Master
Reinforcement Learning 2022
Lecture 10:
Eval & Future

Aske Plaat

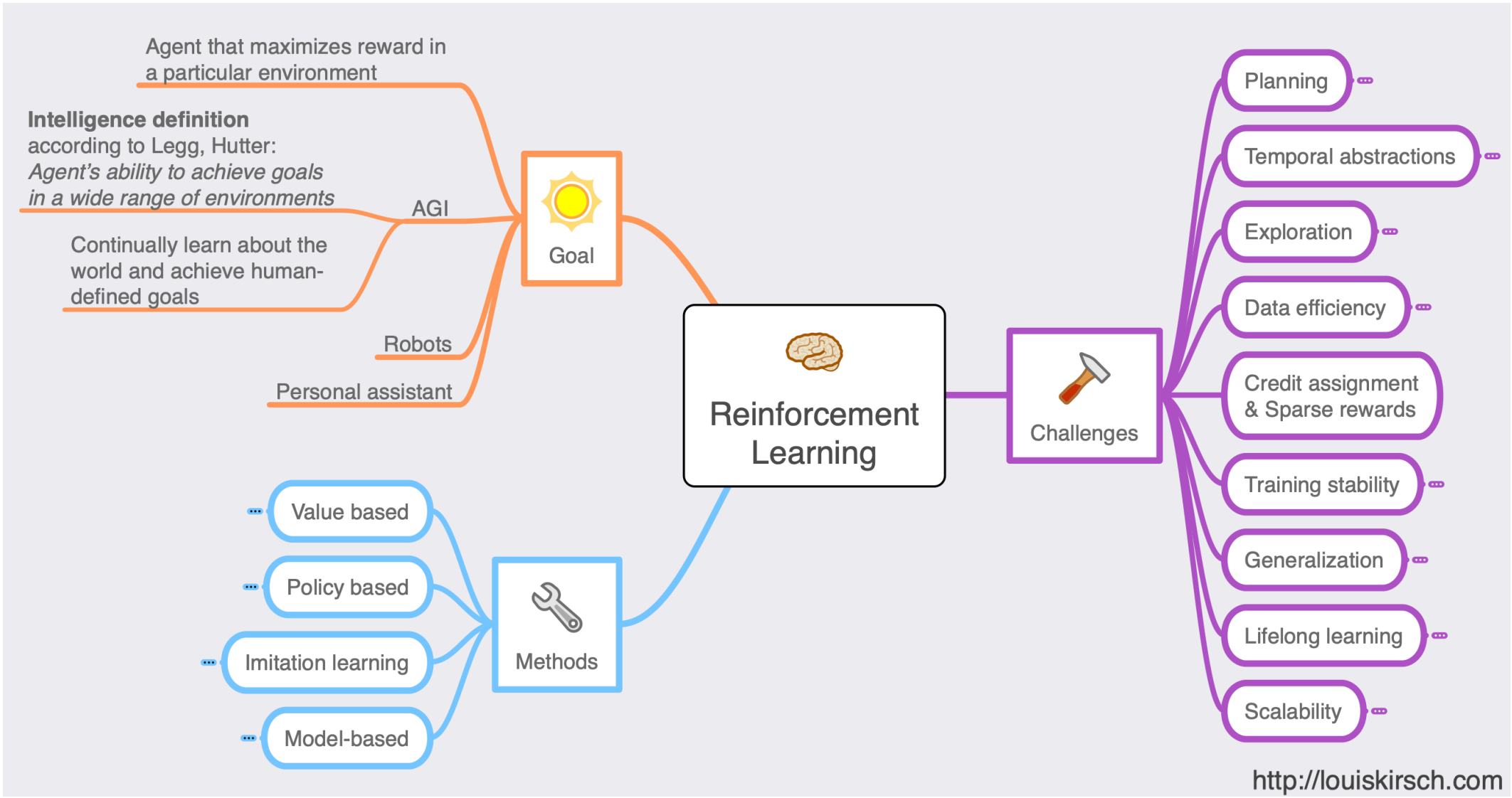
Different Approaches

- Model-free
  - Value-based [2,3]
  - Policy-based [4]
- Model-based
  - Learned [5]
  - Perfect; Two-Agent [6]
- Multi-agent [7]
- Hierarchical Reinforcement Learning (Sub-goals) [8]
- Meta Learning [9]

