THE UNIVERSITY of EDINBURGH
Institute for Neuroscience and Cardiovascular Research

Collaborative Approach to Meta Analysis and Review of Animal Data from Experimental Studies
·C·A·M·A·R·A·D·E·S·

AI-enabled Systematic Review Platforms

How do they perform?

Dr Charis Wong
Honorary Senior Clinical Lecturer, University of Edinburgh
Consultant Neurologist, NHS Fife

charis.wong@ed.ac.uk

UHMLG Spring Forum | 23 April 2026

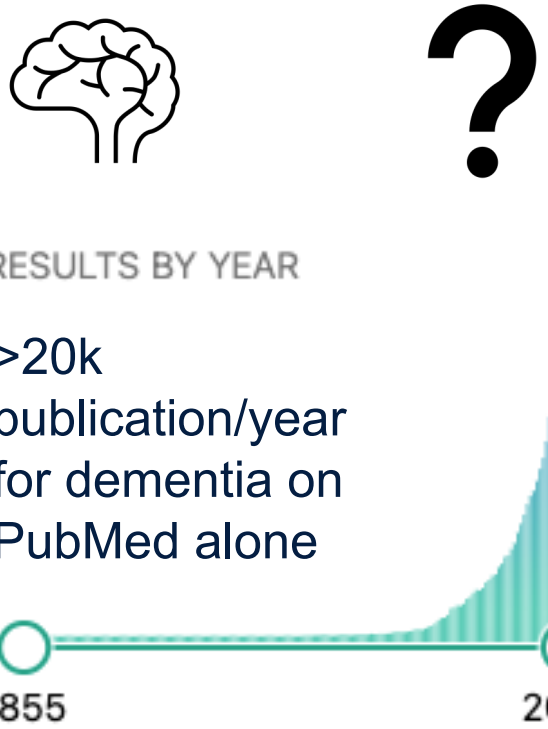
edin.ac/incr

Disclosure

Nested Knowledge has provided us with a free pilot of their ‘Enterprise’ tier for the purpose of this evaluation.

THE UNIVERSITY of EDINBURGH
Institute for Neuroscience and Cardiovascular Research

About me (and how I got interested in this…)



THE UNIVERSITY of EDINBURGH
Institute for Neuroscience and Cardiovascular Research


We need to develop better evidence synthesis tools!
Why don’t we use [insert AI-SRP name]?

Aim

- To evaluate the performance of three AI-enabled systematic review platforms in replicating three recently published systematic reviews

- Protocol pre-registered on OSF 1st July 2025 (https://osf.io/8nf6h/files/qadcw)

THE UNIVERSITY of EDINBURGH
Institute for Neuroscience and Cardiovascular Research

Methods: SR selection

- Identify SR published in past year by
  - CAMARADES (non-human animal SR)
  - JBI
  - Cochrane
- Dual screen SRs against inclusion criteria – SR containing quantitative analysis
- Random selection of 1 SR per category from included SR

THE UNIVERSITY of EDINBURGH
Institute for Neuroscience and Cardiovascular Research

Methods: SRP evaluation

We selected three Systematic Review Platforms (SRP) with embedded AI techniques for evaluation.

Elicit.AI
Nested Knowledge
Scispace

We considered discussions from other researchers/collaborators/evidence synthesis communities, available evaluation data, and publications citing platforms in selection of SRP

We perform evaluation of these platforms on each SR in a 3x3 design.

	SR1	SR2	SR3
SRP1	Evaluation 1	Evaluation 4	Evaluation 7
SRP2	Evaluation 2	Evaluation 5	Evaluation 8
SRP3	Evaluation 3	Evaluation 6	Evaluation 9

THE UNIVERSITY of EDINBURGH
Institute for Neuroscience and Cardiovascular Research

What do they claim?

Elicit.AI

Scale

Elicit can find up to 1,000 relevant papers and analyze up to 20,000 data points at once.

Accuracy

Elicit is the most accurate AI product for scientific research. Learn about how we validated Elicit's accuracy.

Transparency

Elicit supports all AI-generated claims with sentence-level citations from the underlying sources.

More than chat

Elicit goes beyond chat to provide rich, interactive tables and multi-step workflows.

Nested Knowledge

Powerful evidence synthesis tools for medical researchers. Accelerate, collaborate, automate and share.

Decisions at the speed of AI: Get answers in days, not months, with Nested Knowledge®, your all-in-one solution for efficient literature reviews.

