Explainable AI for higher cognitive functions

Rena Bayramova*, Ole Goltermann*, Lioba Enk, Max Hinrichs, Fabian Kamp, Bianca Serio & Simon M. Hofmann

ESCOP Lille

31/08/2022

MAX PLANCK INSTITUTE
FOR HUMAN COGNITIVE AND BRAIN SCIENCES
UKE
HAMBURG
MAX PLANCK SCHOOL of cognition

Outline

1. From a black box to an explanation
2. Four strategies how to get there
3. Conclusions, questions and limitations

Rise of deep learning in cognitive neuroscience

Toward an Integration of Deep Learning and Neuroscience

Adam H. Marblestone[1], Greg Wayne[2] and Konrad P. Kording[3]*

[1] Synthetic Neurobiology Group, Massachusetts Institute of Technology, Media Lab, Cambridge, MA, USA, [2] Google Deepmind, London, UK, [3] Rehabilitation Institute of Chicago, Northwestern University, Chicago, IL, USA

Building machines that learn and think like people

Brenden M. Lake
Department of Psychology and Center for Data Science, New York University, New York, NY 10011
brenden@nyu.edu
http://cims.nyu.edu/~brenden/

Joshua B. Tenenbaum
Department of Brain and Cognitive Sciences and The Center for Brains, Minds and Machines, Massachusetts Institute of Technology, Cambridge, MA 02139
jbt@mit.edu
http://web.mit.edu/cocosci/josh.html

Tomer D. Ullman
Department of Brain and Cognitive Sciences and The Center for Brains, Minds and Machines, Massachusetts Institute of Technology, Cambridge, MA 02139
tomeru@mit.edu
http://www.mit.edu/~tomeru/

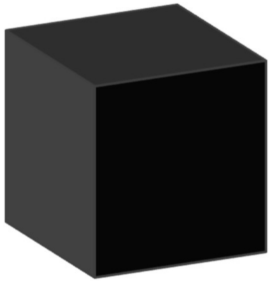
Samuel J. Gershman
Department of Psychology and Center for Brain Science, Harvard University, Cambridge, MA 02138, and The Center for Brains, Minds and Machines, Massachusetts Institute of Technology, Cambridge, MA 02139
gershman@fas.harvard.edu
http://gershmanlab.webfactional.com/index.html

Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing

Nikolaus Kriegeskorte

Medical Research Council Cognition and Brain Sciences Unit, University of Cambridge, Cambridge CB2 7EF, United Kingdom; email: nikolaus.kriegeskorte@mrc-cbu.cam.ac.uk

Black box problem



→ **Explainable Artificial Intelligence (XAI)**

What is an explanation?

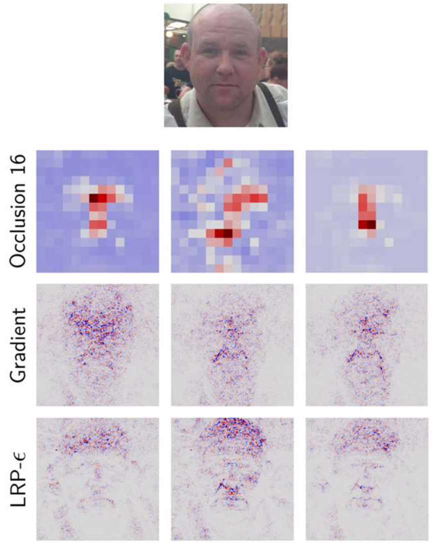
… is any **information** that is **helpful** for the user to understand the **mechanism** behind the described system, by showing what **caused** the system to make decisions it made given a certain input.

Why do we care?

