VALAWAI
TOWARDS VALUE-AWARE AI

VALAWAI

Value-Aware Artificial Intelligence

Giulio Prevedello, Pietro Gravino, Martina Galletti, Lara Verheyen, Remi Van Trijp
Emanuele Brugnoli, Ruggiero Lo Sardo, Vittorio Loreto

Sony CSL

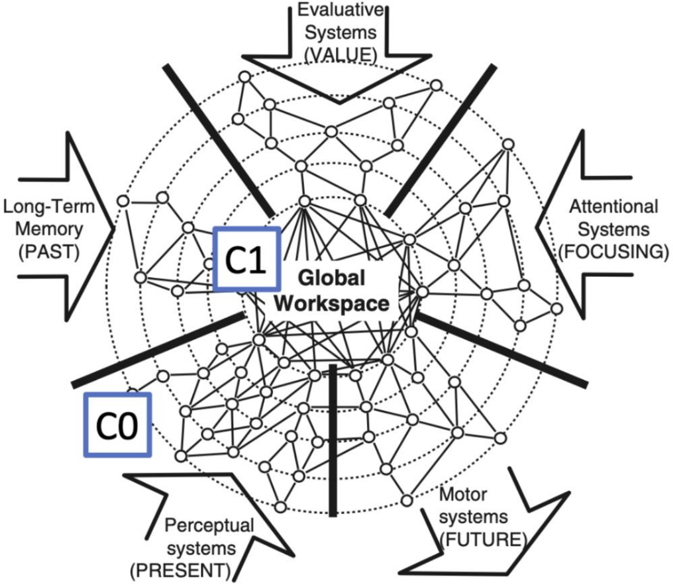
AI-PHI (S2E1)
Paris, 13th February 2025

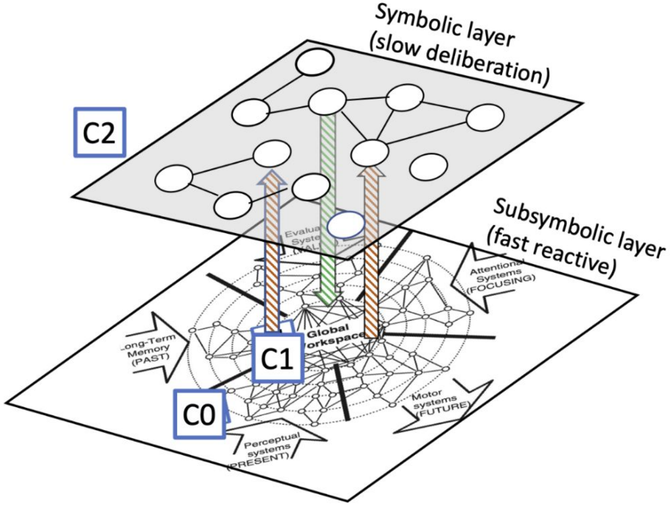
CSIC IIIA-CSIC INSTITUT D'INVESTIGACIÓ EN INTEL·LIGÈNCIA ARTIFICIAL | IMIM Parc de Salut MAR Barcelona Institut Hospital del Mar d'Investigacions Mèdiques | Sony CSL | UNIVERSITEIT GENT | iit ISTITUTO ITALIANO DI TECNOLOGIA | Studio Stelluti

Theoretical framework

Implement a general operational model inspired by the Global Neuronal Workspace* (C0+C1)

Develop a framework for value-aware situation analysis and decision (C2) and prove its utility on real applications


Evaluative Systems (VALUE)
Long-Term Memory (PAST)
Attentional Systems (FOCUSING)
C1
Global Workspace
C0
Perceptual systems (PRESENT)
Motor systems (FUTURE)


Symbolic layer (slow deliberation)
C2
Subsymbolic layer (fast reactive)
Attentional Systems (FOCUSING)
Long-Term Memory (PAST)
Global Workspace
C1
C0
Motor systems (FUTURE)
Perceptual systems (PRESENT)

*Dehaene, S., J-P Changeux and L. Naccache (2011) *The Global Neuronal Workspace Model of Conscious Access: From Neuronal Architectures to Clinical Applications.* Research and Perspectives in Neurosciences.
*Seth, A. K., & Bayne, T. (2022). Theories of consciousness. Nature Reviews Neuroscience

Objectives of Application

- Demonstrate the utility of vale-aware systems in three domains:
  - Social robots
  - Medical protocols
  - **Social media**

- Tackles problems of news fruition in the social media:
  - Polarisation
  - Adversarial behaviour

- Promote users' ethical literacy and awareness
  (plurality in public debate, moral consequences of behaviour, etc.)

Moral Values and Polarisation

Moral value = preference over world-states, in terms of right or wrong

Moral Foundations Theory[1]

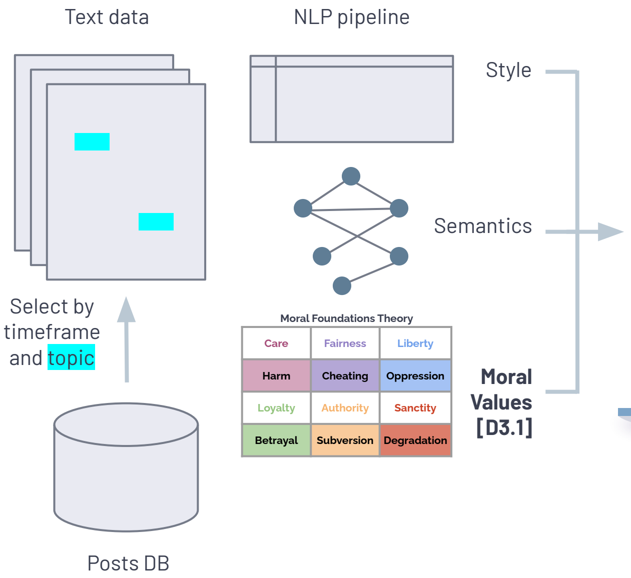
- A psychological theory based on innate mental structures universal to all humans, selected through evolutionary mechanisms
- Independent of social background
- Correlated with political sides and animosity online[2]

Care	Fairness	Loyalty	Authority	Purity
Harm	**Cheating**	**Betrayal**	**Subversion**	**Degradation**

MFT should be useful to identify polarisation[3] in the public debate

[1] Graham J., *et al.* (2013). Moral foundations theory: The pragmatic validity of moral pluralism. Advances in experimental social psychology.
[2] Rathje S., *et al.* (2021). Out-group animosity drives engagement on social media. Proceedings of the National Academy of Sciences.
[3] Jost J. T., *et al.* (2022). Cognitive-motivational mechanisms of political polarization in social-communicative contexts. Nature Reviews Psychology.

