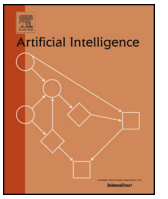
Artificial Intelligence 300 (2021) 103546

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint



Kandinsky Patterns ☆

Heimo Müller *, Andreas Holzinger

Medical University Graz, Austria

ARTICLE INFO

Article history:
Received 16 September 2019
Received in revised form 23 May 2021
Accepted 3 June 2021
Available online 9 June 2021

Keywords:
Explainable AI
Explainability
Synthetic test data
Ground truth

ABSTRACT

Kandinsky Figures and Kandinsky Patterns are mathematically describable, simple, self-contained hence controllable synthetic test data sets for the development, validation and training of visual tasks and explainability in artificial intelligence (AI). Whilst Kandinsky Patterns have these computationally manageable properties, they are at the same time easily distinguishable by human observers. Consequently, controlled patterns can be described by *both* humans and computers. We define a Kandinsky Pattern as a set of Kandinsky Figures, where for each figure an "infallible authority" defines that the figure belongs to the Kandinsky Pattern. With this simple principle we build training and validation data sets for testing explainability, interpretability and context learning. In this paper we describe the basic idea and some underlying principles of Kandinsky Patterns. We provide a Github repository and invite the international AI research community to a challenge to experiment with our Kandinsky Patterns. The goal is to help expand and advance the field of AI, and in particular to contribute to the increasingly important field of explainable AI.

1. Introduction

AI is currently very successful, due to (i) advances in statistical machine learning ("deep learning"), (ii) the availability of large amounts of training data, and (iii) the available computing power [1], [2]. The high complexity, nonlinearity, and high dimensionality of such approaches make them difficult for a human to interpret, and therefore such approaches are considered "black box" models [3].

Boosted by DARPA's Explainable Artificial Intelligence Program [4], the field of explainable AI (xAI) has experienced a tremendous renaissance. Due to the imoprtance of legal and ethical considerations, explainability became enormously relevant and has established itself as an important concept. In roughly simplified terms, explainability technically highlights decision relevant parts of machine representations and/or parts which contributed to model accuracy in training and the xAI community has already developed a variety of successful methods. However, explainability does not refer to a human model. In certain application domains, e.g. in the medical domain, there is a need for going beyond explainability, i.e. there is a need for causability. Causability [5] is neither a typo nor a synonym for Causality [6]. The term Causa-bil-ity was introduced in reference to the well-known term Usa-bil-ity [7]. Causability has been defined as the measurable extent to which an explanation (resulting from an explainable AI method) to a human achieves a specified level of causal understanding

☆ This paper is part of the Special Issue on Explainable AI.

* Corresponding author.
E-mail address: heimo.mueller@medunigraz.at (H. Müller).
URL: http://human-centered.ai (H. Müller).

https://doi.org/10.1016/j.artint.2021.103546

measured with effectiveness, efficiency and satisfaction in a specified context of use - similar to usability. This can be measured with the System Causability Scale (SCS) [8]. Consequently, causability refers to a human model and the understanding can be ensured when mapping explainability with causability. A successful mapping between the two would require new human-AI interfaces which allow domain experts to interactively ask questions and counterfactual questions to gain insight into the underlying *independent* explanatory factors of a result [9]. In an ideal world both human and AI statements would be identical and congruent with the *ground truth,* which is defined for both humans and AI equally [8]. Compared to the map metaphor, the explainability–causability mapping is about establishing connections and relations - not drawing a new map. It is about identifying the *same areas in two completely different maps.* For example, when explaining predictions of deep learning models we apply an explanation method, e.g. simple sensitivity analysis, to understand the prediction in terms of the input variables. The result of such an explainability method can be a heatmap. This heatmap visualization indicates which pixels need to be changed to make the image look (from the AI-systems perspective!) more or less like the predicted class [10]. On the other hand there are the corresponding human concepts, and "contextual understanding" needs effective mapping of them both [11], and is among the future grand goal of human-centered AI [12].