1. Describe
2. **Explain**
3. Predict
4. Change

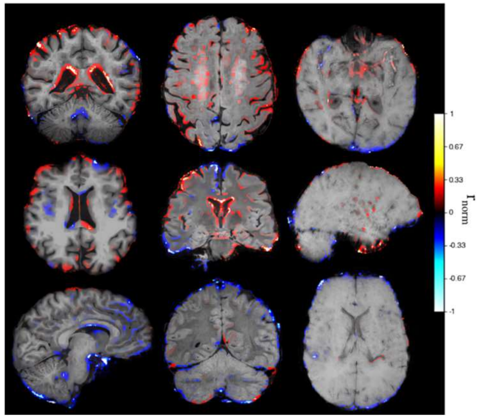
Strategy 1 – Post hoc explanation methods

Image classification



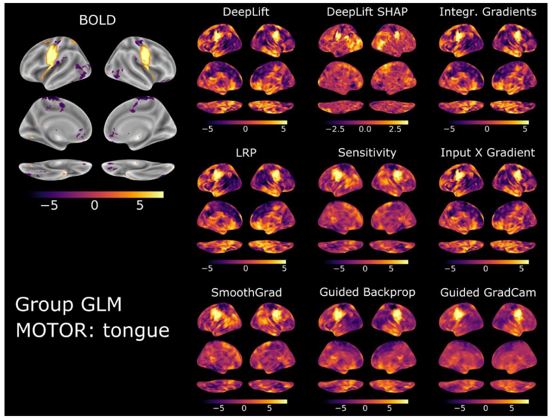
Samek et al. 2021, *IEEE*

Neuroimaging



Hofmann et al. 2022, *Neuroimage*
Goltermann & Hofmann et al. 2022, *OHBM*

Mental state decoding



Thomas et al. 2022, *arxiv preprint*

Strategy 2 – Being cognitive psychologists

Cognitive Psychology for Deep Neural Networks: A Shape Bias Case Study

Samuel Ritter [*1] David G.T. Barrett [*1] Adam Santoro [1] Matt M. Botvinick [1]

Using cognitive psychology to understand GPT-3

Marcel Binz[1,*] and Eric Schulz[1]

[1]MPRG Computational Principles of Intelligence, Max Planck Institute for Biological Cybernetics, 72076 Tübingen, Germany
[*]marcel.binz@tue.mpg.de

PERSPECTIVE
https://doi.org/10.1038/s42256-019-0038-z
nature machine intelligence

Lessons for artificial intelligence from the study of natural stupidity

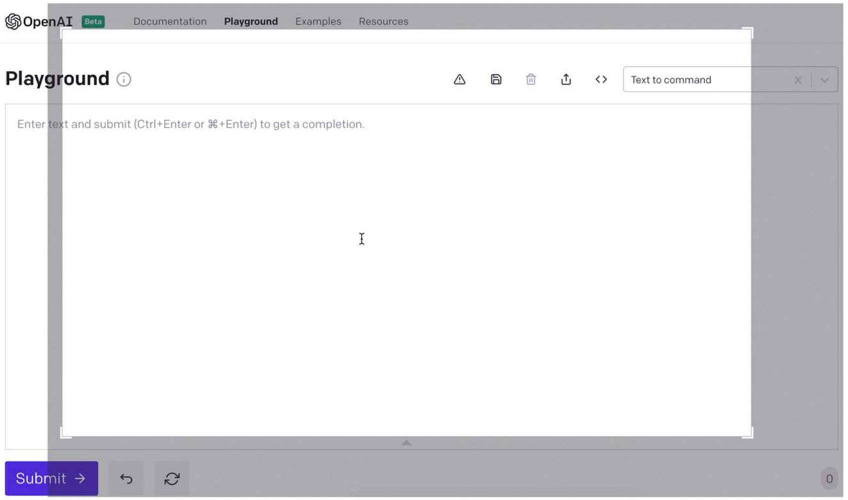
Alexander S. Rich[1,2*] and Todd M. Gureckis[1]

ARTICLES
https://doi.org/10.1038/s42256-022-00458-8
nature machine intelligence
Check for updates

Large pre-trained language models contain human-like biases of what is right and wrong to do

Patrick Schramowski[1], Cigdem Turan[1,2], Nico Andersen[3], Constantin A. Rothkopf[2,4,5] and Kristian Kersting[1,2,5]

