Introduction to the paper

"From birth to loss of representations in artificial neural networks"

CIFMA Workshop

NOVEMBER 2024, PHILIPP STECHER, EBERHARD-KARLS-UNIVERSITY TÜBINGEN, GERMANY

Today's agenda

Context

Deep Dive into the paper

Outlook

Why it's interesting to be attentive in the next ~25 minutes

Not exhaustive

For everyone

Details on the next slide

Cool topic

"Understanding what [representations] are, [...] will be essential for engineering non-brittle AI systems"

Mitchell M., *Abstraction and analogy-making in artificial intelligence.* Annals of the New York Academy of Sciences, 2021

What not to expect and …

A theory ready to be empirically validated

Causal effects or neural mechanisms

… what to expect

A young set of ideas to be matured

A schema of neural input & output states

Today's agenda

Context

Deep Dive into the paper

Outlook

To introduce the paper three questions are addressed

Answers to these question were synthesized using an interdisciplinary literature review

What are representations?

What is representational change?

How do representations change?

To introduce the paper three questions are addressed

What are representations?

What is representational change?

How do representations change?

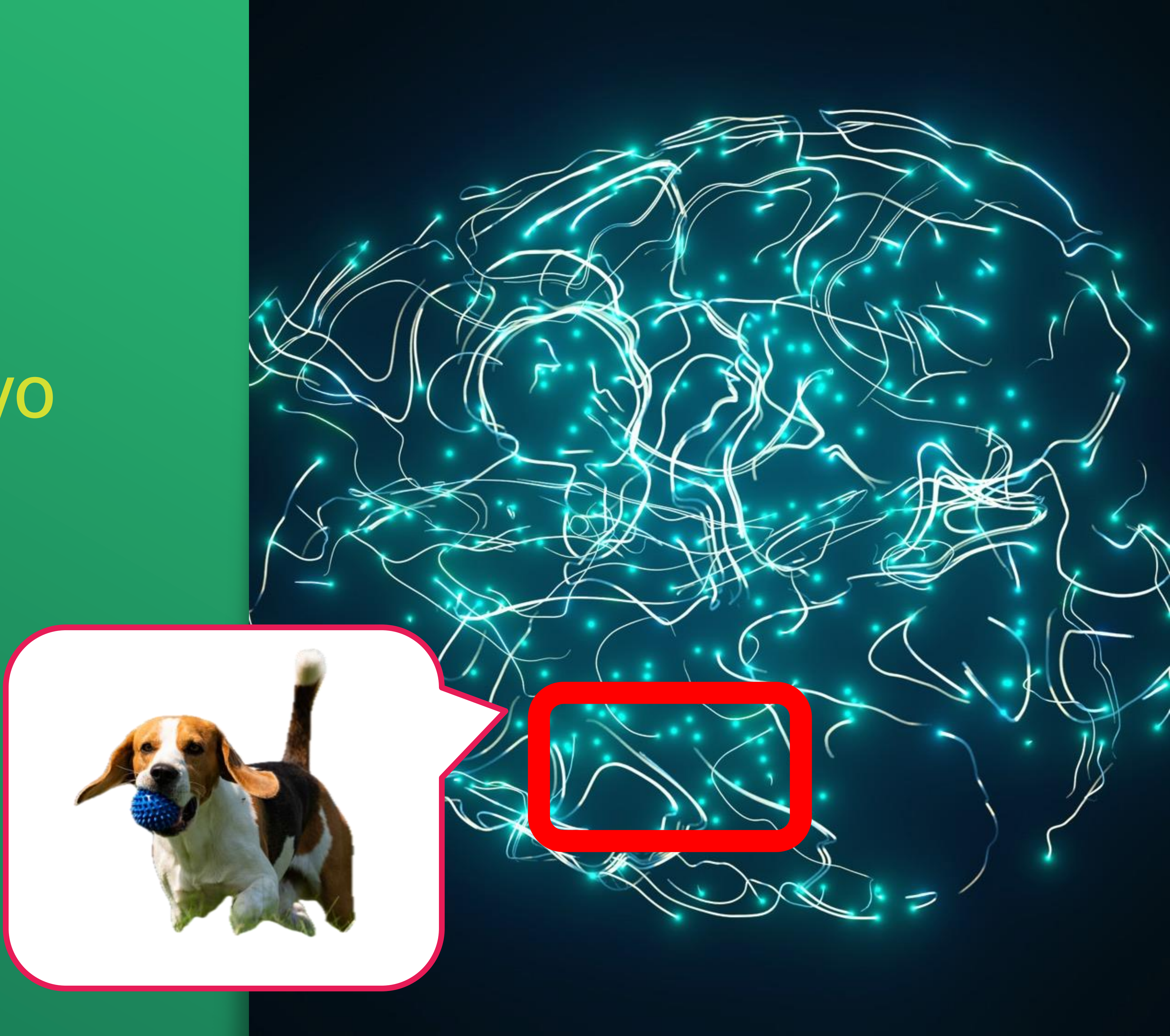
= Representation's lifecycle

The first of two parts of an artificial neural representation

Example of an referential anker: picture of a dog

Source: One interpretation of neuro-representationalism outlined by Hubbard (2008), Anderson and Champion (2022)

