017;
Alkan et al., 2020

Ye et al., 2019;
Zou & Kwok, 2022

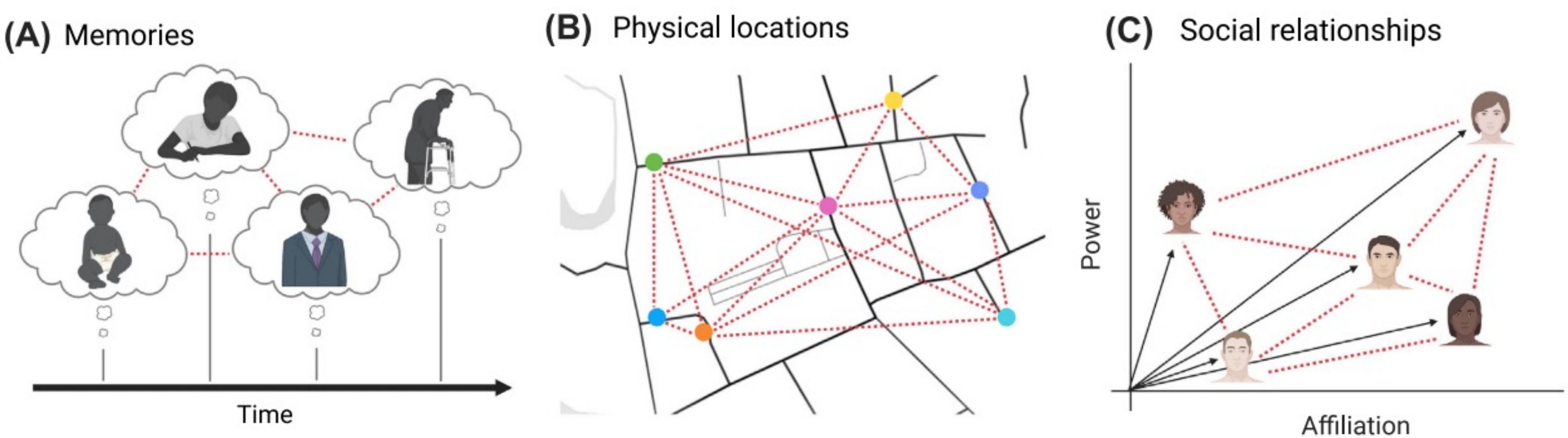
(Adapted from PriMed)

Hypothesis 2

There will be a region-related specificity in associations among different mentalising components and amygdala or hippocampal MMS and rs-FC.

Self-other mentalisation (SO, perspective-taking)

Relational integration theory
(O'Keefe & Nadel, 1978; Rubin et al., 2014)



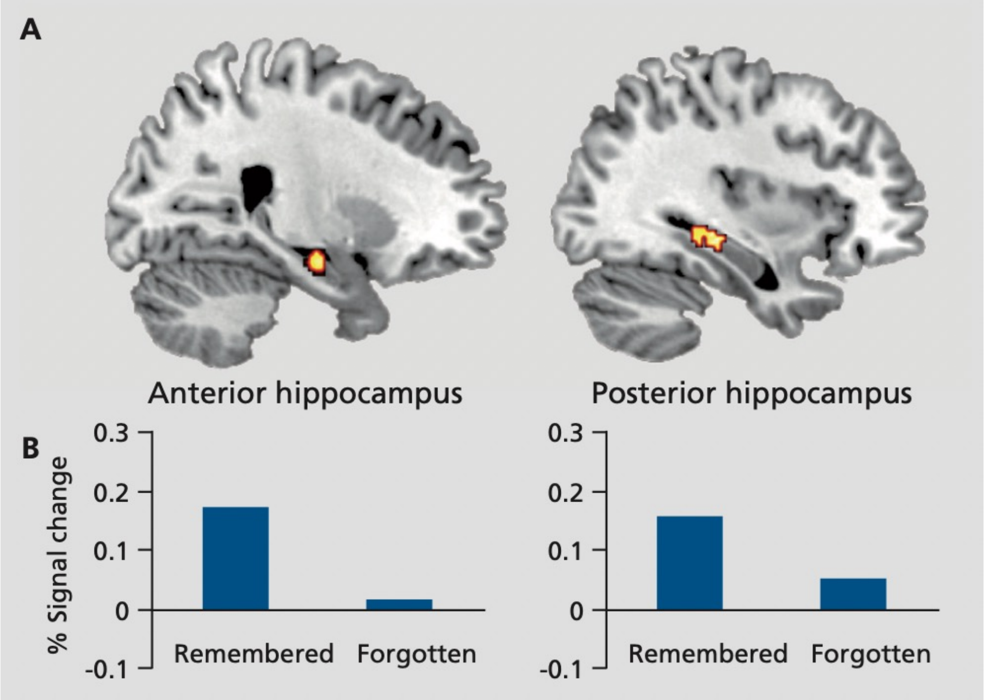
(Adapted from Banker et al., 2021)