The central motivation for this work was the lack of ground truth when testing with real data sets. Image classifiers operate on low-level features (e.g. lines, circles, etc.) rather than high-level concepts, and with domain concepts (e.g. images with a storefront). With our Kandinsky exploration environment we can produce Kandinsky Figures and Kandinsky Patterns along with the ground truth. With these mathematically describable, simple, and controllable synthetic test data sets we enhance the development, validation and training of visual tasks and explainability. Very important is that they are at the same time easily distinguishable by human observers.

2. Kandinsky patterns

Wassily Kandinsky (1866–1944) was an influential Russian painter [13]. As his career progressed, Kandinsky produced increasingly abstract images. For a period from 1922 to 1933 he taught at the famous Bauhaus school in Germany, which celebrated simple colors and forms. Kandinsky was a theorist as well as an artist, and he derived profound meaning from aesthetic experiences. One of Kandinsky's ideas was that there are certain fundamental associations between colors and shapes [14], e.g. he proposed Yellow-Triangle, Blue-Circle, and Red-Square. These associations were formulated introspectively, however, he did conduct his own survey at the Bauhaus in 1923 and postulated a correspondence between color and form. Subsequent empirical studies used preference judgments to test Kandinsky's original color-form combinations, usually yielding inconsistent results. Recent findings suggest that there is no implicit association between the original color-form combinations and hence cannot be considered as a universal property of the visual system [15]. In our work we do not pursue this hypothesis any further, but take only the visual principles of Kandinsky as starting point and eponym for the following definitions.

A **Kandinsky Figure** is a square image containing *1* to *n* geometric objects. Each object is characterized by its shape, color, size and position within this square. Objects do not overlap and are not cropped at the border. All objects must be easily recognizable and clearly distinguishable by a human observer.

The set of all possible Kandinsky Figures k is defined by the general definition together with a specific set of values for shape, color, size, position and the number of geometric objects. In the following examples we use for shape the values circle, square and triangle; for color we use the values red, blue, yellow, and we allow arbitrary positions and size with the restriction that it is still recognizable. Furthermore, we require each Kandinsky Figure to contain exactly 4 objects in the following illustrative examples. In the demo implementation this fact is embedded in the base class "Kandinsky Universe", and in the generator functions,[1] see Fig. 1.

A Statement $s(k)$ about a Kandinsky Figure k is either a mathematical function, $s(k) \rightarrow B$; with $B(0,1)$ or a natural language statement, which is either true or false.

Remark: The evaluation of a natural language statement is always done in a specific *context*. In the followings examples we use well known concepts from human perception and linguistic theory. If $s(k)$ is given as an algorithm, it is essential that the function is a pure function, which is a computational analogue of a mathematical function.

A **Kandinsky Pattern** K is defined as the subset of all possible Kandinsky Figures k with $s(k) \rightarrow 1$ or the natural language statement is true. $s(k)$ and a natural language statement are equivalent, if and only if the resulting Kandinsky Patterns contains the same Kandinsky Figures. $s(k)$ and the natural language statement are defined as the **Ground Truth** of a Kandinsky Pattern.

In a deep learning solution classification algorithm for a visual pattern is usually represented as a highly non-linear, high-dimensional network. One aim of explainable AI is to identify areas of activation within the network structure, which correspond to concepts in the natural language statement.

Problem 1: How can we explain a Kandinsky Pattern, if we do not know the Ground Truth and the membership of Kandinsky Figures to a Kandinsky Pattern is only known for a limited number of Kandinsky Figures.

Problem 2: Generate a natural language statement, which is easily understandable and equivalent to the machine explanation (classification algorithm).

[1] https://github.com/human-centered-ai-lab/app-kandinsky-pattern-generator.