The second of two parts of an artificial neural representation



Source: One interpretation of neuro-representationalism outlined by Hubbard (2008), Anderson and Champion (2022)

Set of characteristics of artificial neural representations

Representations are actually much more complicated

Non exhaustive

- Epistemic tools
- Aspect representing
- Time expanded
- Local / distributed
- (Dis-)entangled
- Situated
- Differentiable
- Structured
- (Multi-)Modal
- Subset of information
    - Physically instantiated
    - Similar (and maybe) equivalent

Aspects that can be represented:

- Pictures of dogs
- Objects
- Functional schemas
- Motor actions
- Pictures
- Sounds
- Smells
- Emotions (for humans)
- Numbers
- Humans
- Relationships
- …

To introduce the paper three questions are addressed

What are representations?

What is representational change?

How do representations change?

= *Representation's lifecycle*

What is representational change?

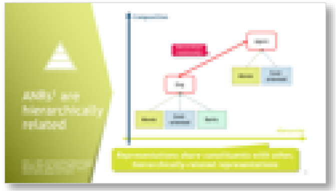
Representational change is …

Presented in the following

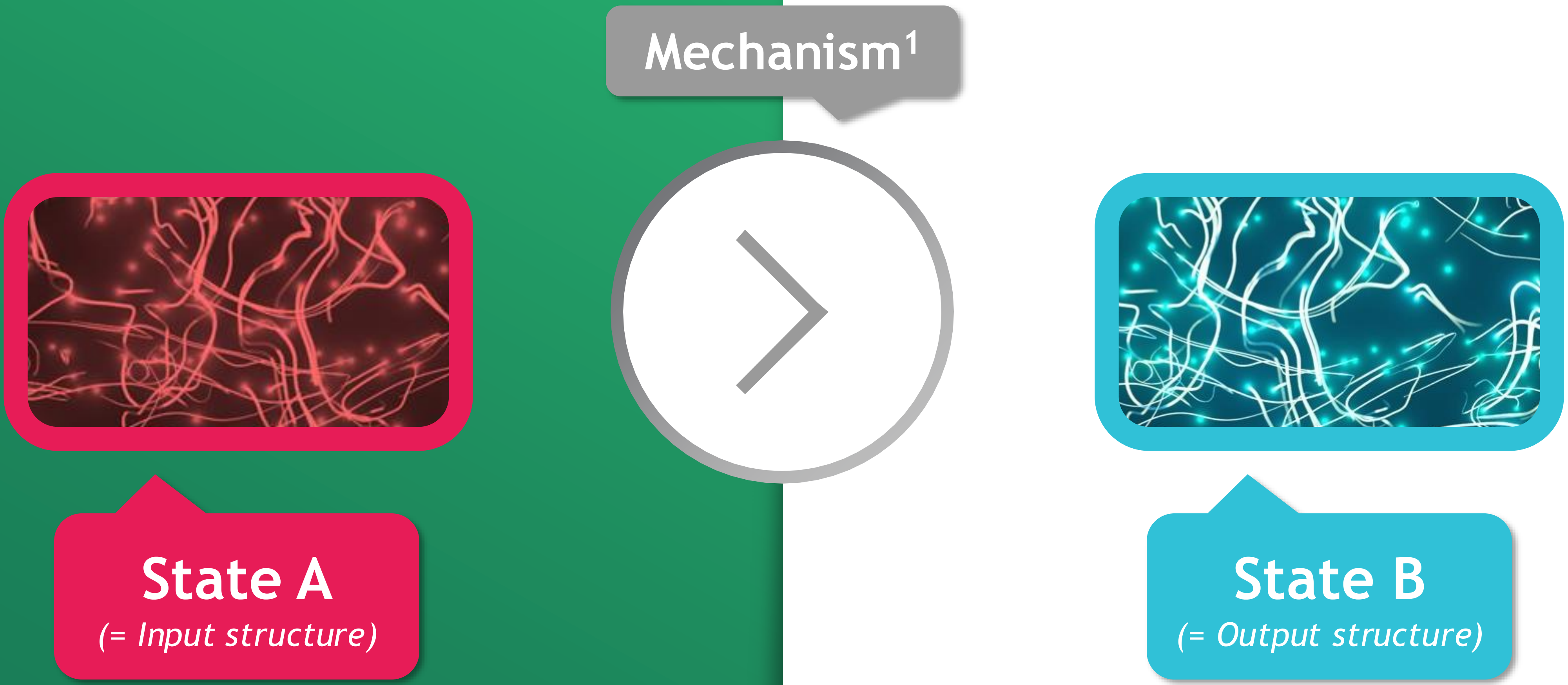
… a “state-to-state” transition …

… describable along a compositional and …

… a hierarchical dimension



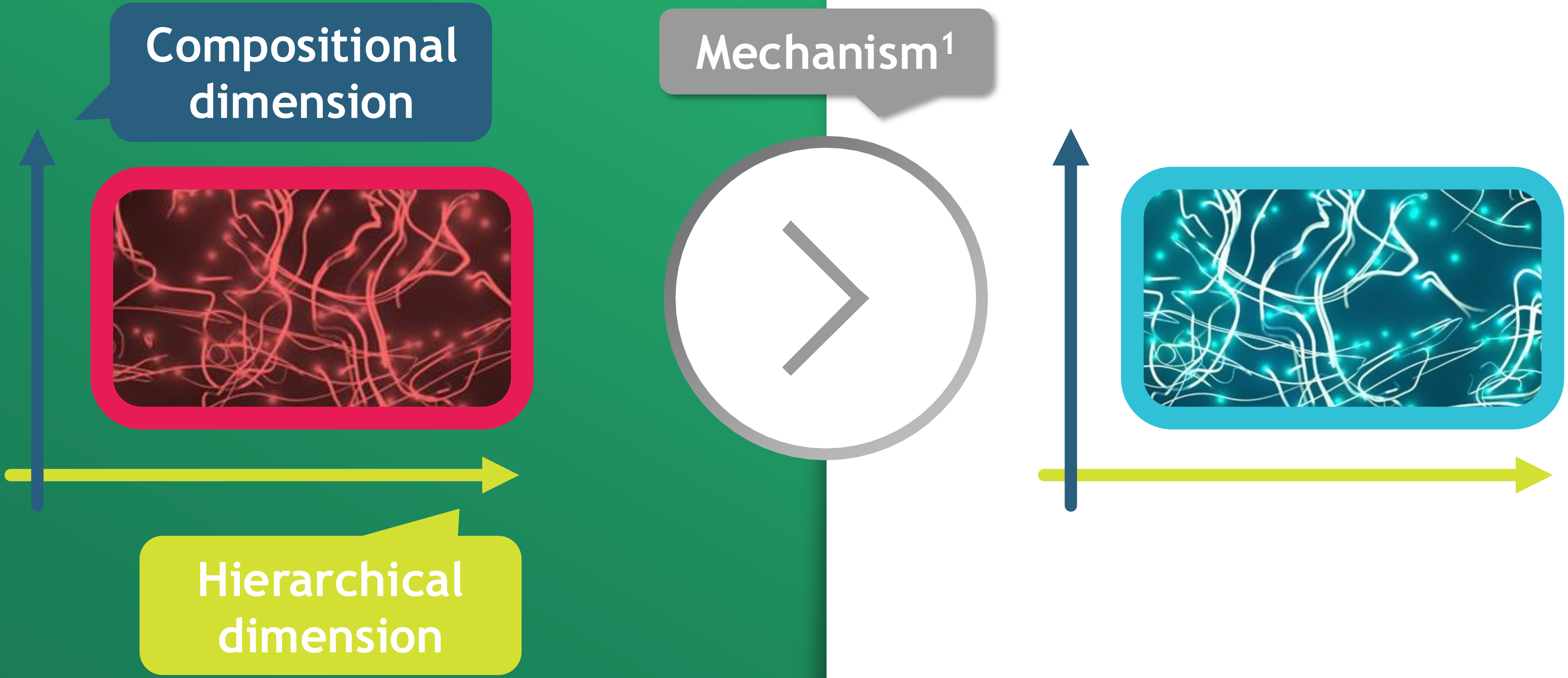
Representational change is a “state-… … to-state transition” including a mechanism