Start with AI, Dive Deeper with Experts

Scispace

Feature	SciSpace Agent
Primary focus	Academic research workflows (writing, reviewing, data analysis, etc.)
Designed for	Researchers, PhD students, R&D professionals, and other academicians
Knowledge base	280M+ papers, 50M+ full-text PDFs
Use case coverage	Manuscript writing, peer review, grant proposal drafting, and more!
Accuracy wrt academic use cases	High (Deep integration with scholarly data)
Integration with academic databases	Integrated with 150+ academic tools & databases

Platform features on versions available in July 2025

Elicit (Team subscription; USD 65/month/user)

- User input research question
- Elicit searches, retrieves, deduplicates
  - Max 500 papers
  - ACS, ACSO, Pubmed, OpenAlex, Semantic Scholar
  - Optional manual upload up to further 500 papers
- Users define screening criteria
- Elicit screens citations; user defines score threshold
- Users define data extraction columns
- Elicit extracts data
- Elicit generates research report
  - Max 40 papers
  - No quantitative analysis

NK (Academic subscription ~£18/month/user; *Enterprise $695/month/user)*

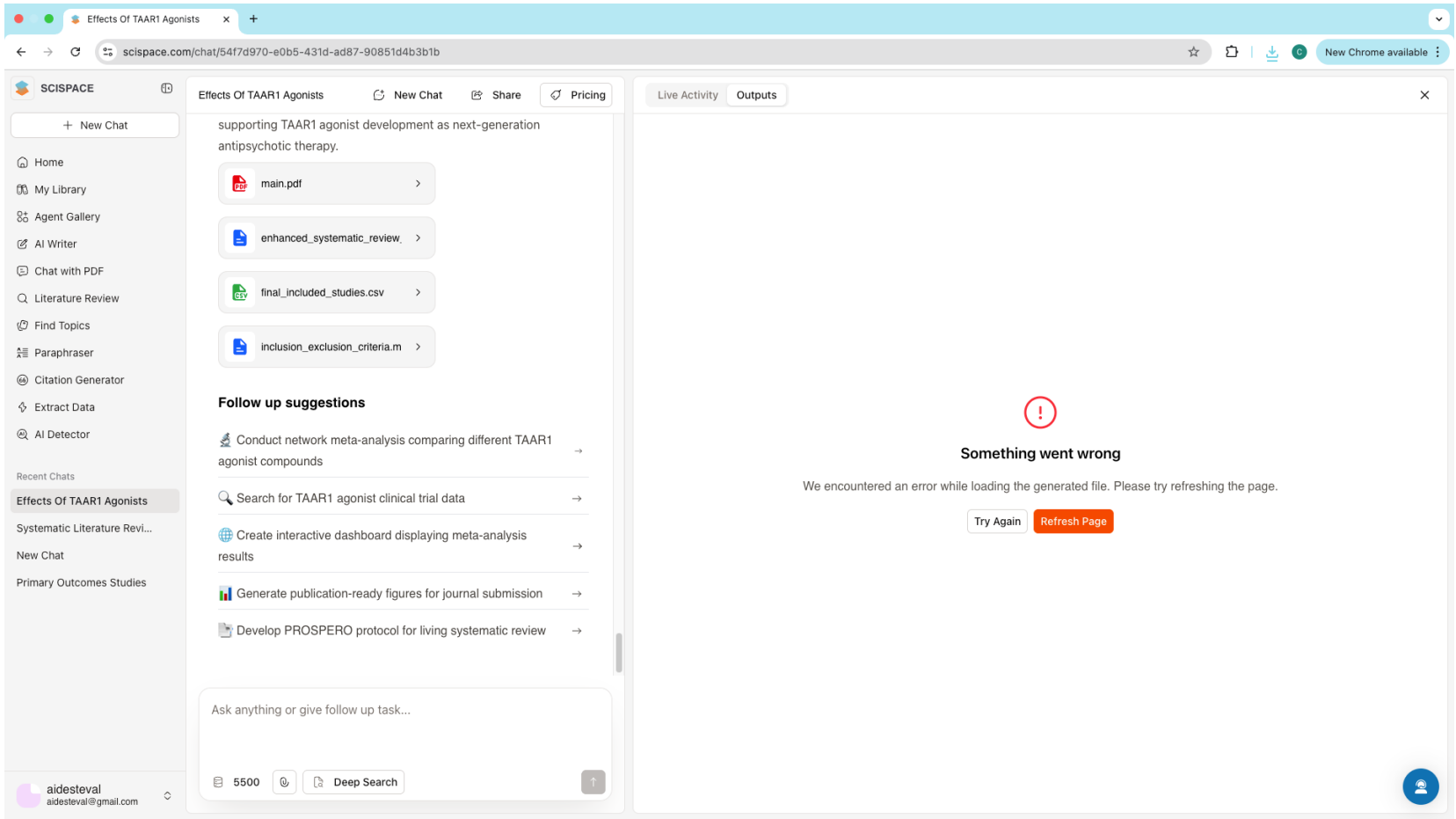
- User input search strings for selected databases* Manual uploads for other databases
  - *Automatic updates: Pubmed, DOAJ, ClinicalTrials.gov, EuropePMC, Maude
  - Max total = 10K citations/Nest
  - *Smart search: User input research question. NK crafts Pubmed queries. User finalise.*
- NK deduplicates
  - Duplicate queue available for review
- Screening: single vs dual stage; single vs dual reviewer
  - Robot Screener for abstract dual screen (min: 50 dual screened decisions, 10 advanced to enable). performance metrics provided
- Users define tag hierarchy and tags.
  - *Core Smart Tags and Adaptive Smart Tags – NK AI suggest tags, human reviews tags*
- User performs data extraction
  - *Smart MA extraction from tables/text (<30 publications). User reviews and submit*
- NK reports data inc. MA