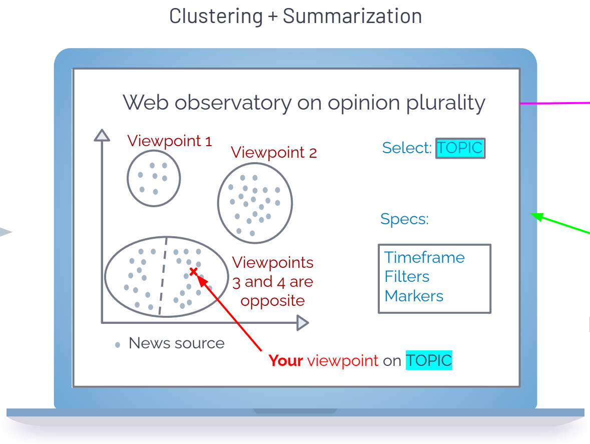
Social Media Observatories



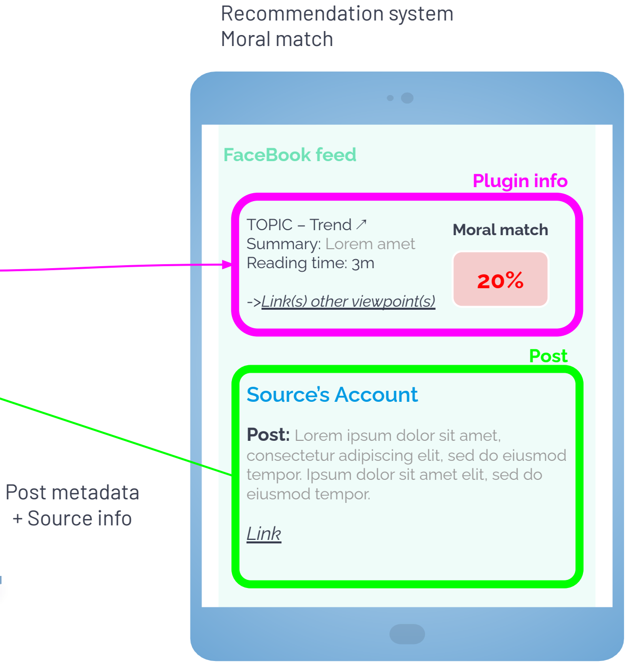
Moral Foundations Theory

Care	Fairness	Liberty
Harm	Cheating	Oppression
Loyalty	Authority	Sanctity
Betrayal	Subversion	Degradation

Moral Values [D3.1]



Web interface for public discourse analysis [D3.2]



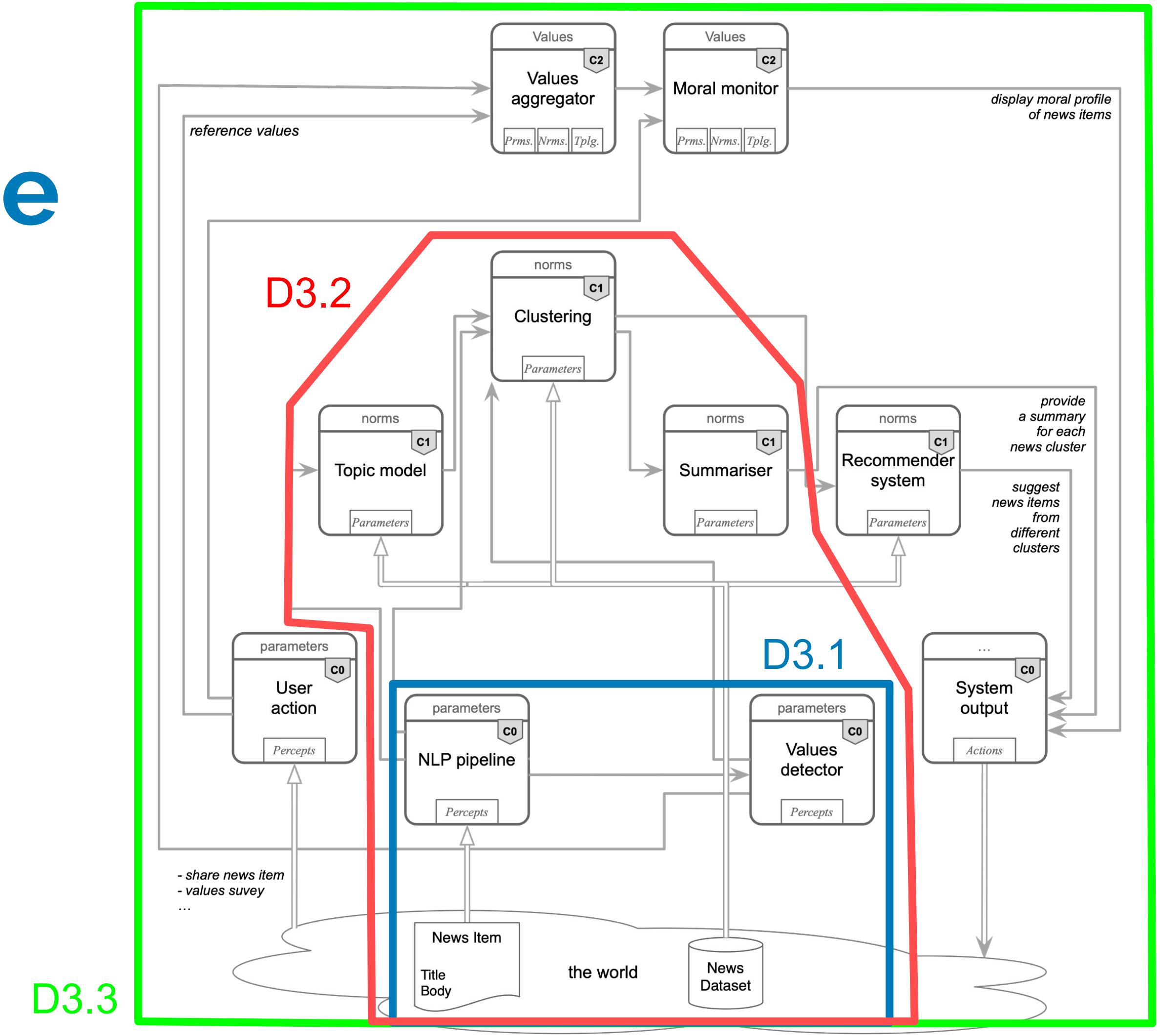
Recommendation system
Moral match
Post metadata + Source info
FaceBook feed
Plugin info
TOPIC – Trend ↗
Summary: Lorem amet
Reading time: 3m
->Link(s) other viewpoint(s)
Moral match
20%
Post
Source's Account
Post: Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor. Ipsum dolor sit amet elit, sed do eiusmod tempor.
Link

