GAC 2021 Proposal

Title: What makes representations “useful”?

Scientific question: Internal representations play a central role in the study of both biological and artificial intelligence, as well as in philosophy of mind, but what precisely defines a representation is challenging to pin down. Across disciplines, one common thread is that representations are typically “useful” in some sense. Centering around this concept of usefulness, we propose a cross-disciplinary GAC to share ideas and develop more precise answers to the following questions:

1. What makes representations “useful,” both in terms of their *content* and their *form*?
2. How does the *use* or downstream causal effect of a representation contribute to its meaning?

We will simplify the scope by primarily discussing these questions in the context of visual perception.

Background: Representations play a central role in cognitive science, neuroscience, philosophy of mind, and machine learning. Painting in broad strokes, “representation” in philosophy often goes hand-in-hand with intentionality, or the idea that minds manipulate *meaningful* symbols, and meaningful representations play an explanatory role in phenomenology. In cognitive science and artificial intelligence, the idea that an agent internalizes representations of the world – in terms of entities (like objects) and structures (like maps) – helps to both explain and design intelligent systems. In machine learning, designing or learning the right representations can significantly impact the efficiency of downstream tasks including both further learning and inference. While some pluralism of ideas is to be expected, there is also much to gain by exploring connections between these different aspects of representation across disciplines.

Researchers in each of these fields are often inspired by ideas or developments in the others, and motivated by some of the same basic questions, such as those surrounding representations. Disagreement and misunderstanding about the term “representation” has the potential to hinder scientific communication and duplicate research efforts [6, 11, 59, 31, 25]. If the main source of confusion is a plurality of serviceable notions of representations, our GAC will help to clarify what these notions are and in which contexts they are appropriate. If, indeed, a deeper confusion about representations stymies innovative research, then our GAC presents an opportunity to clarify and refocus research in multiple fields, and to cultivate cross-disciplinary researchers.

Ideas about “useful” representations are seeing a resurgence in different forms, making it a timely and exciting subtopic for collaboration across disciplines. We will next summarize some active areas related to “usefulness” in each discipline.

An active area in **philosophy** examines *embodied* cognition and how cognitive and perceptual processes might involve *affordances*, or possibilities for action in the perceiver’s environment. It remains unresolved what the implications of embodiment and affordances are for our understanding of representations [47, 54, 27, 3, 50, 52, 49, 19]. Further, there is a (re)animated literature around how *teleology* – roughly, functions or purposes derived from the process of evolution – should inform our understanding of representations and mechanisms in neural systems [15, 4, 53, 26, 14, 41].

In **neuroscience**, “usefulness” appears in many guises. For one, many have pointed out the insufficiency of correlational methods for making representational claims; equally important is how neural activity influences downstream functions, or put another way, how the hypothesized representation is actually *used* [46, 45, 11, 6]. A related problem is leveraging animals’ natural behaviors and ecological niches as a way to constrain theories of neural representation [20, 18, 42]. Other senses of “usefulness” derive from optimization for a task or set of tasks [39, 61, 32, 58], or task-independent prediction of past or future inputs [43, 44, 16, 55]; learning an internal model of the world may be a possible middle-ground [34, 9].

There have been many exciting recent developments in **machine learning** on topics related to representation-learning from unlabeled and minimally structured data. As a result, there is a growing and evolving set of quantitative metrics for what makes a model or representation useful. One area is elaborating classic concepts from Information Theory (IT) to explain the success of deep learning [57, 1, 23], and to address technical limitations that have made IT impractical for studying representations [60]. There have also been recent developments in defining "disentangled" representations [8, 29, 48, 24], along with new design patterns and training objectives to achieve them by unsupervised methods [28, 13, 2, 35]. New ideas to self-supervised learning, learning from structured data like video, and active/causal perception has led to further breakthroughs in representation learning [17, 36, 21, 56, 40, 51]. Another active area in deep learning is uncertainty quantification, where there is renewed interest in questions like *what* types of uncertainty to represent and *why*, and various proposals for how to do so in practice [30, 10, 33, 12][1].

The dominant tools for quantitatively studying representations in machine learning and neuroscience incorporate varying senses of *usefulness* and *meaningfulness*, which are crucial in philosophical accounts of representation. For example, Information Theory and Mutual Information have been extensively applied throughout neuro- and cognitive science [5, 7, 22] as well as in deep learning [30, 57, 1]. However, it has been pointed out that this kind of information is not necessarily "usable" in principle [60, 23] or "used" in practice [45, 6]. Another popular suite of tools compares representations based on their geometry [38, 37], which identifies "useful" representations insofar as the same information is relevant for two systems (e.g. brains and neural networks). Further, these methods are sensitive to statistical dependencies between representational spaces, but not to their downstream causal effects.