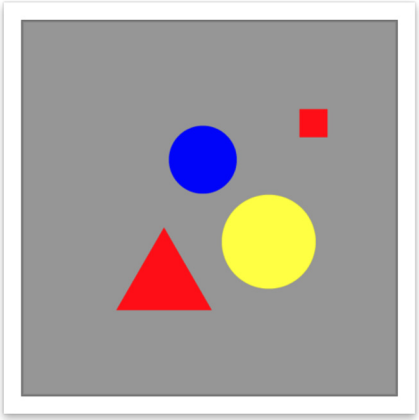


Fig. 1. A Kandinsky Figure with 4 objects. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

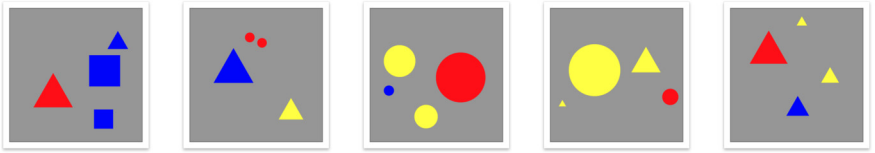


Fig. 2. Five Kandinsky Figures of a Kandinsky Pattern.

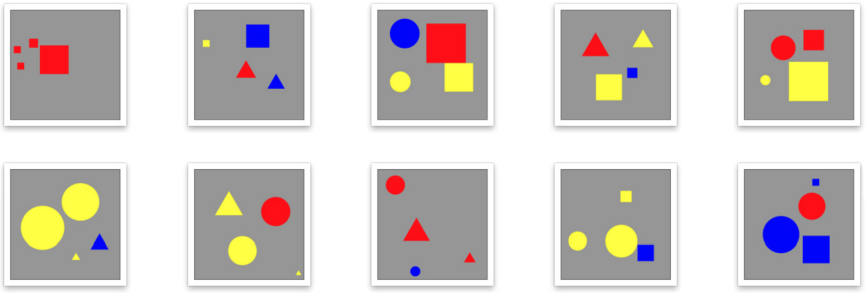


Fig. 3. Kandinsky Figures of $h_1(k)$, the first row shows contradictions.

The process of explanation is the generation and refinement of a hypothesis to find the underlying description. The validation is achieved by the scientific method of asking a question, forming a testable hypothesis, setting up the experimental design, running the experiment and either accepting the hypothesis, rejecting it or, in the third case according to [16], one cannot make any assumption.

The ground truth is used to prove or disprove research hypotheses. "Ground truthing" consequently refers to the process of collecting the proper objective (provable!) data for testing the hypothesis. For a machine learning algorithm an explanation can be seen as the successful classification algorithm of a Kandinsky pattern.

The following example illustrates the above:

The ground truth $gt(k) =$ "*the Kandinsky Figure has two pairs of objects with the same shape, in one pair the objects have the same color, in the other pair different colors, two pairs are always disjunct, i.e. they don't share objects*" defines the Kandinsky Pattern K_{gt}, see Fig. 2.

For a more general hypothesis $h_1(k)$ = *"the Kandinsky Figure has two pairs of objects with the same shape"* we see that $K_{h1} \setminus K_{gt} \neq \emptyset$, i.e. the Kandinsky Pattern of $h_1(k)$ contains Kandinsky Figures which are not in the Kandinsky Pattern of the ground truth. Fig. 3 shows Kandinsky Figures according to $h_1(k)$, the first row is a contradiction to the ground truth, i.e. it falsifies $h_1(k)$.

A specific hypothesis like $h_2(k)$ = *"the Kandinsky Figure consists of two triangles with different color and two circles of same color"* generates a Kandinsky Pattern K_{h2} with $K_{gt} \setminus K_{h2} \neq \emptyset$, i.e. the Kandinsky Pattern of $h_2(k)$ is missing Kandinsky Figures which are in the Kandinsky Pattern of the ground truth. Fig. 4 shows in the first row Kandinsky Figures according to $h_2(k)$ and Kandinsky Figures from K_{gt} in the second row, which falsify $h_2(k)$. $K_h \setminus K_{gt} \cup K_{gt} \setminus K_h$ is the set of contradictions for a given hypothesis h.