Scispace (Advanced subscription; USD 90/month/user)

- User input research question
- Scispace searches and deduplicates
  - OpenAlex, Semantic Scholar, Google Scholar and other trusted repositories (>200 million papers)
- Scispace agent: AI chat bot to interact
- Scispace generates outputs: report, code, data, figures

Data export issues with Scispace

https://scispace.com/chat/54f7d970-e0b5-431d-ad87-90851d4b3b1b



Data export issues with Scispace

PMID	Title	Abstract	Journal	Authors	Publicatior	DOI
dimethoxy	'Fribourg M.	Moreno J.L	Holloway T	Provasi D.	Baki L.	Mahajan R
	A growing understanding of the role of muscarinic receptors in the r	Pre-clinical models, postmortem and neuroimaging studies				
22641180	A new perspective for schizophrenia: TAAR1 agonists reveal antipsyc	Schizophre	Molecular	Revel, Mor	2013	10.1038/mp
	A new perspective for schizophrenia: TAAR1 agonists reveal antipsych	otic-and antidepressant-like activity, improve cognition an				
30366004	Activation of trace amine-associated receptor 1 attenuates schedule	Trace Amin	Neurophar	Sukhanov,	2019	10.1016/j.n
26441502	An animal model of differential genetic risk for methamphetamine i	The questic	Frontiers ir	Phillips, Sh	2015	10.3389/fr
	Avenues for the development of therapeutics that target trace amine	associated receptor 1 (TAAR1) miniperspective				
29520239	Behavioral Phenotyping of Dopamine Transporter Knockout Rats: C	Alterations	Frontiers ir	Cinque, Zor	2018	10.3389/fp
	Beyond dopamine receptor antagonism: new targets for schizophre	Treatment of schizophrenia (SCZ) historically relies on the us				
	Binding of SEP-363856 within TAAR1 and the 5HT1A receptor: impli	Current medications for schizophrenia typically modulate d				
22763617	Brain-specific overexpression of trace amine-associated receptor 1 a	Trace amin	Neuropsycl	Revel, Mey	2012	10.1038/n