1. Not in scope of the paper

Representational change is a "state-…

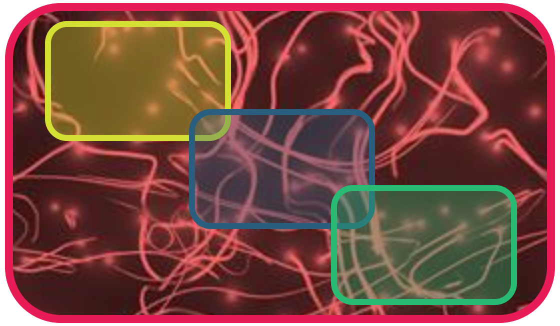
… to-state transition" including a mechanism



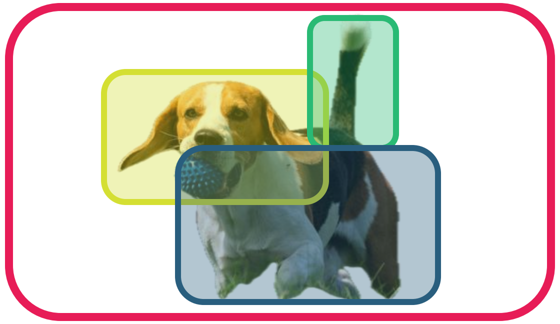
ANRs[1] are compositions

Note 1. ANR = Artificial Neural Representation
Sources: Poldrack, R.A. (2020); Bengio et al (2013); Kästner and Crook, 2023; Nanda et al. (2023)

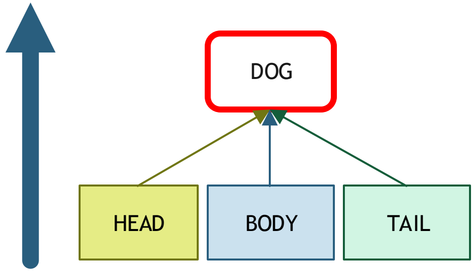
ANRs are **composed of constituents** …



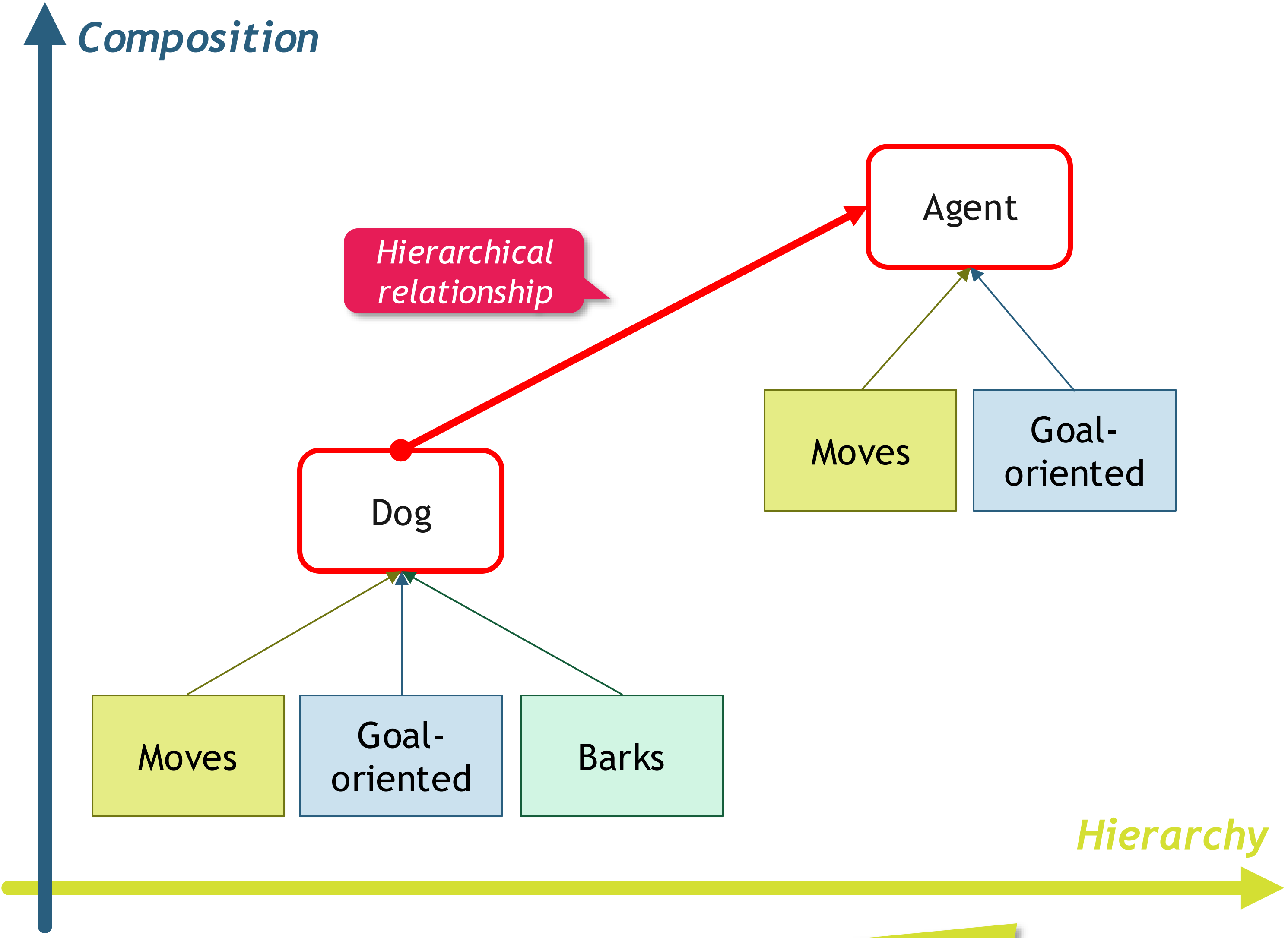
… that **refer to components** of the referential anker …