Eval

- What should we keep?
  - Lectures
  - Assignments
- What can be improved?
  - Lectures
  - Assignments
- Other remarks

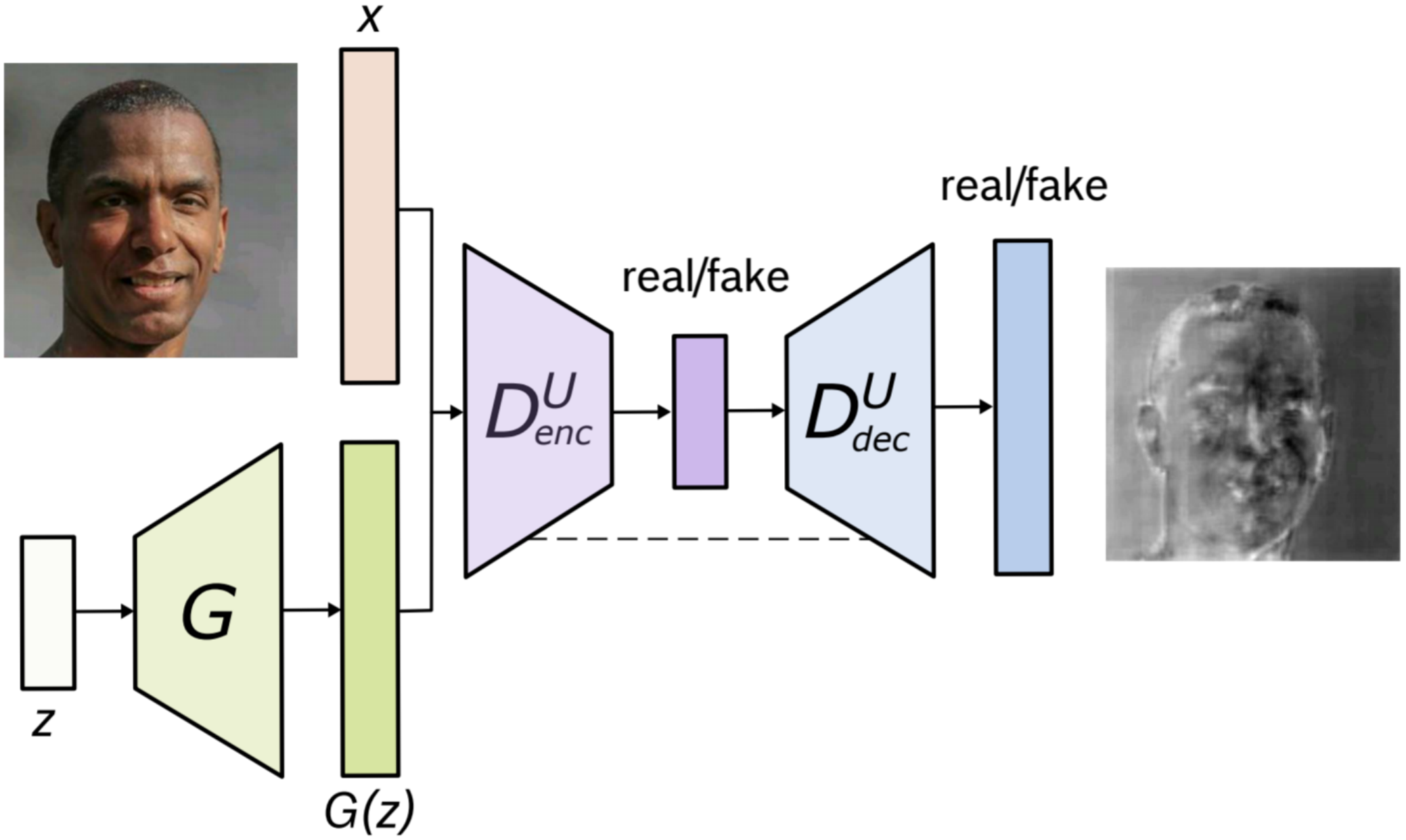
Future


Agent that maximizes reward in a particular environment
Intelligence definition according to Legg, Hutter: Agent's ability to achieve goals in a wide range of environments
AGI
Continually learn about the world and achieve human-defined goals
Robots
Personal assistant
Goal
Reinforcement Learning
Value based
Policy based
Imitation learning
Model-based
Methods
Challenges
Planning
Temporal abstractions
Exploration
Data efficiency
Credit assignment & Sparse rewards
Training stability
Generalization
Lifelong learning
Scalability
http://louiskirsch.com

Extensions

- Model-based
  - Robots
- Multi-Agent
  - Real time strategy games
- Hierarchical
  - Teams
- Meta
  - Lifelong learning