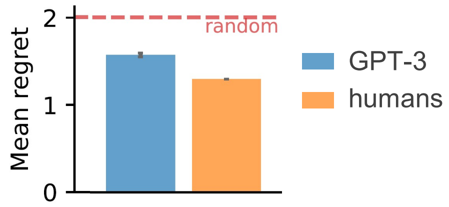
Strategy 2 – Being cognitive psychologists

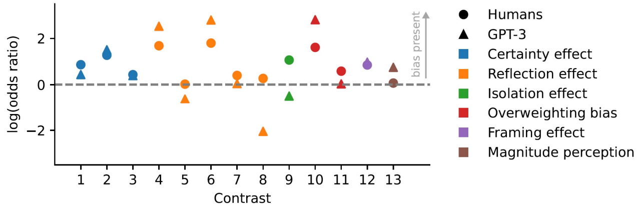


Binz & Schulz, 2022, arxiv *preprint*

Q: Which option do you prefer?

- Option F: 69.0 dollars with 1.0% chance, 26.0 dollars with 99.0% chance.
- Option J: 2.0 dollars with 75.0% chance, 94.0 dollars with 25.0% chance.





MAX PLANCK INSTITUTE FOR HUMAN COGNITIVE AND BRAIN SCIENCES
UKE HAMBURG
MAX PLANCK SCHOOL of cognition

Further approaches

PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Models that learn how humans learn: The case of decision-making and its disorders

Amir Dezfouli, Kristi Griffiths, Fabio Ramos, Peter Dayan, Bernard W. Balleine

Version 2 Published: June 11, 2019 • https://doi.org/10.1371/journal.pcbi.1006903

Article | Open Access | Published: 18 March 2022

Using deep learning to predict human decisions and using cognitive models to explain deep learning models

Matan Fintz, Margarita Osadchy & Uri Hertz

Scientific Reports **12**, Article number: 4736 (2022) | Cite this article

RESEARCH ARTICLE | BIOLOGICAL SCIENCES

Adversarial vulnerabilities of human decision-making

Amir Dezfouli, Richard Nock, and Peter Dayan Authors Info & Affiliations

Edited by James L. McClelland, Stanford University, Stanford, CA, and approved October 3, 2020 (received for review August 10, 2020)

November 4, 2020 | 117 (46) 29221-29228 | https://doi.org/10.1073/pnas.2016921117

REPORT

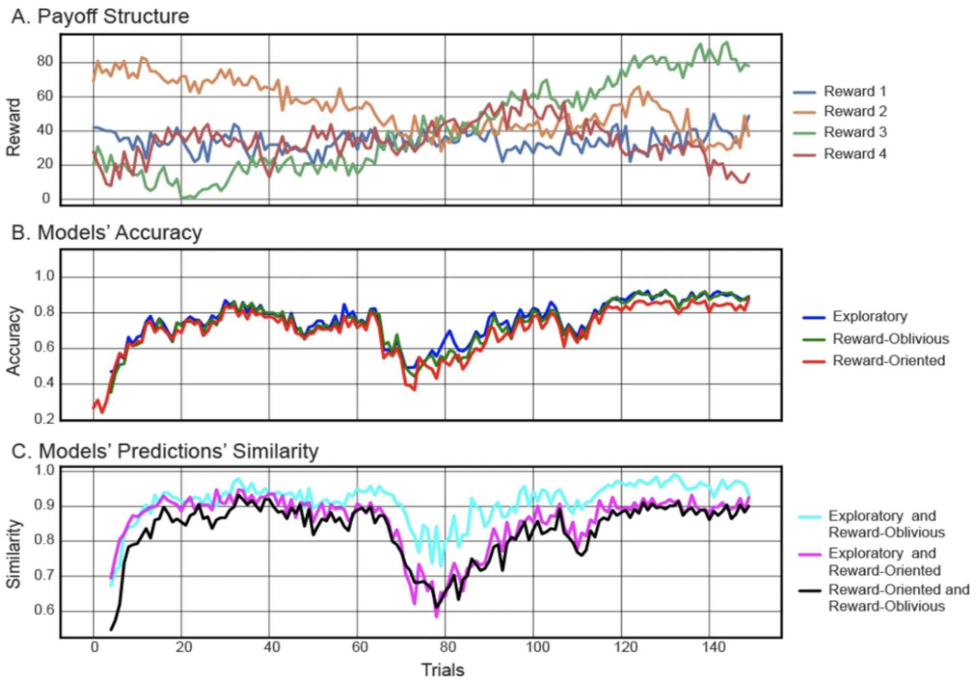
Using large-scale experiments and machine learning to discover theories of human decision-making