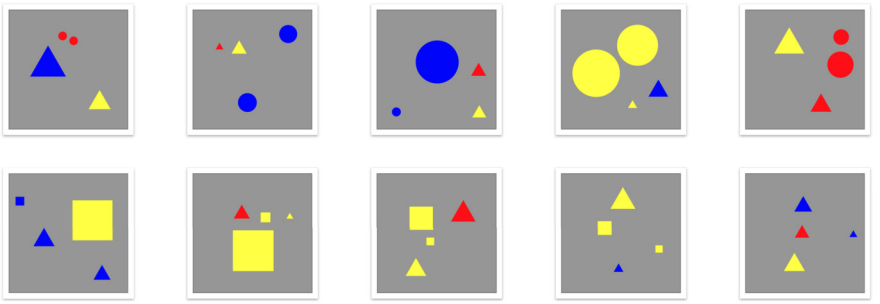


Fig. 4. Kandinsky Figures of $h_2(k)$ in the first row, Kandinsky Figures from K_{gt} in the second row, which falsify $h_2(k)$.

3. Background

In a natural language statement about a Kandinsky Figure humans use a series of basic concepts which are combined through logical operators. The following (incomplete) examples illustrate some concepts of increasing complexity.

- Basic concepts given by the definition of a Kandinsky Figure: a set of *objects*, described by *shape*, *color*, *size* and *position*.
- Existence, numbers, set-relations (*number*, *quantity* or *quantity ratios* of objects), e.g. *"a Kandinsky Figure contains 4 red triangles and more yellow objects than circles"*.
- Spatial concepts describing the arrangement of objects, either absolute (*upper*, *lower*, *left*, *right*, ...) or relative (*below*, *above*, *on top*, *touching*, ...), e.g. *"in a Kandinsky Figure red objects are on the left side, blue objects on the right side, and yellow objects are below blue squares"*.
- Gestalt concepts (see below) e.g. *closure*, *symmetry*, *continuity*, *proximity*, *similarity*, e.g. *"in a Kandinsky Figure objects are grouped in a circular manner"*.
- Domain concepts, e.g. *"a group of objects is perceived as a "flower""*.

In their experiments Hubel & Wiesel (1962) [17] discovered, among others, that the visual system builds an image from very simple stimuli into more complex representations. This inspired the neural network community to see their so-called "deep learning" models as a cascading model of cell types, which follows always similar simple rules: at first lines are learned, then shapes, then objects are formed, eventually leading to **concept representations.** By use of backpropagation such a model is able to discover intricate structures in large data sets to indicate how the internal parameters should be adapted, which are used to compute the representation in each layer from the representation in the previous layer [2]. Building *concept representations* refers to the human ability to learn categories for objects and to recognize new instances of those categories. In machine learning, concept learning is defined as the inference of a Boolean-valued function from training examples of its inputs and outputs [18], in other words it is training an algorithm to distinguish between examples and non-examples (we call the latter contradictions).

Concept learning has been a relevant research area in machine learning for a long time and had it origins in cognitive science, defined as the search for attributes which can be used to distinguish exemplars from non exemplars of various categories [19]. The ability to think in abstractions is one of the most powerful tools humans possess. Technically, humans order their experience into coherent categories by defining a given situation as a member of that collection of situations for which responses *x, y,* etc. are most likely appropriate. This classification is not a passive process and to understand how humans learn abstractions is essential not only to the understanding of human thought, but to building artificial intelligence machines [20]. One interesting study was performed by [21], who presented two sets (A, B) of simple diagrams, where all the diagrams from set A have a common factor or attribute, which is lacking in all the diagrams of set B. The problem is to find the common factor. These problems were also described in the popular book by [22].

In computer vision an important task is to find a likely interpretation W for an observed image I, where W includes information about the spatial location, the extent of objects, the boundaries etc. Let SW be a function associated with an interpretation W that encodes the spatial location and extent of a component of interest, where $SW_{(i,j)} = 1$ for each image location (i, j) that belongs to the component and 0 elsewhere. Given an image, obtaining an optimal or even likely interpretation W, or associated SW, can be difficult. For example, in edge detection previous work [23] asked what is the probability of a given location in a given image belonging to the component of interest.