Data export issues with Scispace

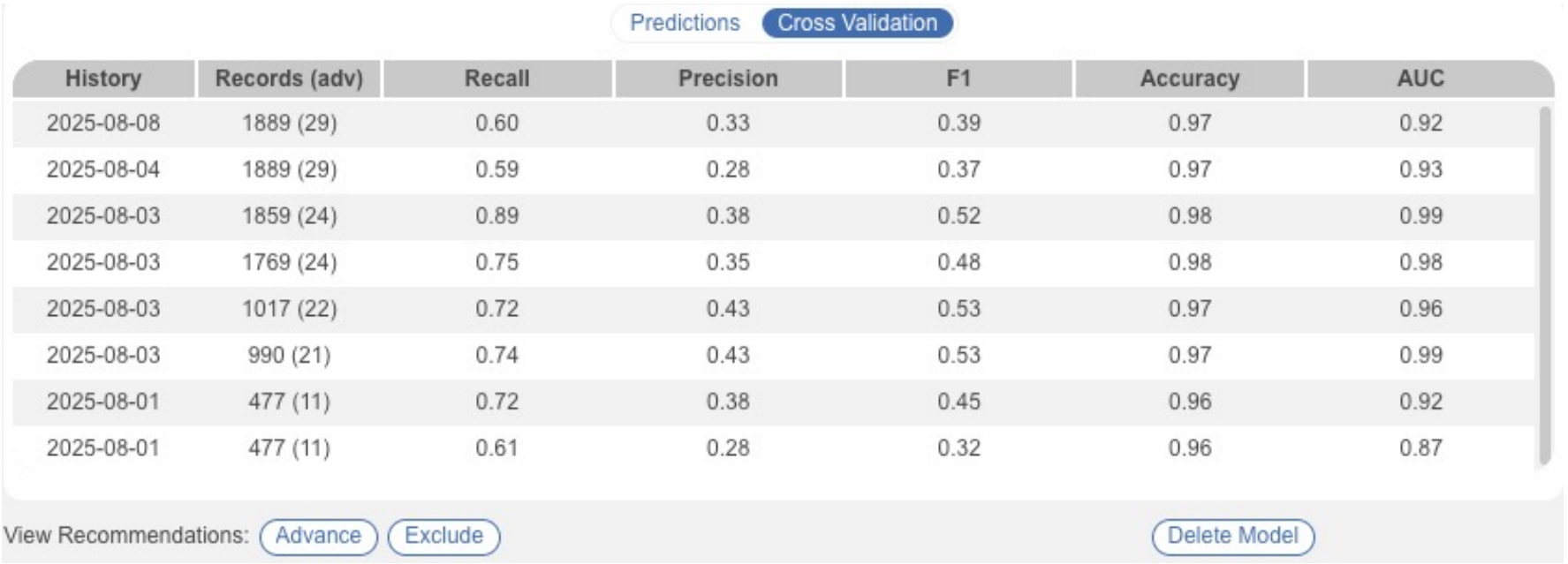
PMID	Title	Abstract	Journal	Authors	H
22641180	A new perspective for schizophrenia: TAAR1 agonists reveal antipsychotic- and antidepressa	Schizophreni	Molecular ps	Revel, Morea	
	A new perspective for schizophrenia: TAAR1 agonists reveal antipsychotic-and antidepressant-like activity, improve cognition and con				
	Binding of SEP-363856 within TAAR1 and the 5HT1A receptor: implications for the design of n	Current medications for schizophrenia ty			
	Effect of co-treatment of olanzapine with SEP-363856 in mice models of schizophrenia	Olanzapine is a commonly used drug in th			
	In-vivo pharmacology of trace-amine associated receptor 1				
32382134	Reproducing the dopamine pathophysiology of schizophrenia and approaches to ameliorate	Patients with	Molecular ps	Kokkinou, Irvi	
	SEP-363856, a novel psychotropic agent with a unique, non-D2 receptor mechanism of actic	For the past 50 years, the clinical efficacy			
36100653	TAAR1 dependent and independent actions of the potential antipsychotic and dual TAAR1/5-	SEP-383856 (	Neuropsycho	Saarinen, Ma	
	TAAR1 dependent and independent actions of the potential antipsychotic and dual TAAR1/5-	HT1A receptor agonist SEP-363856			
	The potential of trace amines and their receptors for treating neurological and psychiatric d	Mining of the human genome has revealed			
	The trace amine-associated receptor 1 modulates methamphetamine's neurochemical and	The newly discovered trace amine-associ			
	Trace amine associated receptor 1 modulates behavioral effects of ethanol	Background: Few treatment options for al			
	Trace amine-associated receptor 1 (TAAR1) agonism as a new treatment strategy for schizophrenia and related disorders				
	Trace amine-associated receptor 1 as a target for the development of new antipsychotics: c	Schizophrenia is a mental illness associat			
	Trace amine-associated receptor 1 contributes to diverse functional actions of O-Phenyl-Iodotyramine in mice but not to the effects o				
	Trace amine-associated receptor 1 modulation of dopamine system	Trace amine-associated receptor 1 (TAAR			
	Trace amine-associated receptor 1 partial agonism reveals novel paradigm for neuropsychiatric therapeutics				
	Trace amine-associated receptor 1: a multimodal therapeutic target for neuropsychiatric di	Introduction: The trace amines, endogenc			
	Unlocking the therapeutic potential of ulotaront as a trace amine-associated receptor 1 agc	All antipsychotics currently used in clinic			