Challenges, competing hypotheses, and proposed approach for resolution: The primary challenge will be identifying specific areas of overlap across fields, given their diverse ideas and vocabularies. We will therefore structure the GAC in two parts: the first will be a brief set of tutorials in which representatives from philosophy, machine learning, and neuroscience will each give an introduction to concepts, seminal studies, operational definitions, and open questions about "useful" representations in their respective fields. In the second part of the GAC, we will hold moderated discussions and debates on specific topics structured around the high-level "scientific questions" outlined at the beginning of this document. These high-level questions are paraphrased in bold below, alongside possible fine-grained discussion topics. This format and the questions below are subject to changes given feedback from the community.

1. **What makes representations "useful" in terms of their *content* and their *form*, and how are these related?**

   What to represent (content): To what extent is it task performance or reward all the way down versus task-free model-building all the way up, and how do these interact? How do embodiment, affordances, and ecology further shape what is useful to represent? In what ways is it useful to represent uncertainty? In what ways is it useful to represent causal relations?

   How to represent it (form): What are the principles of designing usable representations – things like disentanglement, independence, invariance, equivariance, efficiency, smoothness,

[1] Representing uncertainty is also a contentious topic in neuroscience and cognitive science, with two of last year's GACs devoted to questions on representation of probability in the brain.

and decodability – and how do they relate to each other? How do these interact with representational content, if at all? What existing evidence is there for each of these properties in the brain, or what new experiments are needed to answer this?

2. **The role of actual and potential *use*.** Are different questions answered by knowing the potential uses of a representation versus knowing the actual downstream effect that a particular representation has? What are the conceptual, experimental, or technical barriers to quantifying causal interactions among internal representations, or between internal representations and behavior? How do salient actions and threats for an organism broadly influence the representations it uses? How, if at all, do evolutionary histories constrain biological representations, and does something play an analogous role for artificial ones?

Concrete outcomes:

- A taxonomy of useful forms of representations across disciplines, and how they relate.
- A set of empirically testable questions – and experimental methodologies – about representational forms in biological neural systems.
- Updated philosophical theories of representation, and mathematical tools for quantifying representations and representational similarity, taking into account emerging ideas in ML on disentanglement, causality, etc.

Benefit to the community:

- Our collaboration will allow for deeper and broader insight into a far-reaching set of questions than an investigation from within any one discipline is likely to achieve. It will also serve to cross-pollinate ideas between disciplines, and to facilitate future inter-disciplinary collaborations.
- We will strongly encourage didacticism and open-access materials for all participants so that tutorials and discussions will serve as an ongoing teaching resource.
- We will draw from diverse backgrounds and career levels for the participants, creating an impactful and rare opportunity for researchers that face unique challenges, and serving as a model for inclusive science.

Core members: The final team is subject to change, and we welcome new ideas and new members from the community. If this proposal is accepted, the following members have all initially agreed to help organize or advise the GAC, to co-author a summary paper, and to share updates at CCN 2022.

Name	Position/Institution	Expertise	Role
Ben Baker	Postdoc/UPenn	Philo	Organizer
Richard Lange	Postdoc/UPenn	Neuro (ML)	Organizer
Alessandro Achille	Applied Scientist/Amazon	ML	Organizer
Rosa Cao	Professor/Stanford	Philo (Neuro)	Organizer
Niko Kriegeskorte	Professor/Columbia	Neuro (Philo)	Advisor
Odelia Schwartz	Professor/Miami	Neuro	Advisor
Xaq Pitkow	Professor/BCM/Rice	Neuro (ML)	Advisor

Signed: Ben Baker, Richard Lange, Alessandro Achille[2], Rosa Cao, Nikolaus Kriegeskorte, Odelia Schwartz, Xaq Pitkow

[2] Paper co-authorship pending approval from Amazon.

References:

[1] Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Information Theory and Applications Workshop*, 19, 2018.

[2] Alessandro Achille, Tom Eccles, Loic Matthey, Christopher P. Burgess, Nick Watters, Alexander Lerchner, and Irina Higgins. Life-long disentangled representation learning with cross-domain latent homologies. *Advances in Neural Information Processing Systems*, pages 9873–9883, 2018.