[24] presented a model of concept learning that is both computationally grounded and able to fit to human behavior. He argued that two apparently distinct modes of generalizing concepts – abstracting rules and computing similarity to exemplars – should both be seen as special cases of a more general *Bayesian learning framework*. Originally, Bayes (and more specific [25]) explained the specific workings of these two modes, i.e. which rules are abstracted, how similarity is measured, why generalization should appear in different situations. This analysis also suggests why the rules/similarity distinction, even if not computationally fundamental, may still be useful at the algorithmic level as part of a principled approximation to fully Bayesian learning.

Gestalt-Principles ("Gestalt" = German for shape) are a set of empirical laws describing how humans gain meaningful perceptions and make sense of chaotic stimuli of the real-world. As Gestalt-cues they have been used in machine learning for a long time. Particularly in learning classification models for segmentation, the task is to classify between "good" segmentations and "bad" segmentations and to use the Gestalt-cues as features (the priors) to train the learning model. Images segmented manually by humans are used as examples of "good" segmentations (ground truth), and "bad" segmentations are constructed by randomly matching a human segmentation to a different image [26]. Gestalt-principles [27] can be seen as rules, i.e. they discriminate competing segmentations only when everything else is equal, therefore we speak more generally of Gestalt-laws and one particular group of Gestalt-laws are the Gestalt-laws of grouping, called Prägnanz [28], which include the law of Proximity: objects that are close to one another appear to form groups, even if they are completely different, the Law of Similarity: similar objects are grouped together; or the law of Closure: objects can be perceived as such, even if they are incomplete or hidden by other objects.

4. Related work

Reasoning and explanation has a long history within the AI/machine learning community [29] and recently quite a number of authors proposed mechanisms for generating explanations by deep learning models.

Among them are Compositional Language and Elementary Visual Reasoning diagnostics dataset (CLEVR) [30], CLEVERER [31], CLOSURE [32], CURI [33], Bongard-LOGO [34], and V-PROM [35] to mention only some. We present here only a tiny fraction of related work and apologize for any work not mentioned here and refer to future work [36].

Within the machine learning community there is an intensive debate on whether neural networks can learn abstract reasoning or whether they merely rely on pure correlation. In a recent paper the authors [37] propose a data set and a challenge to investigate abstract thinking inspired by a well-known human IQ test: the Raven test, or more specifically the Raven's Progressive Matrices (RPM) and Mill Hill Vocabulary Scales, which were developed in 1936 for use in fundamental research into both the genetic and the environmental determinants of "intelligence" [38]. The premise behind RPMs is simple: one must reason about the relationships between perceptually obvious visual features – such as shape positions or line colors – to choose an image that completes the matrix. For example, perhaps the size of squares increases along the rows, and the correct image is that which adheres to this size relation. RPMs are strongly diagnostic of abstract verbal, spatial and mathematical reasoning ability. To succeed at the challenge, models must cope with various generalization 'regimes' in which the training and test data differ in clearly-defined ways.

Kandinsky Patterns can be used as a validation data set for experiments in explainability, similarly as in the following works: [39] proposed a model that focused on discriminating properties of a visible object and jointly predicts a class label. They explained why the predicted label is appropriate for the respective image on the basis of a loss function based on sampling and reinforcement learning that learns to generate sentences that realize a global sentence property, such as class specificity. [40] proposed a technique for producing 'visual explanations' following Gradient-weighted Class Activation Mapping (Grad-CAM), which uses the gradients of any target concept (e.g. logits for "dog" or a caption, etc.), influencing the final convolutional layer to produce a coarse localization map to highlight relevant regions in the image for predicting the concept. [41] introduced so called Concept Activation Vectors (CAVs), which provide an interpretation of a neural networks internal state in terms of more human-friendly concepts. Their key idea is to view the high-dimensional internal state of a neural network as an aid and not as an obstacle. [30] presented a different approach, called CLEVR, which contains a diagnostic data set for testing visual reasoning abilities. The code can be used to render synthetic images and compositional questions for those images e.g. *"How many small spheres are there?"*. Each question in CLEVR is represented both in natural language and as a functional program, the latter representation allows for precise determination of the reasoning skills required to answer each question. Questions in CLEVR test various aspects of visual reasoning including attribute identification, counting, comparison, spatial relationships, and logical operations. However, this work does not deal with concept learning. [42] described a similar approach already earlier by addressing the task of learning novel visual concepts and their interactions with other concepts from a few images with sentence descriptions.