Generative



Input
*From **real life***

HiFaceGAN
ACMMM 20

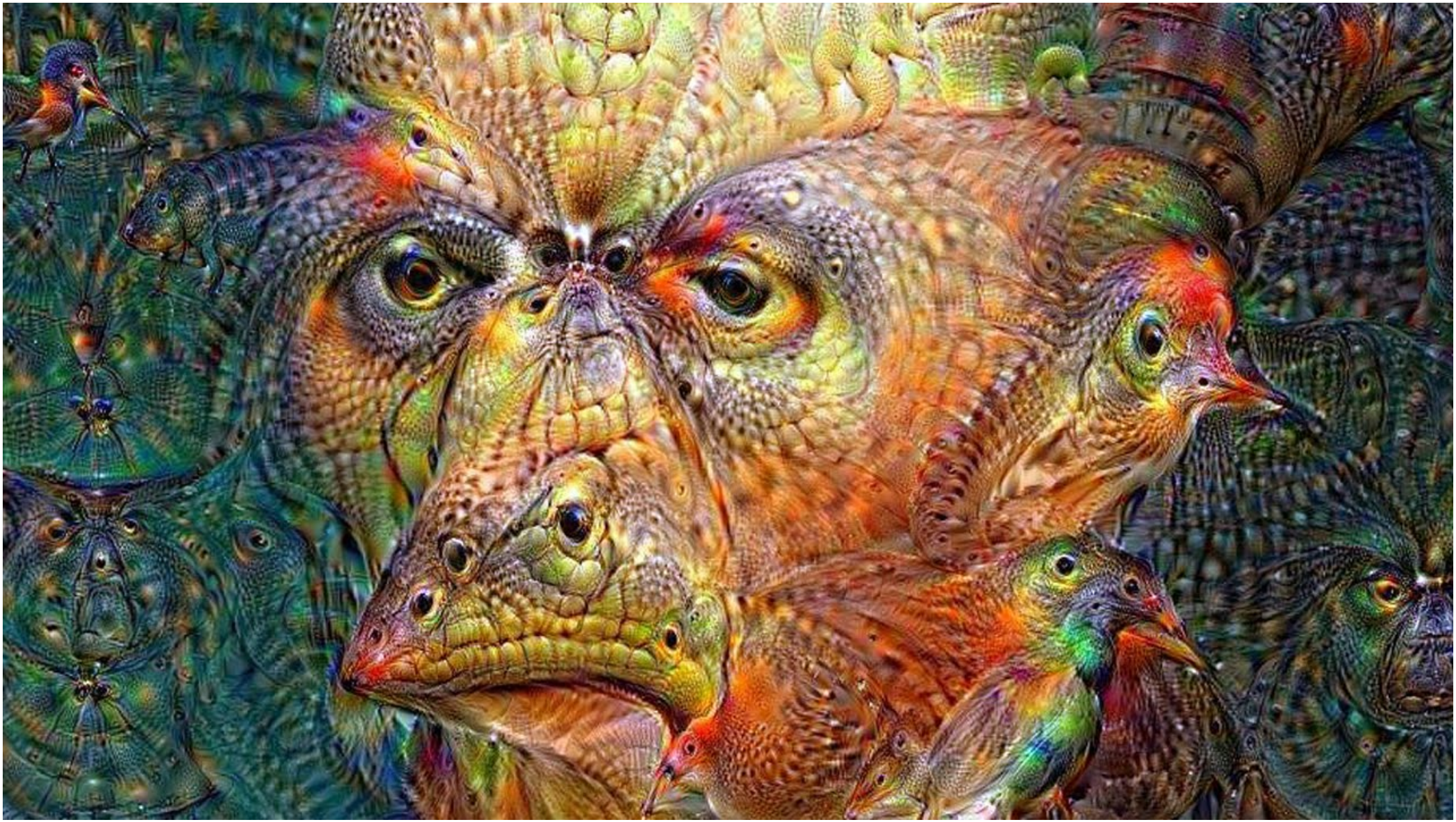
DFDNet
ECCV 20

Wan *et al.*
CVPR 20

PULSE
CVPR 20

GFP-GAN
Ours


Fake



Multi Modal

TEXT PROMPT an armchair in the shape of an avocado. an armchair imitating an avocado.

AI-GENERATED IMAGES

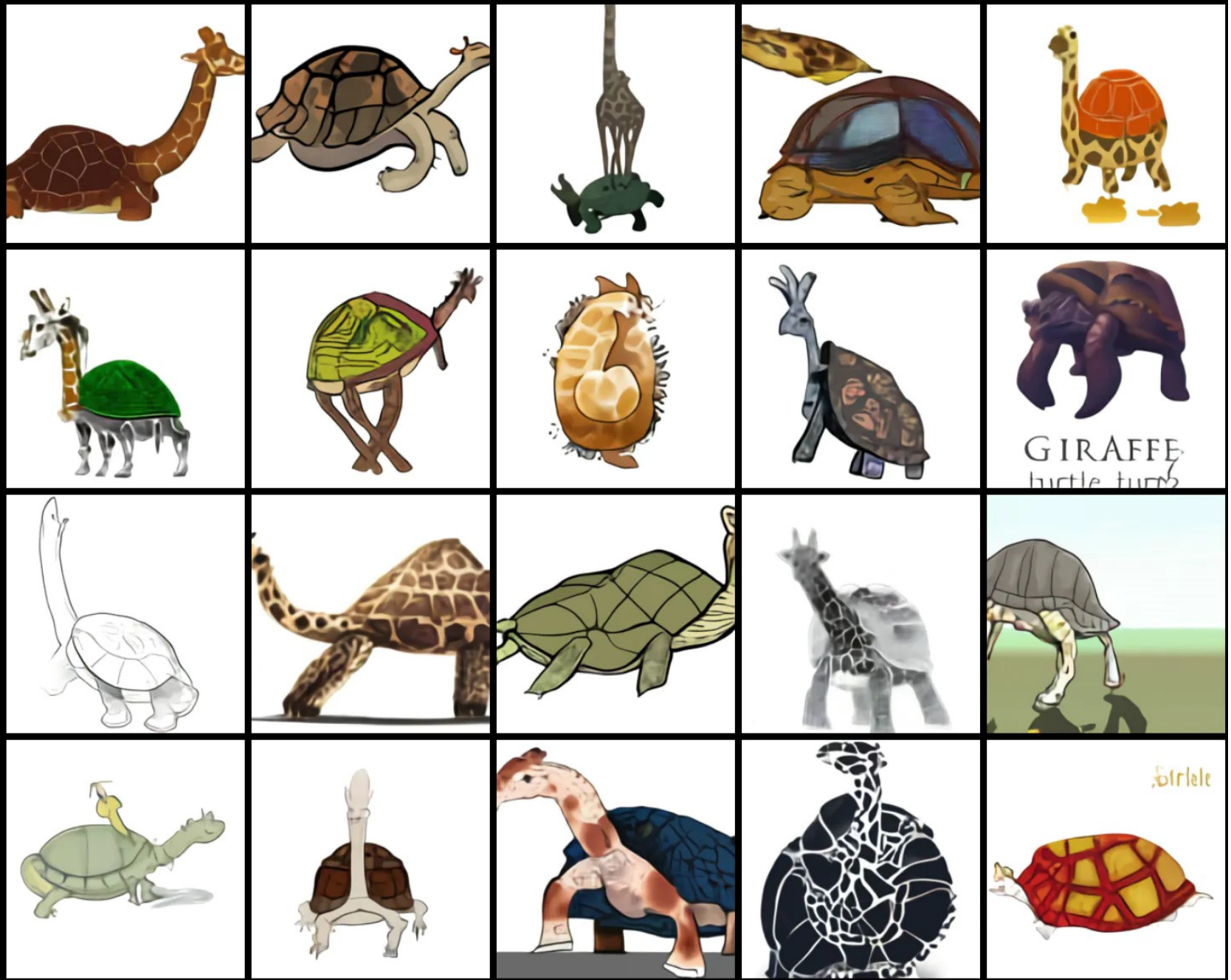
In the preceding visual, we explored DALL·E's ability to generate fantastical objects by combining two unrelated ideas. Here, we explore its ability to take inspiration from an unrelated idea while respecting the form of the thing being designed, ideally producing an object that appears to be practically functional. We found that prompting DALL·E with the phrases "in the shape of," "in the form of," and "in the style of" gives it the ability to do this.