… and are **illustrated as conceptual graphs** in my paper



ANRs[1] are hierarchically related



Representations share constituents with other, hierarchically-related representations

Note: 1. ANR = Artificial Neural Representation
Sources: Saxe et al. (2019); Thagard, P. (2024); Barsalou et al. (2018); Bengio et al. (2013); LeCun (2015); Bengio et al. (2011)

To introduce the paper three questions are addressed

What are representations?

What is representational change?

How do representations change?

= *Representation's lifecycle*

Representational changes are organized along three phases

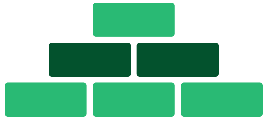
Innate

Description

Refers to a priori[1] integration of so-called primitives (= “innate representations”) into the system’s architecture

Representational changes

- Perceptual primitives’ integration
- Abstract primitives’ integration



Form & Change

Includes formation and change of representations through combination of data and existing representations

- Assembly
- Abstraction
- Differentiation

Deletion

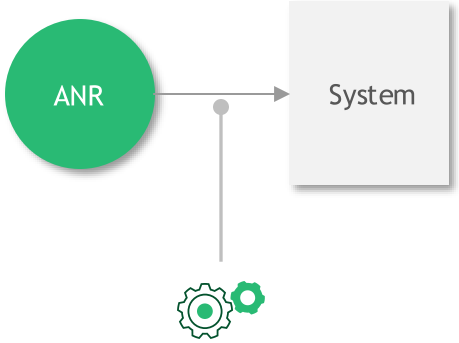
Concerned with the deletion of representations or parts of the representations due to e.g., significantly changing input data

- Deletion

Changes synthesized on the basis of a interdisciplinary literature review using AI research and broader cognitive science (e.g., psychology, philosophy, neuroscience, linguistics etc.)

Representational changes are organized along three phases

Innate



A priori integration of concepts in systems through e.g.,
- Evolution (Humans)
- Developers (AI)

Form & Change

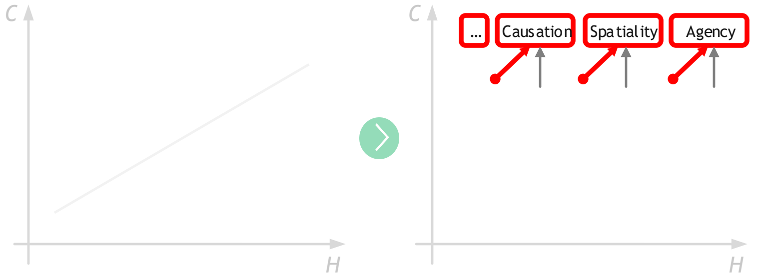
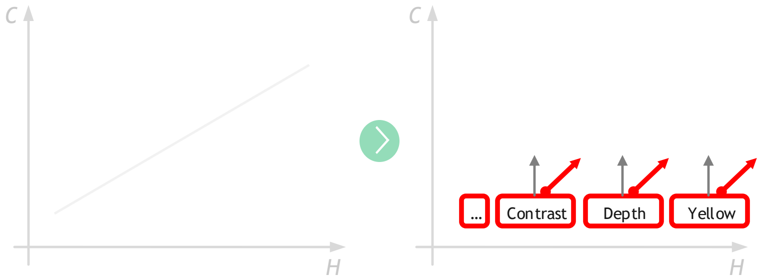
A posteriori formation & change of concepts in systems through e.g.,
1 utilizing perceptual data
2 combining existing concepts

Deletion

A posteriori deletion of concepts in systems through e.g.,
- structured deletion
- variable change