Search and screening results for SR1

non human SR in “Trace amine-associated receptor 1 (TAAR1) agonism for psychosis: a living systematic review and meta-analysis of human and non-human data” https://wellcomeopenresearch.org/articles/9-182 15 FT included in original review

Original review - FT included	SRP decision	Elicit	NK (full text dual-human screened; Robot Screener used for TiAb dual screen)	NK Robot screener (TiAb)	Scispace
In	In	6	15	5	6
In	Out	9	0	2	9
Out	In	21	6	14	11
Sensitivity		40%	100%	71%	40%
Precision		22%	94%	26%	35%

Nested Knowledge Robot Screener Internal Cross Validation Metrics for SR1



How the Screening Model Works

At a high level, the model is a Decision Tree- a series of Yes/No questions about characteristics of records that lead to different probabilities of inclusion/advancement.

In more detail, the model is a gradient-boosted decision tree ensemble. Its hyperparameters, particularly around model complexity (number of trees, tree depth) are optimized using a cross validation grid search. The model produces posterior probabilities and is optimized on logistic loss. SMOTE oversampling is employed as a correction to highly imbalanced classes frequently seen in screening.

SR1: Additional inclusions

Publications before original search performed	Appropriate inclusion	Elicit	NK	Scispace
TRUE	Yes	2[a,b]	1[b]	0
	No	15	0	11
FALSE	Yes	1	5	0
	No	3	0	0

(a) 1 paper was identified within the TAAR1 agonist review update 2024 via psychosis-SOLES. (b) 1 paper was identified in the original search but excluded following human reviewer dual screening – this paper was included following further review by senior reviewer prompted by this evaluation.

Search and screening results for SR2

JBI review - “Comparison of diagnostic accuracy of rapid antigen tests for COVID-19 compared to the viral genetic test in adults: a systematic review and meta-analysis. DOI: 10.11124/JBIES-23-00291
Original review screening stages – TiAb / full text / critical appraisal. 142 publications included in analysis

Original review - FT included	SRP decision	Elicit	NK (full text dual-human screened; Robot Screener used for TiAb dual screen)	NK Robot screener (TiAb)	NK Robot screener against original TiAb
In	In	58	99	111	323
In	Out	84	43	37	256
Out	In	367	85	549	337
Sensitivity		40%	70%	75%	56%
Precision		14%	54%	17%	49%

SR2-Elicit

JBI review - “Comparison of diagnostic accuracy of rapid antigen tests for COVID-19 compared to the viral genetic test in adults: a systematic review and meta-analysis. DOI: 10.11124/JBIES-23-00291

Included publications by Elicit not within original review		
Publications before original search performed (Last update 12 July 2022)	Appropriate inclusion	N
TRUE	Yes	50
	No	116
FALSE	Yes	64
	No	23

Multiple duplicates within Elicit search (5 using DOI alone; 23 further if matching preprint to published DOI); Duplicates (including preprint with linked publications) have been removed for this analysis.

Annotations

Percentage of tags in original review accurately tagged by SRP

Evaluation	Population	Intervention	Comparison	Outcomes	ROB	Overall accuracy
SR1-Elicit	60.0%	55.6%	66.7%	NA	51.7%	54.5%
SR1-NK	53.3%	33.3%	66.7%	20.0%	NA	43.3%
SR2-Elicit	88.9%	85.8%	61.7%	72.0%	55.1%	70.3%

Percentage of publications given additional tags by SRP (percentage appropriate in parenthesis)

SR1-Elicit	80% (0%)	55.6% (20%)	66.7% (0%)	NA (NA)	0% (0%)
SR1-NK	26.7% (75%)	60% (11%)	25% (33%)	73.3% (45%)	NA
SR2-Elicit	3.9% (0%)	5.8% (0%)	0% (0%)	12.1% (61%)	44.8% (6%)

Challenges in tagging pharmacologically induced models (assigning intervention rather than population tag); tagging unrelated terms in rest of text.

Many of additional outcomes tagged by SRPs were not primary outcomes of original SR and hence were appropriately not tagged there.

Multiple tags per category and hierarchical tagging likely to affect planned calculations of kappa agreement.