Hypothesis 2

There will be a region-related specificity in associations among different mentalising components and amygdala or hippocampal MMS and rs-FC.

Self-other mentalisation (SO, perspective-taking)

Constructive memory theory
(Schacter, 2012)



Hippocampal responses to encoding simulations of future events

(Adapted from Schacter, 2012)

Hypothesis 2

There will be a region-related specificity in associations among different mentalising components and amygdala or hippocampal MMS and rs-FC.

Other-self mentalisation (OS, the ability to see 'ourselves from the outside')

Wu et al., 2022

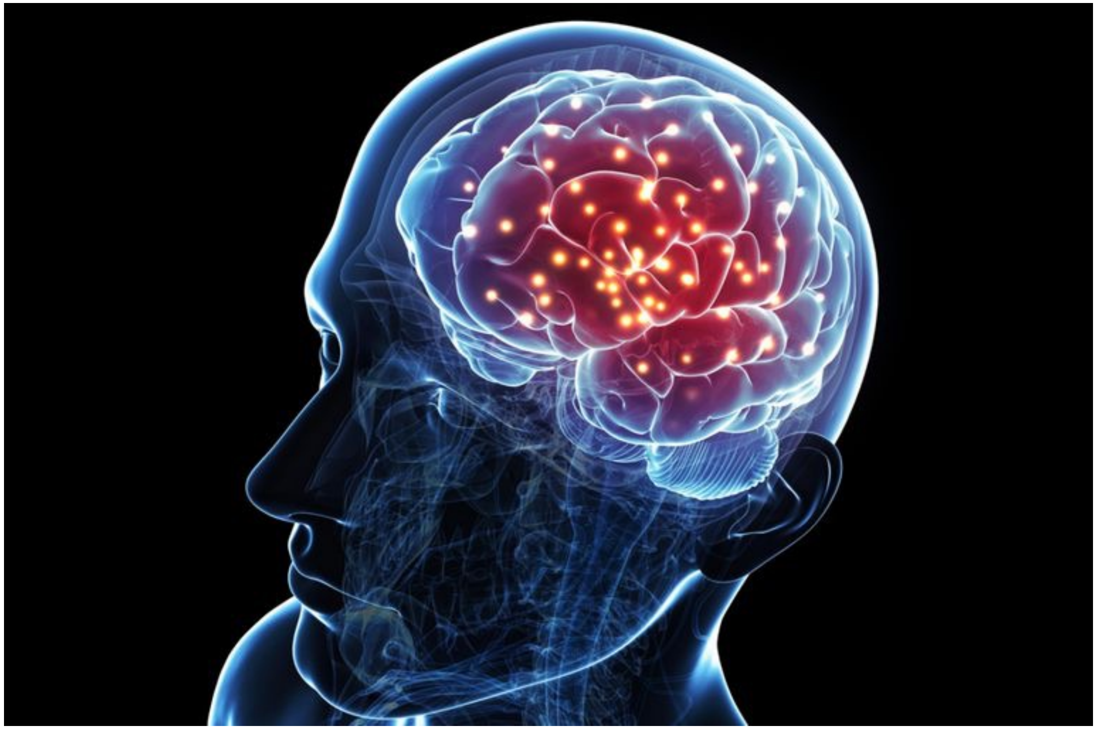
Koscik & Tranel, 2011;
Haas et al., 2015;
Santos et al., 2016;
Eskander et al., 2020

(Adapted from Earth.com)

Hypothesis 3

Subject pairs with similar hippocampal MMS will have even greater SS and SO similarity if they are also similar in hippocampal rs-FC.