Social media plug-in [D3.3]

Goals

- Enable a contextualised fruition based on values
- Moral monitoring adds friction to misaligned actions
- Easy access to other viewpoints to bridge communities

RGNW architecture

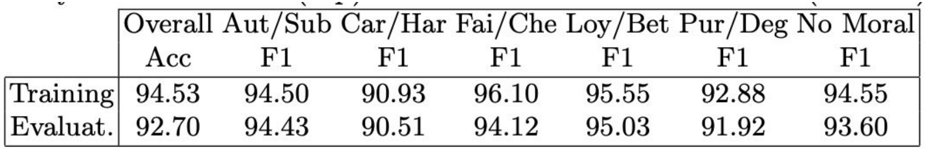


Moral Model – Deep Learning

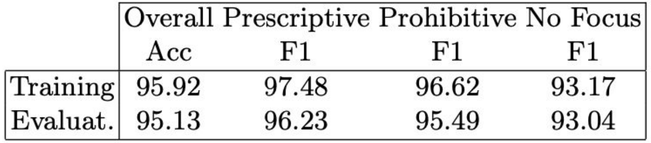
End-to-end approach

Main takeaways:

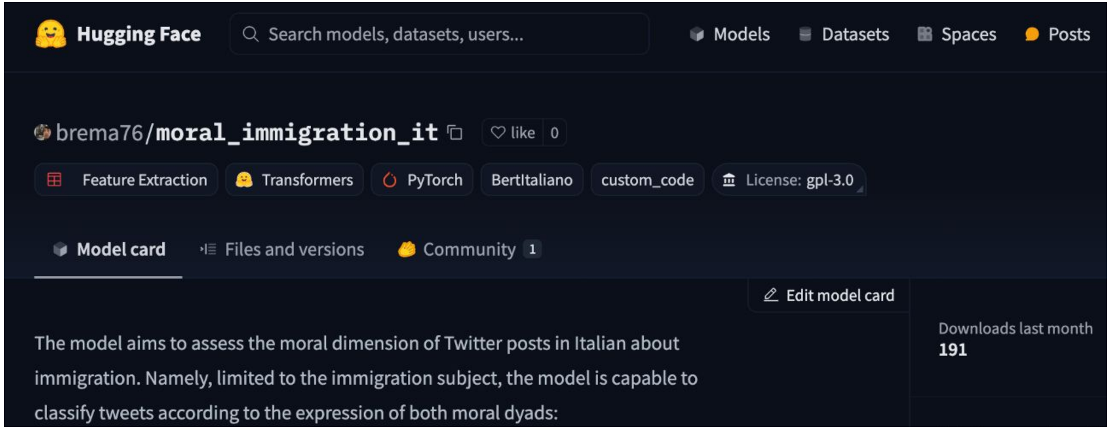
- Very accurate
- Limited to topics in the training set*
- Automatically annotate our Twitter IT dataset



	Overall Acc	Aut/Sub F1	Car/Har F1	Fai/Che F1	Loy/Bet F1	Pur/Deg F1	No Moral F1
Training	94.53	94.50	90.93	96.10	95.55	92.88	94.55
Evaluat.	92.70	94.43	90.51	94.12	95.03	91.92	93.60



	Overall Acc	Prescriptive F1	Prohibitive F1	No Focus F1
Training	95.92	97.48	96.62	93.17
Evaluat.	95.13	96.23	95.49	93.04



brema76/moral_immigration_it

The model aims to assess the moral dimension of Twitter posts in Italian about immigration. Namely, limited to the immigration subject, the model is capable to classify tweets according to the expression of both moral dyads:

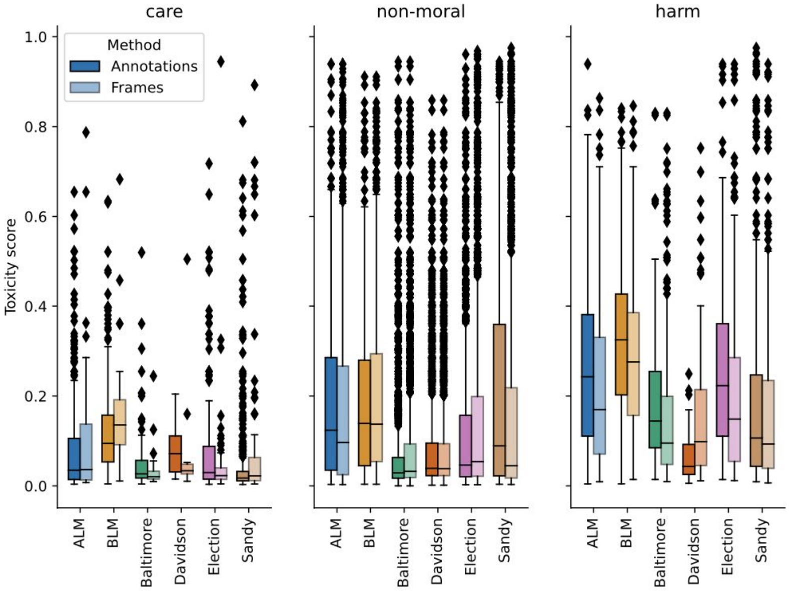
*Stranisci M., *et al.* (2021). The expression of moral values in the twitter debate: a corpus of conversations. IJCoL
Brugnoli E., Gravino P., Prevedello G. (2024). Moral values in social media for disinformation and hate speech analysis. LNCS .