Discussion: AI-enabled systematic review platforms

- Variation in
  - Performance in SR tasks – sensitivity an issue currently across platforms relying on AI methods; annotation/extraction variable; analysis limited
  - Transparency
  - Explainability
  - Human supervision / Customisability
  - Hallucinations
  - Interoperability / export of code/data/metadata
  - Scalability
  - Costs
- Generalisability: need more studies

THE UNIVERSITY of EDINBURGH
Institute for Neuroscience and Cardiovascular Research

Discussion: Evaluating systematic review platforms

- Challenging!
- Challenges with replicating SRs
  - Data and metadata FAIRness can affect replication (e.g. lack of bibliographic data of excluded studies; incomplete/inaccurate metadata; systematic search strategies / results / bibliographic data in PDF format)
  - SR selection matters: Consider size, scope, how well questions are defined vs feasibility / resources
- Challenges with evaluating SRPs
  - Variation in extent of human effort needed to perform SR in SRPs
  - Lack of gold standards on data, metadata / code, format leads to significant variation in SRP outputs; cleaning/extracting data for evaluation of SRP can be time and resource intensive
  - Results limited to version of SRP evaluated; quickly superseded

THE UNIVERSITY of EDINBURGH
Institute for Neuroscience and Cardiovascular Research

Discussion: Looking ahead (1)

- Responsible use of AI in evidence synthesis crucial. See RAISE guidance.
- Each SR provides data and opportunity to validate AI digital evidence synthesis tools both retrospectively and prospectively – we need to maximise this.
  - increasing FAIRness of data/metadata at each step of the SR to improve replicability:
  - Roles of individuals, groups, community, institutions
  - Systems/infrastructure to maximise this – using ‘routinely collected’ data in SRs for evaluation of AI evidence synthesis tools
- Methods development in evaluating AI-enabled tools and platforms

THE UNIVERSITY of EDINBURGH
Institute for Neuroscience and Cardiovascular Research

Further questions

.

- Should SRPs be regulated? if so, who by? (*possible trade off with innovation?*)
- Who should be responsible for evaluation of tools and platforms? Cf. onus on drug developers/pharma to demonstrate evidence of efficacy/safety for regulatory submissions
- Ethical, legal and environmental implications of use of these platforms?

Disclaimers

THE SERVICE AND ALL RELATED MATERIALS AND CONTENT ARE PROVIDED “AS-IS” AND “AS AVAILABLE,” AND WE DISCLAIM ANY AND ALL WARRANTIES, WHETHER EXPRESS OR IMPLIED, INCLUDING WITHOUT LIMITATION IMPLIED WARRANTIES OF TITLE, MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, AND NON-INFRINGEMENT. WE CANNOT GUARANTEE AND DO NOT PROMISE AVAILABILITY OF THE SERVICE OR ANY SPECIFIC RESULTS FROM USE OF THE SERVICE.

WE ARE NOT RESPONSIBLE FOR ANY OF YOUR CONTENT OR ANY OTHER USER’S CONTENT, INCLUDING CONTENT THAT VIOLATES THESE TERMS OF SERVICE OR THAT IS INCORRECT, INCOMPLETE, INACCURATE, OR OFFENSIVE, WHETHER SUCH CONDITION IS CAUSED BY USERS OF THE SERVICE, MEMBERS, OR BY ANY OF THE SOFTWARE OR INFRASTRUCTURE ASSOCIATED WITH OR USED TO PROVIDE THE SERVICE. WE ARE NOT RESPONSIBLE FOR ANY THIRD PARTY’S SERVICES OR PERFORMANCE, INCLUDING ANY THIRD PARTY PROVIDING SERVICES, DATA, OR INFRASTRUCTURE FOR THE SERVICE.

UNDER NO CIRCUMSTANCES WILL WE BE RESPONSIBLE FOR ANY LOSS OR DAMAGE, INCLUDING PERSONAL INJURY OR DEATH, RESULTING FROM ANYONE'S USE OF THE SERVICE, ANY CONTENT POSTED ON THE SERVICE OR TRANSMITTED TO OTHER USERS, OR ANY INTERACTIONS BETWEEN USERS OF THE SERVICE, WHETHER ONLINE OR OFFLINE.

Acknowledgements

Harris Khalid, Kat Stefanska, Florence Do, Mohammed Abomostafa, Gemma Dalgety, Malcolm Macleod

Screeners for Nested Knowledge review:

Leonor Rodriguez, Matheus Gallas-Lopez, Stephen Malden, Kaitlyn Hair, Gillian Currie, Yvonne Khor, Nadia Soliman, Khrystyna Stoska

charis.wong@ed.ac.uk