When generating some of these objects, such as "an armchair in the shape of an avocado", DALL·E appears to relate the shape of a half avocado to the back of the chair, and the pit of the avocado to the cushion. We find that DALL·E is susceptible to the same kinds of mistakes mentioned in the previous visual.

TEXT PROMPT

a professional high quality illustration of a giraffe turtle chimera. a giraffe imitating a turtle. a giraffe made of turtle.

AI-GENERATED IMAGES



Introducing the Center for Research on Foundation Models (CRFM)

This new center at Stanford convenes scholars from across the university to study the technical principles and societal impact of foundation models.

Aug 18, 2021

Computer Science > Machine Learning

[Submitted on 16 Aug 2021 (v1), last revised 18 Aug 2021 (this version, v2)]

On the Opportunities and Risks of Foundation Models

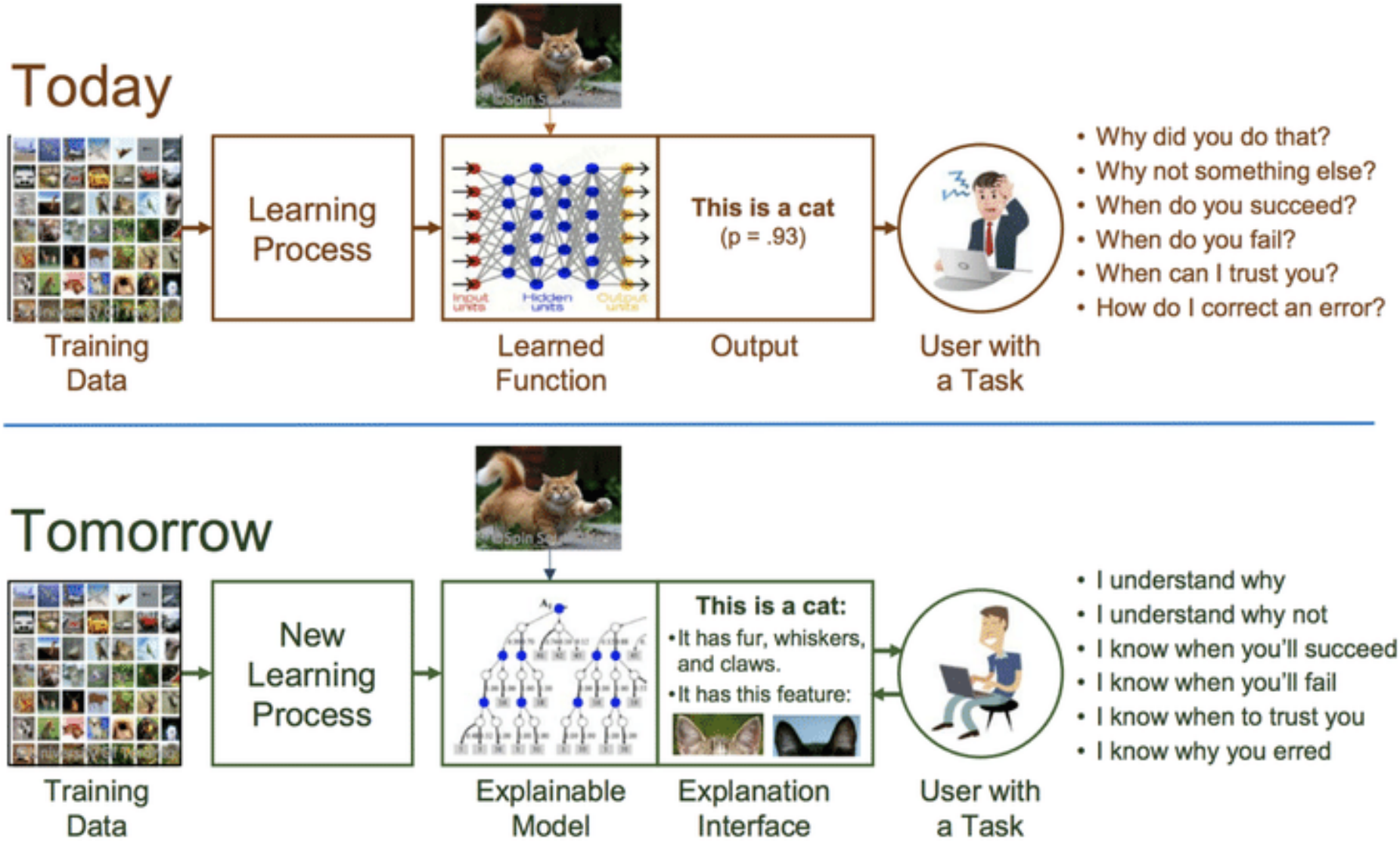
Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang et al. (14 additional authors not shown)