Moral Model – Frame Extraction

Topic-scalable solution

Main takeaways:

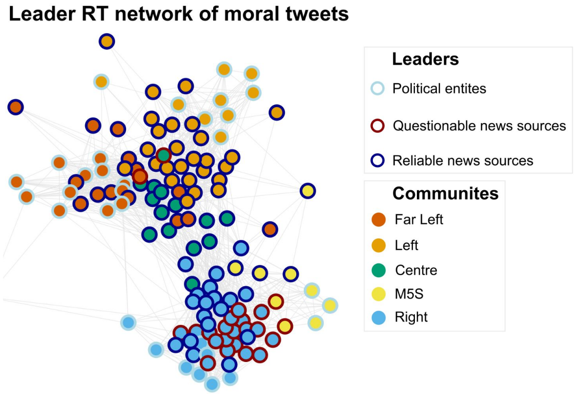
- Low sensitivity
- High specificity
- Levels of toxicity comparable to manual annotations
- FCG limited to frames for few values



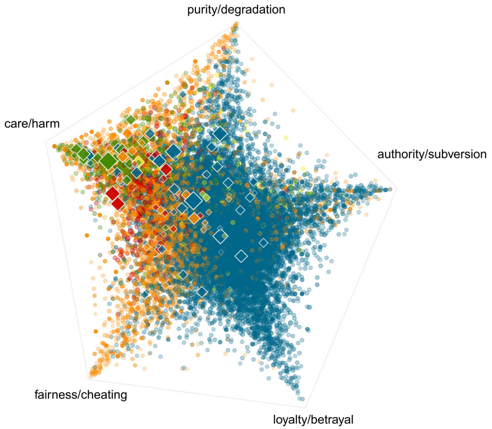
*De Giorgis, Stefano, et al. (2022) Basic human values and moral foundations theory in valuenet ontology. IC on knowledge engineering and knowledge management
*Hoover, J., et al. (2020). Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. Social Psychological and Personality Science
Prevedello G., Verheyen L., Brugnoli E., Lo Sardo R., Van Trijp R. (2024). Adversarial Behavior in Moral Value Expression: A Statistical and Frame-Semantic Analysis of Social Media. VECOMP workshop @ ECAI24

MFT for Public Discourse Analysis

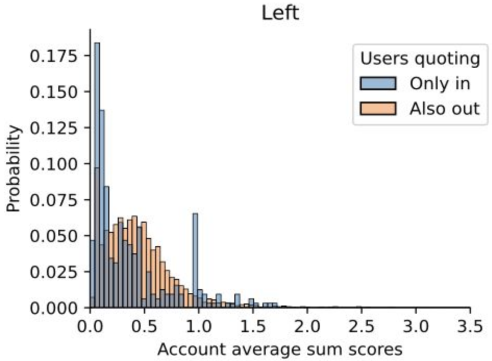
Tweets from Italian news sources and political organisations accounts (~95% of online engagement 2018-2022) with retweets and quotes.



Improved clustering



Moral expression profile



In-group out-group bias

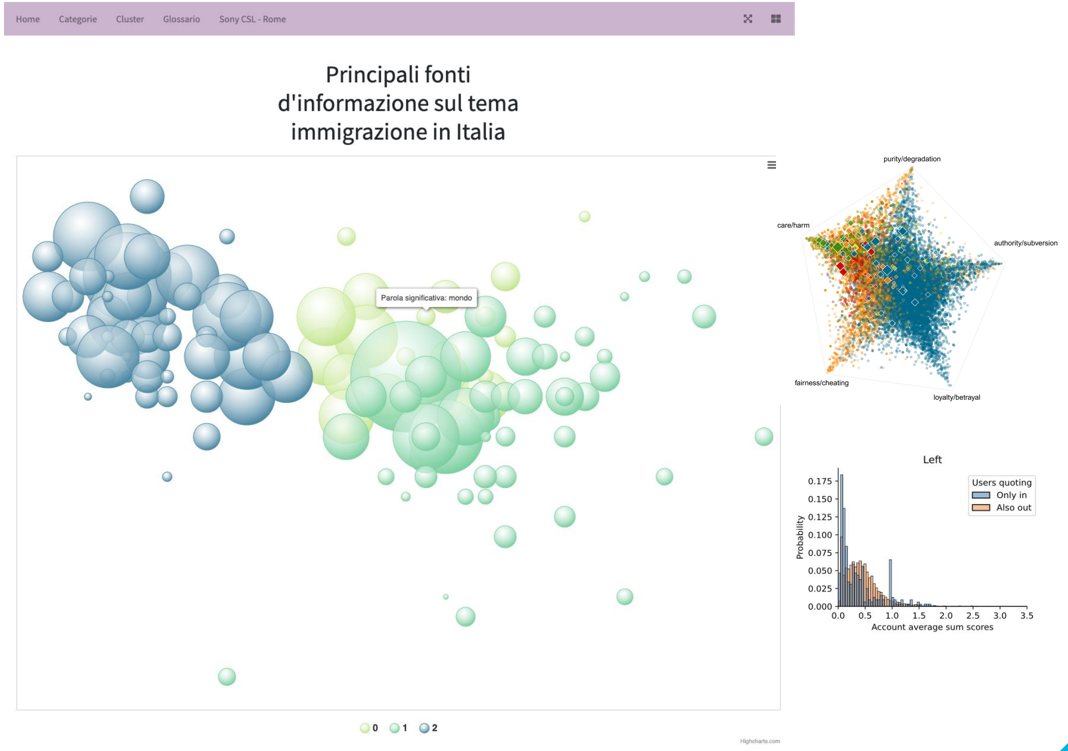
Brugnoli, E., Gravino, P., Sardo, D. R. L., Loreto, V., & Prevedello, G. (2024). Fine-Grained Clustering of Social Media: How Moral Triggers Drive Preferences and Consensus. ICAART

D3.2 – Prototype

Pipeline:

- Moral values
- Keywords Extraction*
- Clustering
- Toxicity analysis

Tool for reporting on the status of public discourse.