Note: ANR = Artificial Neural Representation

Innate

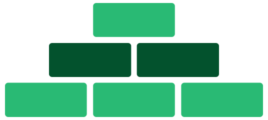
Process	Short description	Illustration of representation-altering process
		Input structure *Output structure**
Abstract Primitives' Integration	In the output structure the representational capacity is **initiated by one or multiple abstract primitive(s)**	
Concrete Primitives' Integration	In the output structure the representational capacity is **initiated by one or multiple concrete primitive(s)**	

Sources (non-exclusive):
Versace, E., et al., Priors in Animal and Artificial Intelligence: Where Does Learning Begin? Trends Cogn Sci, 2018., Marcus, G., Innateness, AlphaZero, and Artificial Intelligence. 2018.
Locke, J., An essay concerning human understanding. Vol. 3. 1689: Oxford University Press. 601-605.
Silver, D., et al., Mastering the game of Go without human knowledge. Nature, 2017. 550(7676): p. 354-359.
Barabasi, D.L., et al., Complex computation from developmental priors. Nat Commun, 2023. 14(1): p. 2226.
Mandler, J.M., On the Birth and Growth of Concepts. Philosophical Psychology, 2008. 21(2): p. 207-230.
Mandler, J.M., The spatial foundations of the conceptual system. Language and Cognition, 2014. 2(1): p. 21-44.
Carey, S., The Origin of Concepts. Journal of Cognition and Development, 2000. 1(1): p. 37-41.

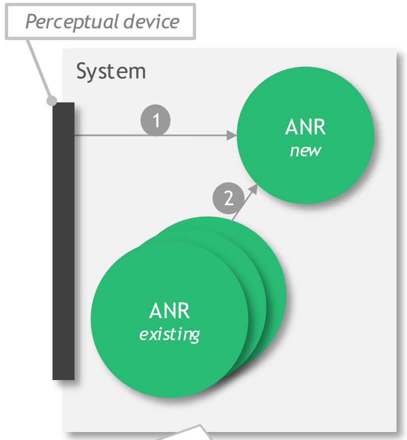
Representational changes are organized along three phases

Innate

A priori integration of concepts in systems through e.g.,
- Evolution (Humans)
- Developers (AI)



Form & Change



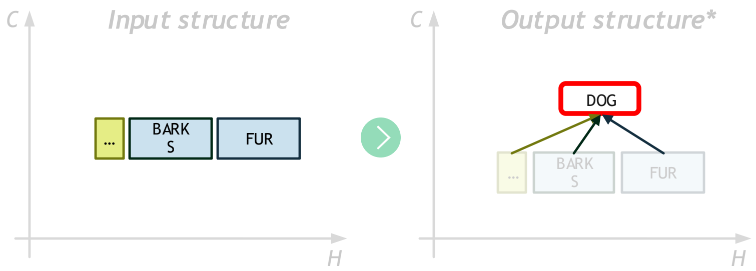
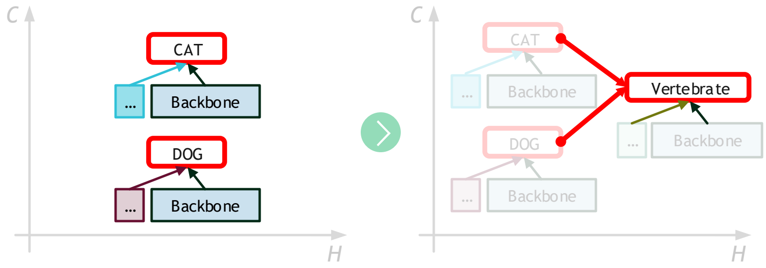
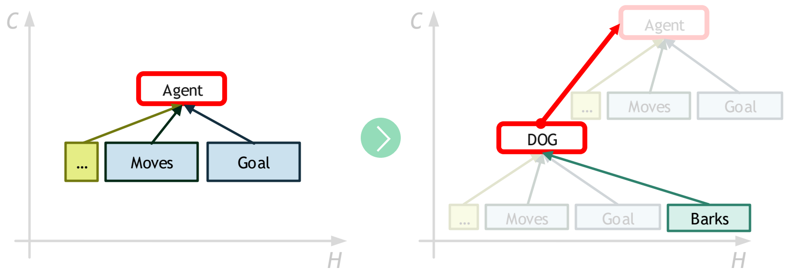
A posteriori formation & change of concepts in systems through e.g.,
1. utilizing perceptual data
2. combining existing concepts

Deletion

A posteriori deletion of concepts in systems through e.g.,
- structured deletion
- variable change