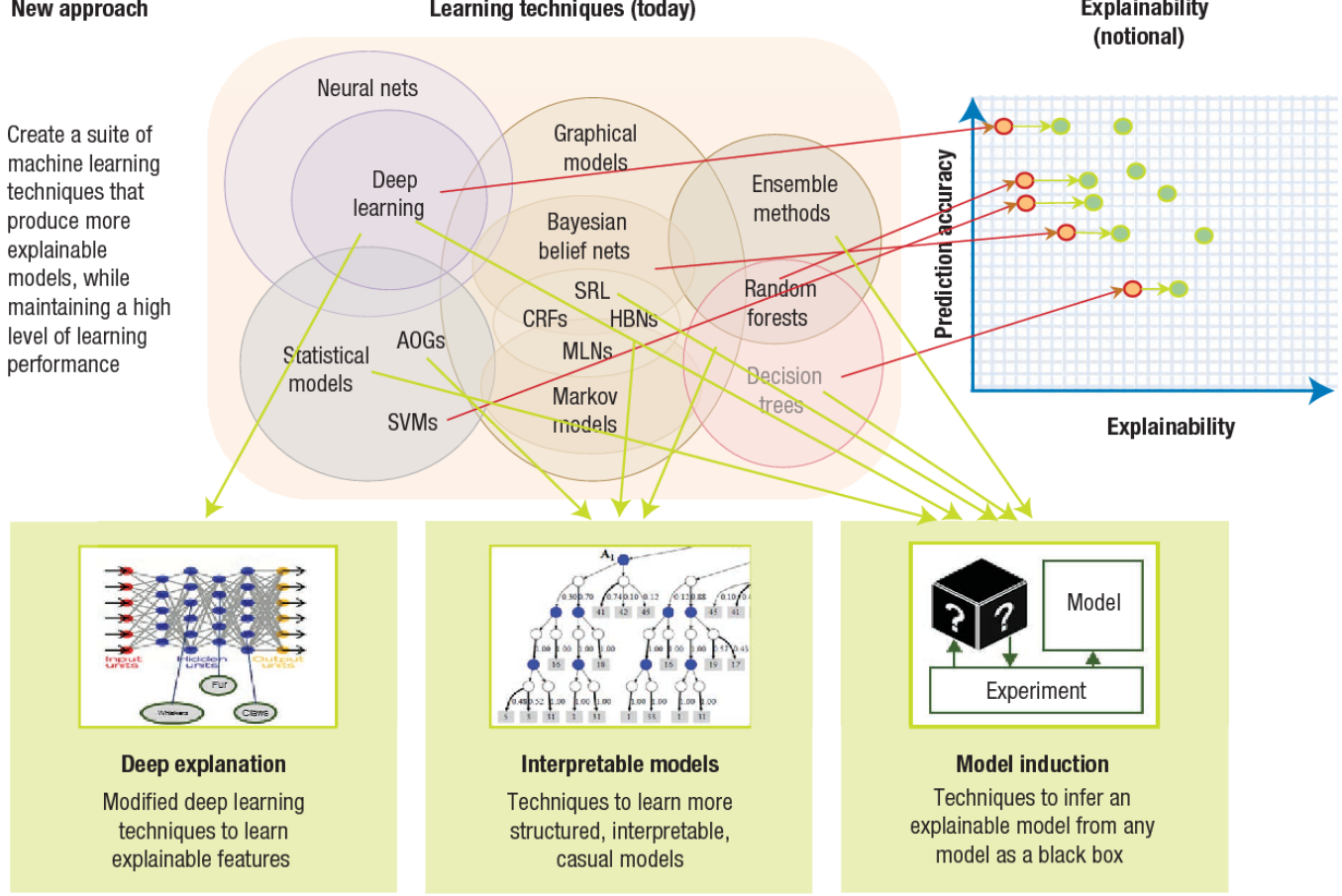
AI is undergoing a paradigm shift with the rise of models (e.g., BERT, DALL-E, GPT-3) that are trained on broad data at scale and are adaptable to a wide range of downstream tasks. We call these models foundation models to underscore their critically central yet incomplete character. This report provides a thorough account of the opportunities and risks of foundation models, ranging from their capabilities (e.g., language, vision, robotics, reasoning, human interaction) and technical principles(e.g., model architectures, training procedures, data, systems, security, evaluation, theory) to their applications (e.g., law, healthcare, education) and societal impact (e.g., inequity, misuse, economic and environmental impact, legal and ethical considerations). Though foundation models are based on standard deep learning and transfer learning, their scale results in new emergent capabilities,and their effectiveness across so many tasks incentivizes homogenization. Homogenization provides powerful leverage but demands caution, as the defects of the foundation model are inherited by all the adapted models downstream. Despite the impending widespread deployment of foundation models, we currently lack a clear understanding of how they work, when they fail, and what they are even capable of due to their emergent properties. To tackle these questions, we believe much of the critical research on foundation models will require deep interdisciplinary collaboration commensurate with their fundamentally sociotechnical nature.