Principali fonti d'informazione sul tema immigrazione in Italia

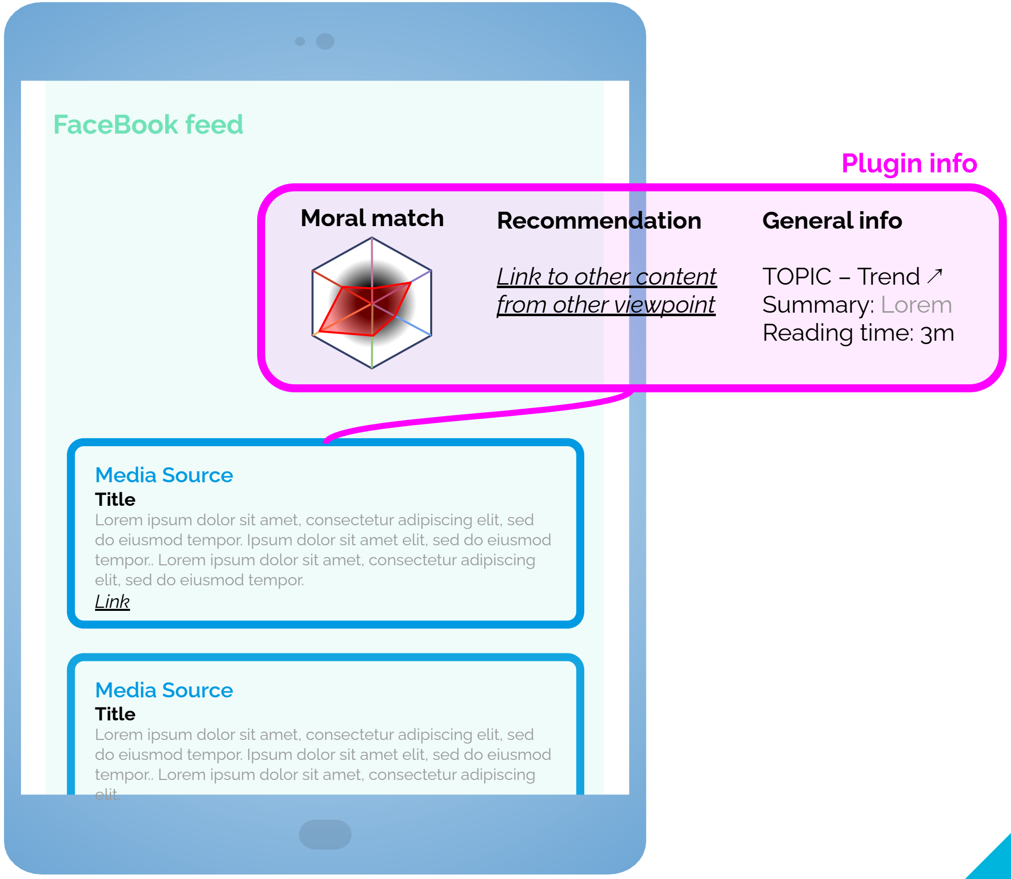
*Galletti, M., et al. (2025) Are Your Keywords Like My Queries? A Corpus-Wide Evaluation of Keyword Extractors with Real Searches." *Proceedings of the 31st COLING*.

Future Work [D3.3]

AIM: Design & test interventions

Plug-in for enhanced social media experience

- User-Content moral match display (raise awareness on moral agency)
- Recommend content from out-group with high chances of positive interaction (create bridges between communities)



FaceBook feed

Plugin info

Moral match

Recommendation

Link to other content from other viewpoint

General info

TOPIC – Trend ↗
Summary: Lorem
Reading time: 3m

Media Source
Title
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor. Ipsum dolor sit amet elit, sed do eiusmod tempor.. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor.
Link

Media Source
Title
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor. Ipsum dolor sit amet elit, sed do eiusmod tempor.. Lorem ipsum dolor sit amet, consectetur adipiscing elit.

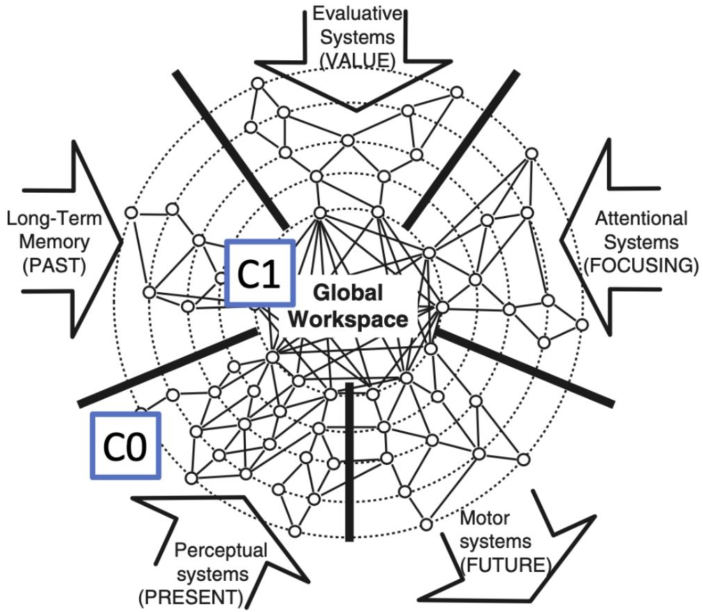
Awareness in AI

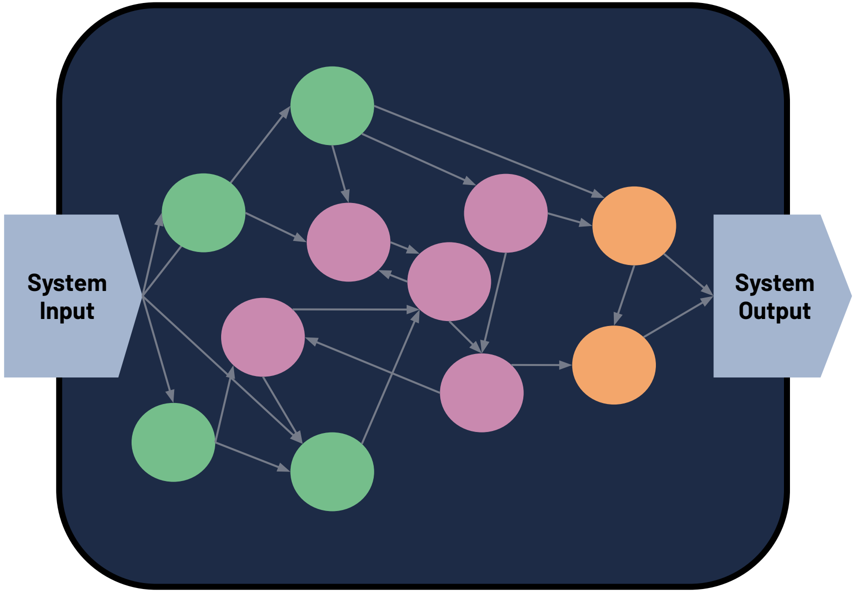
- Alignment < awareness < consciousness
  - Depends not only on output/behaviour, also on design and process

- Definition of awareness based on:
  - **Access of information**
  - Coherence of representation (internal to the system)
  - Expressibility of the representation

* Prevedello G., Marcos-Vidal L,, Brugnoli E., Gravino P. Notes on Measures for Information Access in Neuroscience and AI Systems AWAI *workshop at ICAART 2024*.