In a similar vein, subject pairs with similar amygdala MMS will have even greater OS similarity if they are also similar in amygdala rs-FC.

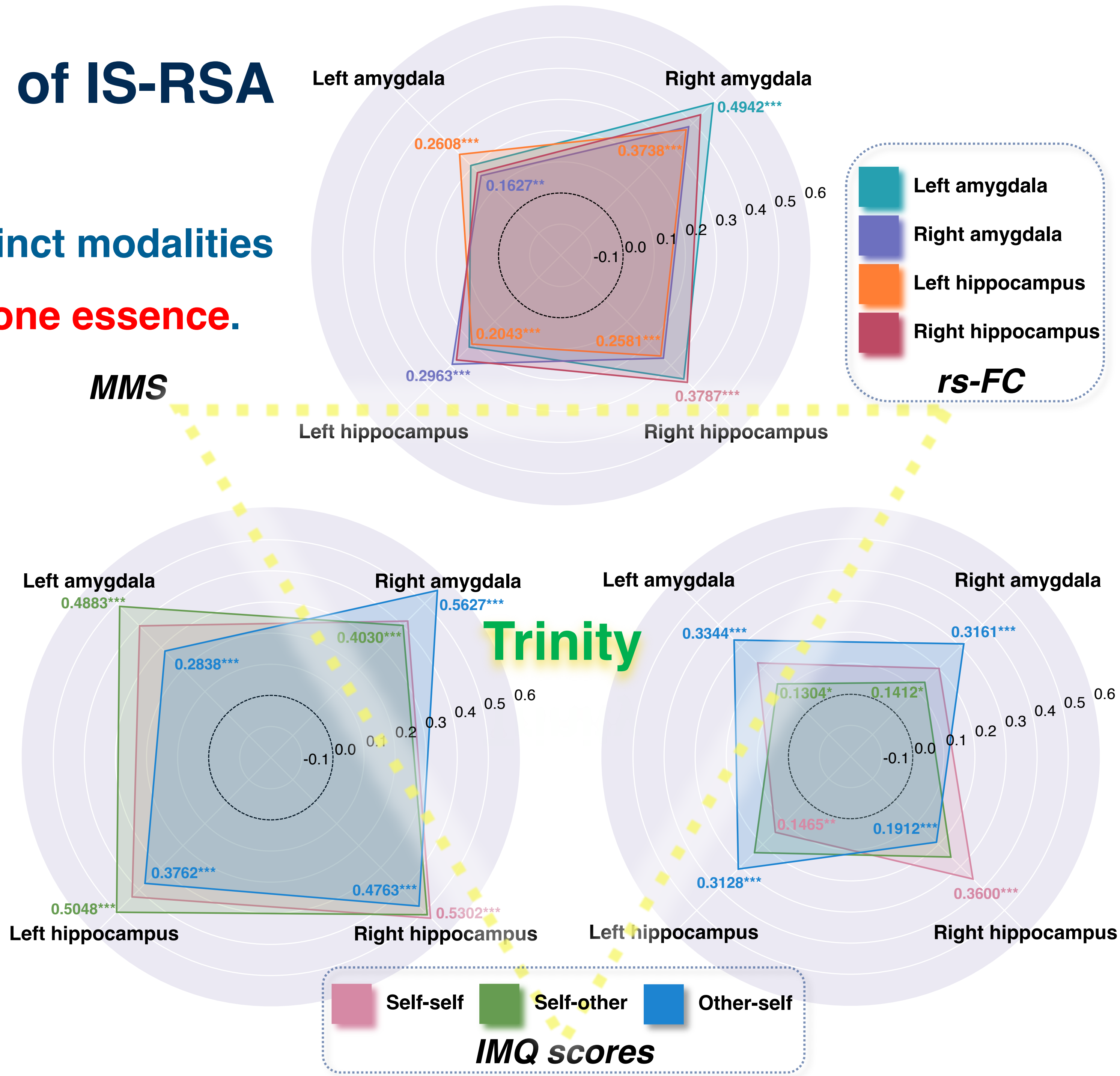


(Adapted from thoughtco.com)

Results of IS-RSA

Three distinct modalities

shared one essence.



Results of IS-RSA

A region-related mentalising specificity emerged from the trinity.

Comb.	*rho*	Mean (*95% CI*)	p_{FDR}
SS			
LA	0.3981	0.3677 (0.3569-0.3785)	<.001***
RA	0.4228	0.3947 (0.3861-0.4034)	<.001***
LH	0.4347	0.4127 (0.4055-0.4199)	<.001***
RH	0.5302	0.5168 (0.5051-0.5284)	<.001***
SO			
LA	0.4883	0.4607 (0.4478-0.4736)	<.001***
RA	0.4030	0.3821 (0.3751-0.3891)	<.001***
LH	0.5048	0.4678 (0.4601-0.4755)	<.001***
RH	0.5156	0.4766 (0.4657-0.4875)	<.001***
OS			
LA	0.2838	0.2890 (0.2801-0.2980)	<.001***
RA	0.5627	0.5153 (0.5051-0.5255)	<.001***
LH	0.3762	0.3548 (0.3453-0.3643)	<.001***
RH	0.4763	0.4433 (0.4321-0.4544)	<.001***

(a) Results of similarities between IMQ scores and MMS.