JOSHUA C. PETERSON, DAVID D. BOURGIN, MAYANK AGRAWAL, DANIEL REICHMAN, AND THOMAS L. GRIFFITHS Authors Info & Affiliations

SCIENCE • 11 Jun 2021 • Vol 372, Issue 6547 • pp. 1209-1214 • DOI: 10.1126/science.abe2629

Strategy 3 - Using simple models to explain a DNN model

Used a DNN model as an exploratory tool to predict different types of human behaviour in a multi-armed bandit task, and explicit, theory-driven models, to explain the DNN model



Fintz, Osadchy, & Hertz, Scientific Reports, 2022

Strategy 3 - Using simple models to explain a DNN model

- DNNs can shed light on previously ignored human behaviours, and simpler models can be used to explain DNNs
- Applicable in fields where the input space is multidimensional and inference is made from noisy data
- Experimental manipulations

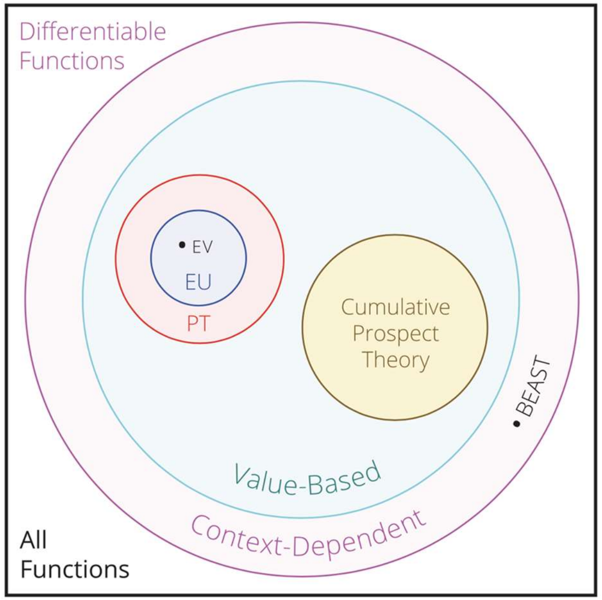
Limitations:

- Rather general explanations

Fintz, Osadchy, & Hertz, Scientific Reports, 2022

Strategy 4 - Using large-scale experiments and ML to discover theories of human decision-making

1. Evaluated most competitive theories at each level of the hierarchy ⟹ best-performing theory belongs to the most complex (less interpretable) class
2. Conducted a second pass of the method to explain the selected model