Assumption: network system


Evaluative Systems (VALUE)
Long-Term Memory (PAST)
Attentional Systems (FOCUSING)
C1
Global Workspace
C0
Perceptual systems (PRESENT)
Motor systems (FUTURE)



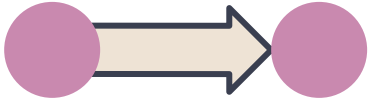
Dehaene, S., J-P Changeux and L. Naccache (2011) *The Global Neuronal Workspace Model of Conscious Access: From Neuronal Architectures to Clinical Applications.* Research and Perspectives in Neurosciences.

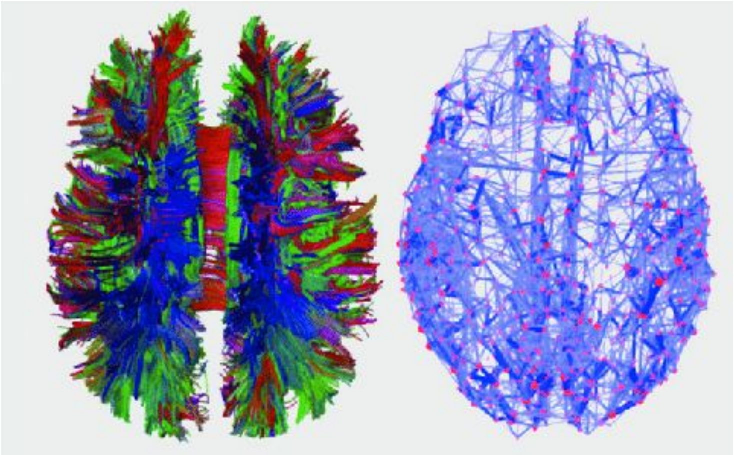
Structural connectivity - Brain

Where/Whether information can flow between components

Measured by targeting the **topology** of information flow

- Synapses establish links between neurons, enabling information transmission
- Mapping the structural connectome is still a challenge

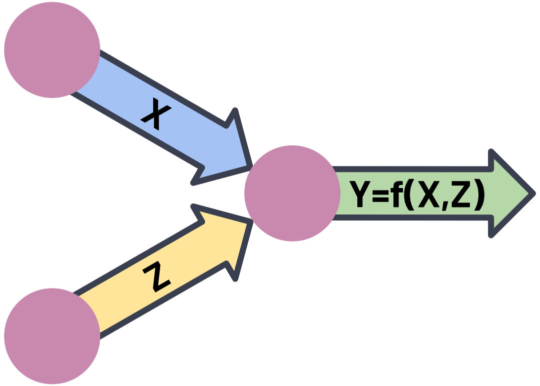


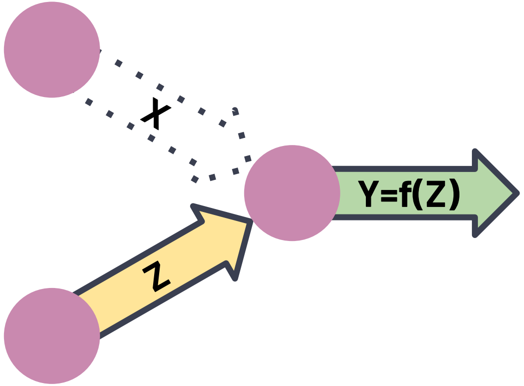


Sporns, O. (2013). Structure and function of complex brain networks. *Dialogues in clinical neuroscience*, 15(3), 247-262.

Structural connectivity - AI

- Available by design
- Accessible even if it changes during operations



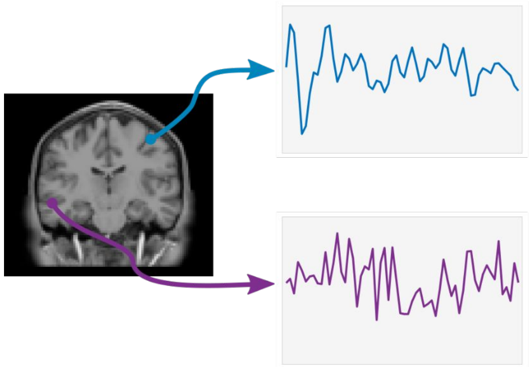


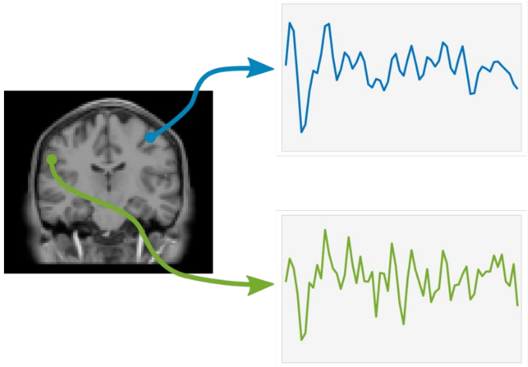
Functional connectivity - Brain

Where information flow affects downstream components

Measured by targeting the **statistical dependence** of information flow

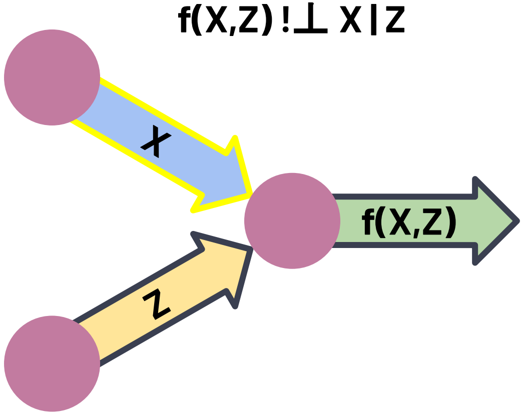
- Enables inference of functional connectome from brain activity (e.g., EEG)
- Synchronicity and correlation of different neural regions