Comments: Authored by the Center for Research on Foundation Models (CRFM) at the Stanford Institute for Human-Centered Artificial Intelligence (HAI)
Subjects: **Machine Learning (cs.LG)**; Artificial Intelligence (cs.AI); Computers and Society (cs.CY)
Cite as: arXiv:2108.07258 **[cs.LG]**
(or arXiv:2108.07258v2 **[cs.LG]** for this version)

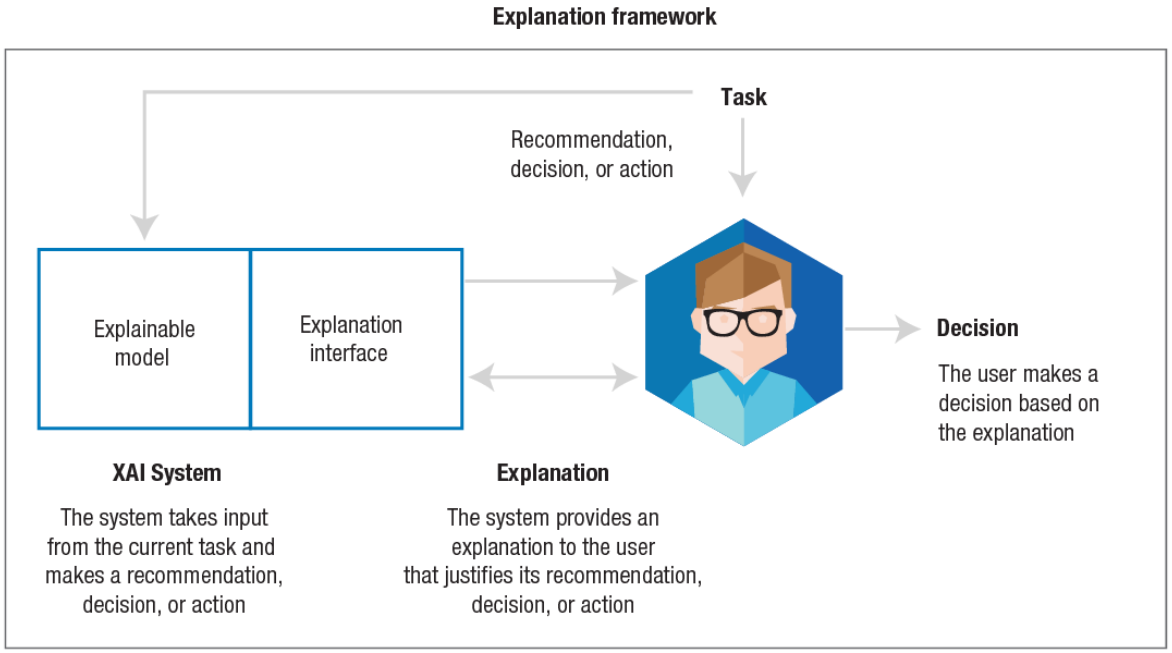
Submission history

Explainable




New approach
Learning techniques (today)
Explainability (notional)
Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance
Neural nets
Deep learning
Graphical models
Bayesian belief nets
Ensemble methods
SRL
CRFs
HBNs
MLNs
Random forests
Statistical models
AOGs
SVMs
Markov models
Decision trees
Prediction accuracy
Explainability
Model
Experiment
Deep explanation
Modified deep learning techniques to learn explainable features
Interpretable models
Techniques to learn more structured, interpretable, casual models
Model induction
Techniques to infer an explainable model from any model as a black box

(a)


Explanation framework
Task
Recommendation, decision, or action
Explainable model
Explanation interface
Decision
The user makes a decision based on the explanation
XAI System
The system takes input from the current task and makes a recommendation, decision, or action
Explanation
The system provides an explanation to the user that justifies its recommendation, decision, or action

(b)


thank you