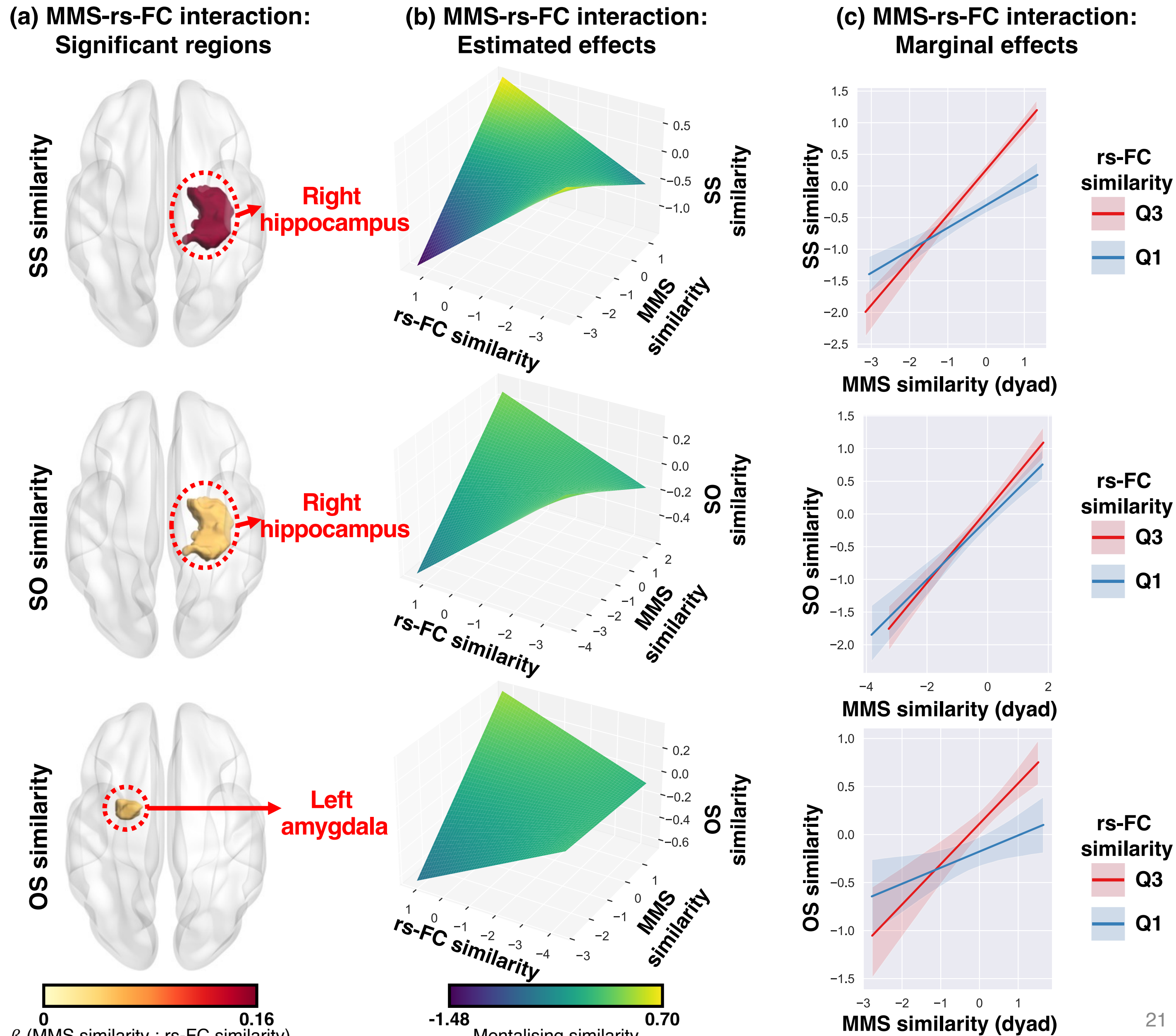
Comb.	*rho*	Mean (*95% CI*)	p_{FDR}
SS			
LA	0.2272	0.2094 (0.1995-0.2194)	<.001***
RA	0.2025	0.1747 (0.1668-0.1826)	<.001***
LH	0.1465	0.1256 (0.1162-0.1350)	.007**
RH	0.3600	0.3434 (0.3348-0.3520)	<.001***
SO			
LA	0.1304	0.1239 (0.1169-0.1310)	.016*
RA	0.1412	0.1359 (0.1266-0.1452)	.010*
LH	0.2383	0.2254 (0.2147-0.2360)	<.001***
RH	0.2580	0.2427 (0.2347-0.2508)	<.001***
OS			
LA	0.3344	0.3164 (0.3078-0.3250)	<.001***
RA	0.3161	0.2890 (0.2788-0.2993)	<.001***
LH	0.3128	0.2861 (0.2742-0.2980)	<.001***
RH	0.1912	0.1682 (0.1538-0.1825)	<.001***

(b) Results of similarities between IMQ scores and rs-FC.

'LA' for left amygdala; 'RA' for right amygdala; 'LH' for left hippocampus; 'RH' for right hippocampus

Results of dyadic regression analysis

Rs-FC gates the MMS predicted similarity in mentalising ability.



Summary

1. The current work defines an integrative trinity framework that provides a testable basis for understanding individual differences in brain morphometry, connectivity and mentalising ability.

2. Our study reveals the existence of a region-related specificity: the variation of SS and SO are more related to individual differences in hippocampal MMS and rs-FC, whereas the variation of OS shows a closer link with individual differences in amygdala MMS and rs-FC.

3. Our data suggest that rs-FC gates the MMS predicted similarity in mentalising ability, revealing the intertwining role brain morphometry and connectivity play in social cognition.

Acknowledgement & contact

10.1101/2022.04.11.4

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

好奇帮

@ANDlab3

***A**ffective **N**euroscience and*
D**ecision-making **Lab
andlab-um.com

Preprint:
https://doi.org/10.1101/2022.04.11.487870

The data and code used are available at
https://github.com/andlab-um/trinity

yc17319@umac.mo

@lizhn7

@Das-Boot

@Zhaoning Li