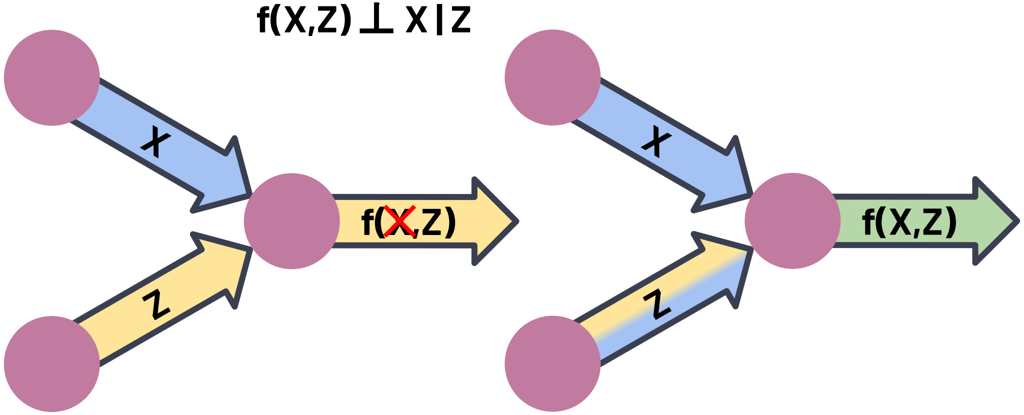




Functional connectivity - AI

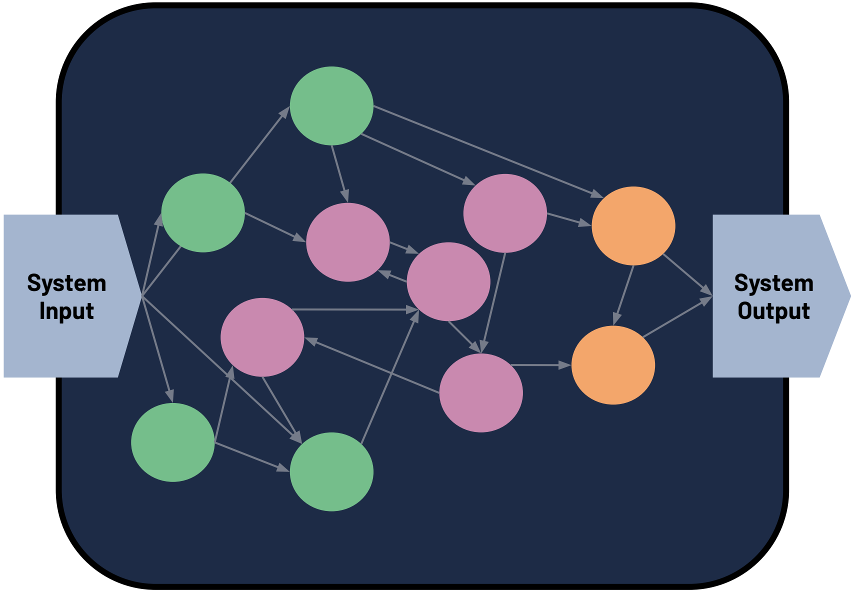
- Transmission of non-redundant information is locally measured with Conditional Mutual Information (CMI)
- CMI can assess functionality of all structural connections

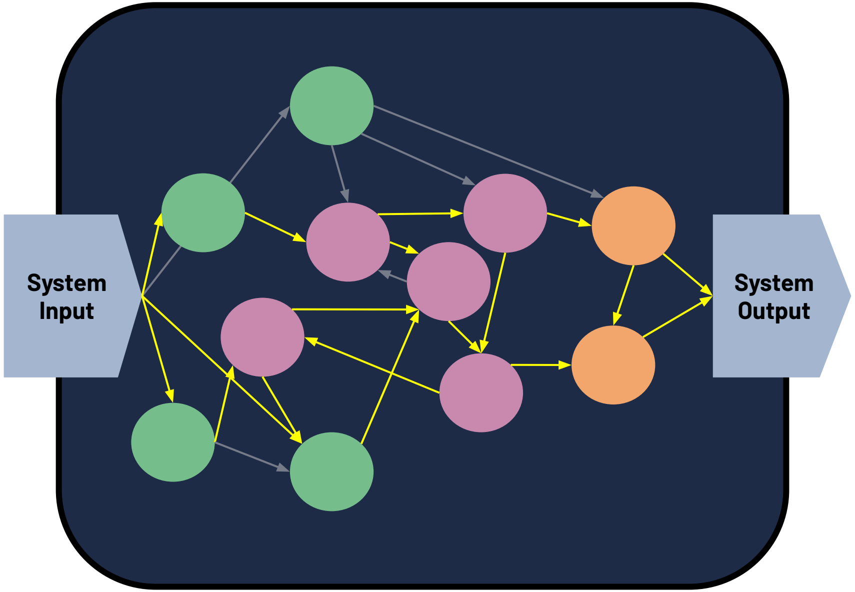




Functional connectivity - AI

Functional connectivity can be recovered iterating the CMI over structural network.





Operational connectivity - Brain

Where information is presently flowing

Measured by targeting **scalability and efficiency** of information flow

- Neuronal avalanches display power laws, hallmark of scalability
- Branching parameter of **σ** = 1 is an hallmark of optimal information transmission

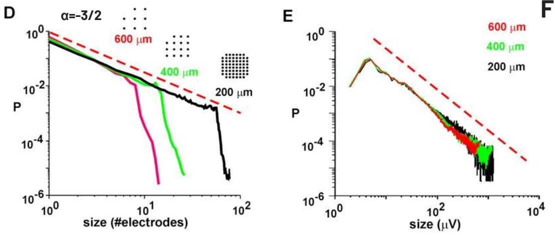


Fig 4

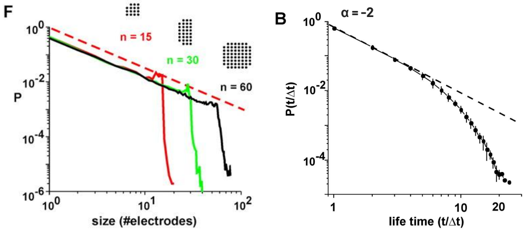
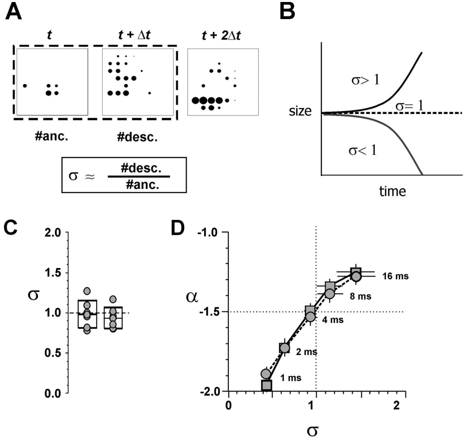