Note: ANR = Artificial Neural Representation

Form & Change

Process	Short description	Illustration of representation-altering process
Assembly	In the output structure the representational capacity is **complemented by an assembled representation** subsuming one or multiple previously unrelated representations of the input structure as constituents	
Abstraction	In the output structure the representational capacity is **complemented by an abstracted representation** subsuming (essential) constituents of related concrete representations of the input structure	
Differentiation	In the output structure the representational capacity is **complemented by a differentiated representation** subsuming constituents of its related abstract parent of the input structure and additional distinguishing constituents	

Sources (non-exhaustive):
Chalmers, D.J., R.M. French, and D.R. Hofstadter, *High-level perception, representation, and analogy: A critique of artificial intelligence methodology*. Journal of Experimental & Theoretical AI, 1992. **4**(3): p. 185-211.
Scholkopf, B., et al., *Toward Causal Representation Learning*. Proceedings of the IEEE, 2021. **109**(5): p. 612-634.
Aslin, R.N. and L.B. Smith, *Perceptual development*. Annu Rev Psychol, 1988. **39**: p. 435-73.
Martin, K.A., *A brief history of the "feature detector"*. Cereb Cortex, 1994. **4**(1): p. 1-7.
(fully view on the following slides)

Sources for the form & change phase

Sources (non-exhaustive):
Chalmers, D.J., R.M. French, and D.R. Hofstadter, *High-level perception, representation, and analogy: A critique of artificial intelligence methodology*. Journal of Experimental & Theoretical AI, 1992. **4**(3): p. 185-211.
Scholkopf, B., et al., *Toward Causal Representation Learning*. Proceedings of the IEEE, 2021. **109**(5): p. 612-634.
Aslin, R.N. and L.B. Smith, *Perceptual development*. Annu Rev Psychol, 1988. **39**: p. 435-73.
Martin, K.A., *A brief history of the "feature detector"*. Cereb Cortex, 1994. **4**(1): p. 1-7.
Eimas, P.D. and J.D. Corbit, *Selective Adaptation of Linguistic Feature Detectors*. Cognitive Psychology, 1973. **4**: p. 99-109.
Pelli, D.G., et al., *Feature detection and letter identification*. Vision Res, 2006. **46**(28): p. 4646-74.
Li, Y., et al., *A survey of recent advances in visual feature detection*. Neurocomputing, 2015. **149**: p. 736-751.
Yu, L. and H. Liu, *Efficient Feature Selection via Analysis of Relevance and Redundancy*. Journal of Machine Learning Research, 2004. **5**: p. 1205–1224.
Higgins, I., et al., *SCAN: Learning hierachical compositional visual concepts*, in *International Conference on Learning Representations*. 2018: Vancouver, Canada.
Wasserman, E.A. and R.R. Miller, *What's elementary about associative learning?* Annu. Rev. Psychology, 1997. **48**: p. 573-607.
Solomon, K., D. Medin, and E. Lynch, *Concepts do more than categorize*. Trends Cogn Sci, 1999. **3**(3): p. 99-105.
Welling, H., *Four Mental Operations in Creative Cognition: The Importance of Abstraction*. Creativity Research Journal - CREATIVITY RES J, 2007. **19**.
Gibson, J. and E. Gibson, *Perceptual learning: differentiation or enrichment?* Psychological Review, 1955. **62**(1): p. 32-41.
Caviezel, M.P., et al., *The Neural Mechanisms of Associative Memory Revisited: fMRI Evidence from Implicit Contingency Learning*. Front Psychiatry, 2019. **10**: p. 1002.
Kiefer, M. and L.W. Barsalou, *Grounding the Human Conceptual System in Perception, Action, and Internal States*, in *Action Science*. 2013. p. 381-407.
Barsalou, L.W., *Grounded cognition*. Annu Rev Psychol, 2008. **59**: p. 617-45.
Mitchell, C.J., J. De Houwer, and P.F. Lovibond, *The propositional nature of human associative learning*. Behav Brain Sci, 2009. **32**(2): p. 183-98; discussion 198-246.
Asmuth, J. and D. Gentner, *Relational categories are more mutable than entity categories*. Q J Exp Psychol (Hove), 2017. **70**(10): p. 2007-2025.
Ullman, S., et al., *Atoms of recognition in human and computer vision*. Proc Natl Acad Sci U S A, 2016. **113**(10): p. 2744-9.
Voulodimos, A., et al., *Deep Learning for Computer Vision: A Brief Review*. Comput Intell Neurosci, 2018. **2018**: p. 7068349.
Li, D., et al., *Visual Feature Learning on Video Object and Human Action Detection: A Systematic Review*. Micromachines (Basel), 2021. **13**(1). Gentner, D. and C. Hoyos, Analogy and Abstraction. Top Cogn Sci, 2017
Frankland, S.M. and J.D. Greene, Concepts and Compositionality: In Search of the Brain's Language of Thought. Annu Rev Psychol, 2020. 71: p. 273-303.
Burgoon, E.M., M.D. Henderson, and A.B. Markman, There Are Many Ways to See the Forest for the Trees: A Tour Guide for Abstraction. Perspect Psychol Sci, 2013. 8(5): p. 501-20.
Borghi, A.M., et al., The challenge of abstract concepts. Psychol Bull, 2017. 143(3): p. 263-292.
Buckner, C., Deep learning: A philosophical introduction. Philosophy Compass, 2019. 14(10).
Voudouris, K., et al., Direct Human-AI Comparison in the Animal-AI Environment. Front Psychol, 2022. 13: p. 711821.
Smith, C., S. Carey, and M. Wiser, On differentiation: A case study of the development of the concepts of size, weight and density. Cognition, 1985. 21: p. 177-237.
Goldstone, R.L., Perceptual learning. Annu. Rev. Psychol., 1998. 49: p. 585-612.