[3] Ken Aizawa. Cognition and behavior. *Synthese*, 194(11):4269–4288, 2017.

[4] Marc Artiga. Teleosemantic modeling of cognitive representations. *Biology and Philosophy*, 31 (4):483–505, 2016.

[5] Fred Attneave. Some informational aspects of visual perception. *Psychological Review*, 61(3): 183–193, 1954.

[6] Ben Baker, Benjamin Lansdell, and Konrad Kording. A Philosophical Understanding of Representation for Neuroscience. *arXiv*, 2021.

[7] H. Barlow. Redundancy reduction revisited. *Network: Computation in Neural Systems*, 12(3): 241–253, 2001.

[8] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[9] Pietro Berkes, Gergo Orbán, Máté Lengyel, and József Fiser. Spontaneous Cortical Activity Reveals Hallmarks of an Optimal Internal Model of the Environment. *Science*, 331(January): 83–87, 2011.

[10] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. *ICML*, 2015.

[11] Romain Brette. Is coding a relevant metaphor for the brain? *Behavioral and Brain Sciences*, 42 (e215), 2019.

[12] Nicolas Brosse, Carlos Riquelme, Alice Martin, Sylvain Gelly, and Éric Moulines. On last-layer algorithms for classification: Decoupling representation from uncertainty estimation. *arXiv*, 2020.

[13] Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE. *Advances in Neural Information Processing Systems*, 2018.

[14] Patrick Butlin. Representation and the active consumer. *Synthese*, 197(10):4533–4550, 2020.

[15] Rosa Cao. A teleosemantic approach to information in the brain. *Biology and Philosophy*, 27(1): 49–71, 2012.

[16] Matthew Chalk, Olivier Marre, and Gašper Tkačik. Toward a unified theory of efficient, predictive, and sparse coding. *Proceedings of the National Academy of Sciences*, 115(1):186–191, 2018.

[17] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *ICML*, 37, 2020.

[18] Paul Cisek. Resynthesizing behavior through phylogenetic refinement. *Attention, Perception, and Psychophysics*, 81(7):2265–2287, 2019.

[19] Axel Constant, Andy Clark, and Karl J. Friston. Representation wars: Enacting an armistice through active inference. *Frontiers in Psychology*, 11, 2021.

[20] Sandeep Robert Datta, David J. Anderson, Kristin Branson, Pietro Perona, and Andrew Leifer. Computational Neuroethology: A Call to Action. *Neuron*, 104:11–24, 2019.

[21] Emily Denton and Vighnesh Birodkar. Unsupervised learning of disentangled representations from video. *Advances in Neural Information Processing Systems*, pages 4415–4424, 2017.

[22] Alexander G. Dimitrov, Aurel A. Lazar, and Jonathan D. Victor. Information theory in neuroscience. *Journal of Computational Neuroscience*, 30(1), 2011.

[23] Yann Dubois, Douwe Kiela, Ramakrishna Vedantam, and David J. Schwab. Learning optimal representations with the decodable information bottleneck. *arXiv*, 2020.

[24] Cian Eastwood and Christopher K.I. Williams. A framework for the quantitative evaluation of disentangled representations. *ICLR*, 2018.

[25] Frances Egan. How to think about mental content. *Philosophical Studies*, 170(1):115–135, 2014.

[26] Justin Garson and David Papineau. Teleosemantics, selection and novel contents. *Biology and Philosophy*, 34(3):36, 2019.

[27] Antje Gentsch, Arne Weber, Matthis Synofzik, Gottfried Vosgerau, and Simone Schütz-Bosbach. Towards a common framework of grounded action cognition: Relating motor control, perception and cognition. *Cognition*, 146:81–89, 2016.

[28] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. *ICLR*, 2017.

[29] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a Definition of Disentangled Representations. *arXiv*, 2018.

[30] Geoffrey E. Hinton and Drew van Camp. Keeping Neural Networks Simple by Minimizing the Description Length of the Weights. *ACM COLT*, 1993.

[31] Ilenna Simone Jones and Konrad Paul Kording. Quantifying the role of neurons for behavior is a mediation question. *Behavioral and Brain Sciences*, 42(e233), 2019.

[32] Alexander JE Kell and Josh H. McDermott. Deep neural network models of sensory systems: windows onto the role of task constraints. *Current Opinion in Neurobiology*, 55:121–132, 2019.