Fig 5

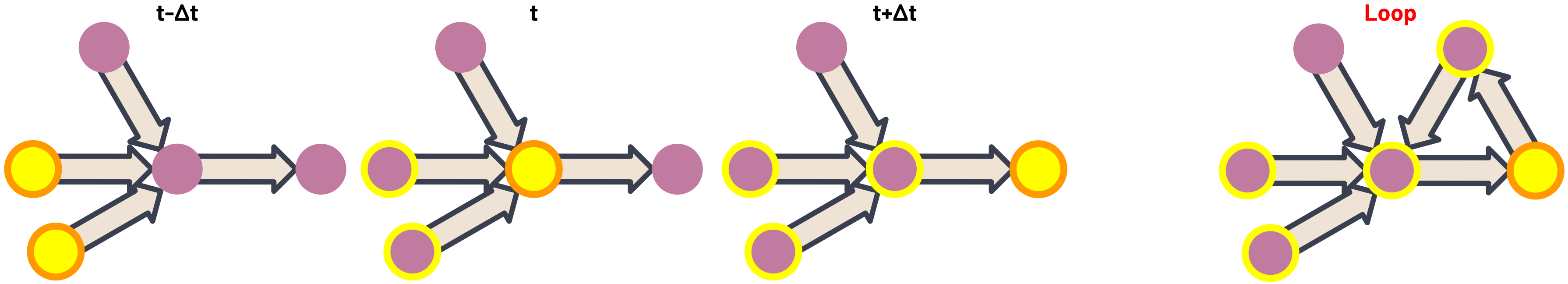
Fig 7



Beggs, J. M., & Plenz, D. (2003). Neuronal avalanches in neocortical circuits. *Journal of neuroscience*, *23*(35), 11167-11177.

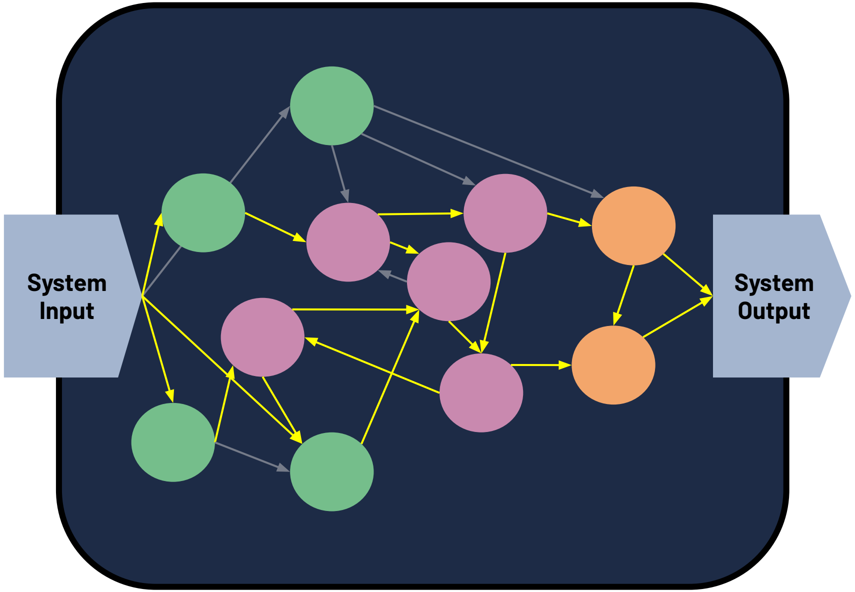
Operational connectivity – AI

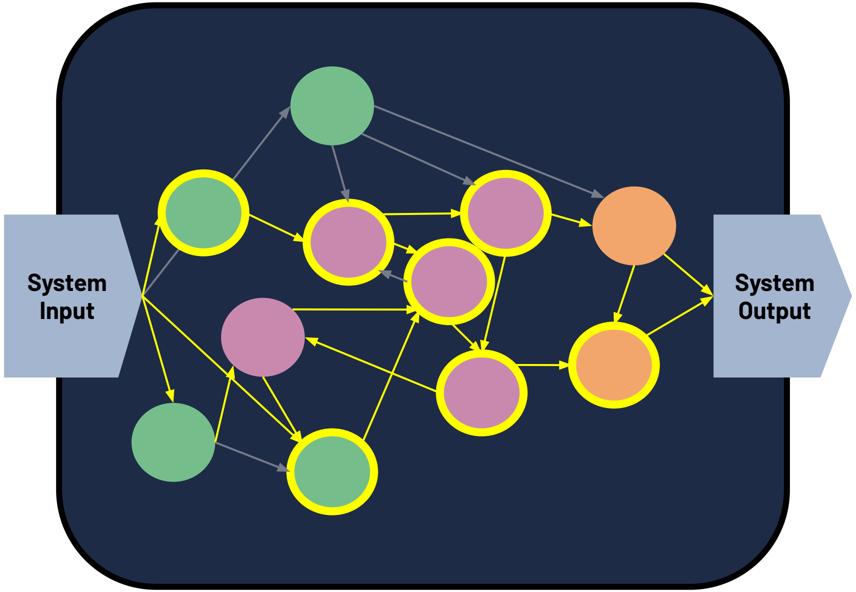
- Limited to systems whose components turn on/off interactively
- Relate input with activation pattern duration, size and propagation
- Highlight limitations and potential conflicts



Operational connectivity – AI

Operational connectivity can be recovered tracing the activation path for a given input.





Awareness in AI

- Alignment < awareness < consciousness
  - Depends not only on output/behaviour, also on design and process

- Definition of awareness based on:
  - Access of information
  - Coherence of representation (internal to the system)
  - Expressibility of the representation

WOULD LOVE TO HEAR YOUR OPINION

THANKS!

Any questions?

giulio.prevedello@sony.com

VALAWAI
TOWARDS VALUE-AWARE AI

valawai.eu

@valawaiEU

VALAWAI is funded by the EU's Horizon EIC programme under Grant Agreement # 101070930.