Representational changes are organized along three phases

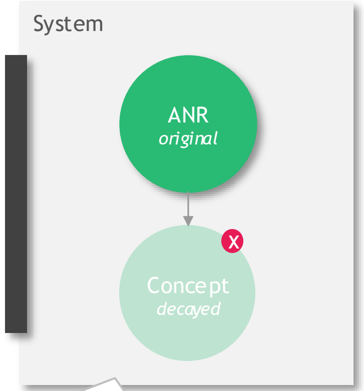
Innate

A **a priori integration** of concepts in systems through e.g.,
- Evolution (Humans)
- Developers (AI)

Form & Change

A **posteriori formation & change** of concepts in systems through e.g.,
1. utilizing perceptual data
2. combining existing concepts

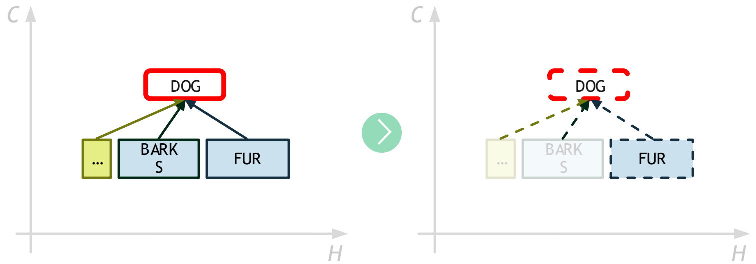
Deletion



A **posteriori deletion** of concepts in systems through e.g.,
- structured deletion
- variable change

Note: ANR = Artificial Neural Representation

Decay

Process	Short description	Illustration of representation-altering process
Deletion	In the output structure the representational capacity has **removed a representation and / or a constituent**	*Input structure* / *Output structure** 

Sources (not exhaustive):
Bjotk, E., R.A. Bjork, and M. Anderson, Varieties of goal directed forgetting, in Intentional forgetting: Interdisciplinary approaches, J.M. Golding and C.M. MacLeod, Editors. 1998
Williams, M., et al., The benefit of forgetting. Psychon Bull Rev, 2013. 20(2): p. 348-55.
Timm, I.J., et al., Intentional Forgetting in Artificial Intelligence Systems: Perspectives and Challenges, in KI 2018: Advances in Artificial Intelligence. 2018. p. 357-365.
Ellwart, T. and A. Kluge, Psychological Perspectives on Intentional Forgetting: An Overview of Concepts and Literature. KI - Künstliche Intelligenz, 2018. 33(1): p. 79-84.
Markovitch, S. and P.D. Scott, The role of forgetting in learning, in Proceedings of the fifth international conference on machine learning. 1998: Ann Arbor, Michigan.
Jung, H., et al., Less-forgetful learning for domain expansion in deep neural networks, in Thirty-Second AAAI Conference on Artificial Intelligence. 2018.
Kirkpatrick, J., et al., Overcoming catastrophic forgetting in neural networks. Proc Natl Acad Sci U S A, 2017. 114(13): p. 3521-3526.
Ebrahimi, S., et al., Remembering for the right reasons: Explanations reduce catastrophic forgetting. Applied AI Letters, 2021. 2(4).

Today's agenda

- Context
- Deep Dive into the paper
- Outlook

Not exhaustive

Three things could be addressed to enrich the framework

Better understand similarity of representations in ANNs

Formalizing the conceptual framework

Exploring mechanisms

Thank you