[33] Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 2017-Decem(Nips):5575–5585, 2017.

[34] Daniel Kersten, Pascal Mamassian, and Alan Yuille. Object perception as bayesian inference. *Annual Review of Psychology*, pages 271–304, 2004.

[35] Hyunjik Kim and Andriy Mnih. Disentanging by Factorising. *ICML*, 35, 2018.

[36] David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards Nonlinear Disentanglement in Natural Data with Temporal Sparse Coding. *arXiv*, 2020.

[37] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of Neural Network Representations Revisited. *ICML*, 36, 2019.

[38] Nikolaus Kriegeskorte. Relating population-code representations between man, monkey, and computational models. *Frontiers in Neuroscience*, 3(3):363–373, 2009.

[39] Nikolaus Kriegeskorte. Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, 1:417–446, 2015.

[40] Francesco Locatello, Stefan Bauer, Mario Lucic, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. *arXiv*, 2019.

[41] Ruth Garrett Millikan. Neuroscience and teleosemantics. *Synthese*, forthcoming.

[42] Simon Musall, Anne E. Urai, David Sussillo, and Anne K. Churchland. Harnessing behavioral diversity to understand neural computations for cognition. *Current Opinion in Neurobiology*, 58: 229–238, 2019.

[43] Bruno a Olshausen and David J. Field. Sparse coding with an incomplete basis set: a strategy employed by V1?, 1997.

[44] Stephanie E. Palmer, Olivier Marre, Michael J. Berry, and William Bialek. Predictive information in a sensory population. *Proceedings of the National Academy of Sciences*, 112(22):6908–6913, 2015.

[45] Stefano Panzeri, Christorpher D. Harvey, Eugenio Piasini, Peter E. Latham, and Tommaso Fellin. Cracking the neural code for sensory perception by combining statistics, intervention and behavior. *Neuron Perspective*, 93(3):491–507, 2017.

[46] a J Parker and William T. Newsome. Sense and the single neuron: probing the physiology of perception. *Annual review of neuroscience*, 21:227–277, 1998.

[47] Simon Prosser. Affordances and phenomenal character in spatial perception. *Philosophical Review*, 120(4):475–513, 2011.

[48] Karl Ridgeway and Michael C. Mozer. Learning deep disentangled embeddings with the F-statistic loss. *Advances in Neural Information Processing Systems*, pages 185–194, 2018.

[49] Sonia F. Roberts, Daniel E. Koditschek, and Lisa J. Miracchi. Examples of gibsonian affordances in legged robotics research using an empirical, generative framework. *Frontiers in Neurorobotics*, 14:12, 2020.

[50] Markus Schlosser. Embodied cognition and temporally extended agency. *Synthese*, 195(5): 2089–2112, 2018.

[51] Bernhard Scholkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward Causal Representation Learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[52] Lawrence Shapiro. *Embodied Cognition, 2nd ed*. Routledge, 2019.

[53] Nicholas Shea. *Representation in Cognitive Science*. Oxford University Press, 2018.

[54] Susanna Siegel. Affordances and the contents of perception. In Berit Brogaard, editor, *Does Perception Have Content?*, pages 39–76. Oxford University Press, 2014.

[55] Yosef Singer, Yayoi Teramoto, Ben D.B. Willmore, Andrew J. King, Jan W.H. Schnupp, and Nicol S. Harper. Sensory cortex is optimised for prediction of future input. *eLife*, 7, 2018.

[56] Valentin Thomas, Emmanuel Bengio, William Fedus, Jules Pondard, Philippe Beaudoin, Hugo Larochelle, Joelle Pineau, Doina Precup, and Yoshua Bengio. Disentangling the independently controllable factors of variation by interacting with the world. *arXiv*, 2018.

[57] Naftali Tishby and N Zaslavsky. Deep Learning and the Information Bottleneck Principle. *arXiv*, 2015.

[58] Maxwell H. Turner, Luis Gonzalo Sanchez Giraldo, Odelia Schwartz, and Fred Rieke. Stimulus- and goal-oriented frameworks for understanding natural vision. *Nature Neuroscience*, 22:15–24, 2019.

[59] Daniel Williams. Predictive processing and the representation wars. *Minds and Machines*, 28 (1):141–172, 2018.

[60] Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A Theory of Usable Information Under Computational Constraints. *ICLR*, 2020.

[61] Daniel L K Yamins and James J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.