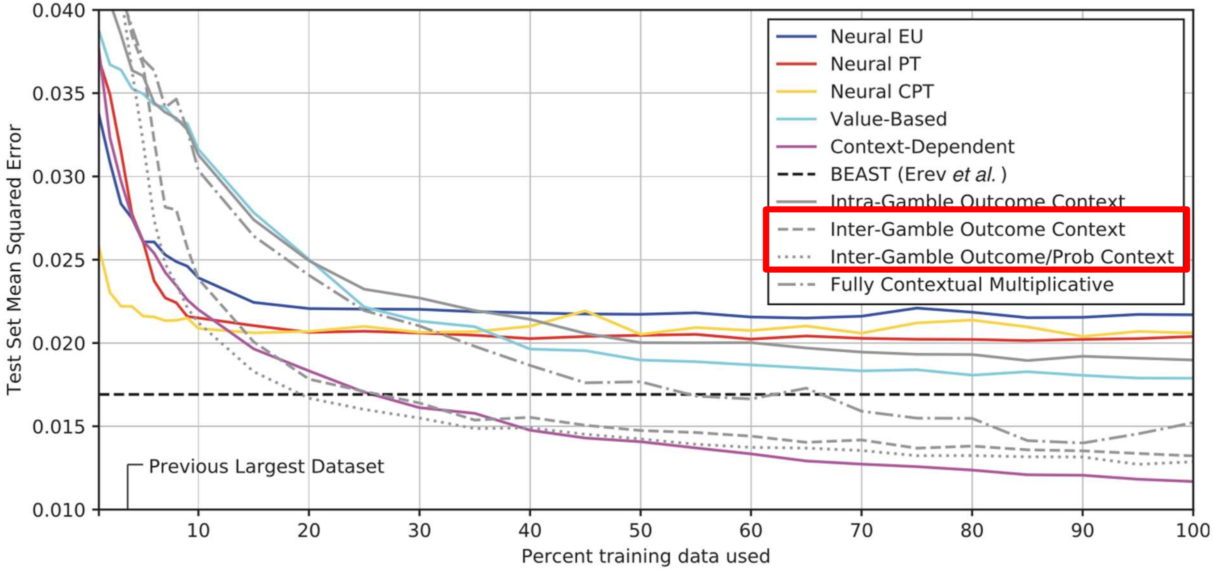


Hierarchy of theoretical assumptions

Peterson et al., Science, 2021

Strategy 4 - Using large-scale experiments and ML to discover theories of human decision-making

2. Conducted a second pass of the method to identify the aspects of context responsible for better model performance



Peterson et al., Science, 2021

Strategy 4 - Using large-scale experiments and ML to discover theories of human decision-making

- When differentiated, psychological theories can be combined with gradient-based optimization approaches from machine learning to broadly search the space of theories and obtain clear scientific explanations

Limitations:

- Need for big datasets

Peterson et al., Science, 2021

Summary

- XAI for cognitive research is in its infancy
- Most approaches involve experimental manipulations, tests on different cognitive tasks, and comparison with simpler models
- These approaches provide a general overview of the process
- There is scope to explore how to apply existing interpretations algorithms to cognitive models to provide more detailed explanations (e.g., how individual decisions come about) and to test proposed methods on a wider range of tasks and stimuli

Implications

Only by providing thorough explanations of how a particular model came to its prediction, we can understand its contribution to the existing body of knowledge, thereby:

- Having a point of reference and advancing our theoretical knowledge on the given cognitive process
- Getting closer to the 'real' AI that learns and thinks like humans (Lake et al., 2017)

Thank you!

Lioba
Enk

Max
Hinrichs

Fabian
Kamp

Bianca
Serio

Simon M.
Hofmann

References

Binz, M., & Schulz, E. (2022). Using cognitive psychology to understand GPT-3, *PsyArXiv*

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … Amodei, D. (2020). *Language Models are Few-Shot Learners* (arXiv:2005.14165). arXiv. https://doi.org/10.48550/arXiv.2005.14165

Dezfouli, A., Griffiths, K., Ramos, F., Dayan, P., & Balleine, B. W. (2019). Models that learn how humans learn: the case of decision-making and its disorders. *PLoS computational biology*, *15*(6), e1006903.

Fintz, M., Osadchy, M., & Hertz, U. (2022). Using deep learning to predict human decisions and using cognitive models to explain deep learning models. *Scientific reports*, *12*(1), 1-12.

Hofmann, S. M., Beyer, F., Lapuschkin, S., Goltermann, O., Loeffler, M., Müller, K.-R., Villringer, A., Samek, W., & Witte, A. V. (2022). Towards the interpretability of deep learning models for multi-modal neuroimaging: Finding structural changes of the ageing brain. *NeuroImage, 261*, 119504. https://doi.org/10.1016/j.neuroimage.2022.119504

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, *40*.

Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science, 372(*6547), 1209–1214. https://doi.org/10.1126/science.abe2629

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K. R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, *109*(3), 247-278.

Thomas, A. W., Ré, C., & Poldrack, R. A. (2022). Comparing interpretation methods in mental state decoding analyses with deep learning models (arXiv:2205.15581). *arXiv*. https://doi.org/10.48550/arXiv.2205.15581