5. Data sets and challenges

Kandinsky Patterns can be used as test data sets for various research questions, e.g. to address and evaluate the following topics:

1. Describe classes of Kandinsky Patterns according to their ability to be classified by machine learning algorithms in comparison to human explanation strategies.
2. Investigate transfer learning of concepts as *numbers*, *geometric positions* and *Gestalt principles* in the classification and explanation of Kandinsky Patterns.
3. Develop mapping strategies from an algorithmic classification to a known human explanation of a Kandinsky Pattern.
4. Automatic generation of a human understandable explanation of a Kandinsky Pattern.

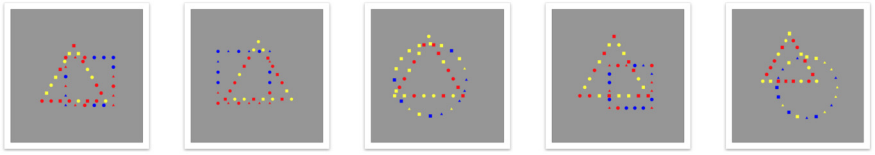


Fig. 5. Kandinsky Figures according to ground truth of challenge 1.

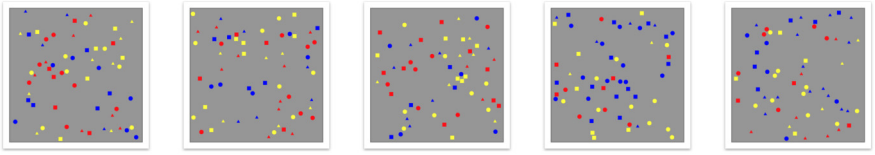


Fig. 6. Kandinsky Figures not belonging to the Kandinsky Pattern of challenge 1.

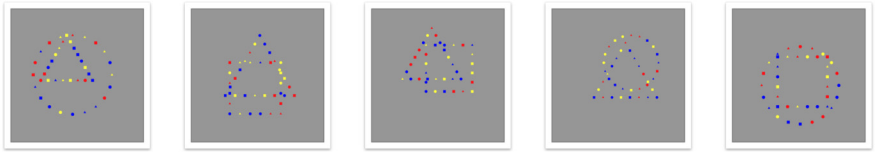


Fig. 7. Kandinsky Figures which falsify a simple hypothesis for challenge 1.

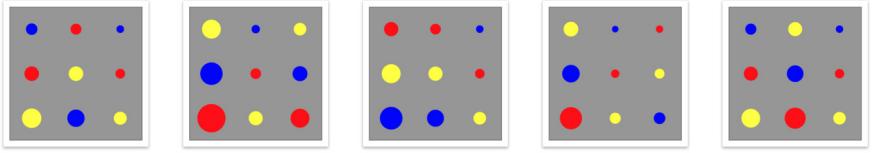


Fig. 8. Kandinsky Figures according to ground truth of challenge 2.

We invite the international machine learning community to experiment with our Kandinsky data set,[2] and re-use and contribute to the Kandinsky software tools.[3]

Please note that the main aim of the training data sets and the following challenges is not in the evaluation of machine learning algorithms, but most of all *in explaining the successful classification by human understandable statements.*

5.1. Challenge 1 - objects and shapes

In the challenge **Objects and Shapes** the ground truth gt(k) is defined as "*in a Kandinsky Figure small objects are arranged on big shapes that are the same as object shapes, in the big shape of type X, no small object of type X exists. Big square shapes only contain blue and red objects, big triangle shapes only contain yellow and red objects and big circle shapes contain only yellow and blue objects*".

Fig. 5 shows Kandinsky Figures according to above ground truth. Fig. 6 shows random Kandinsky Figures with approximately the same number of objects not belonging to the Kandinsky Pattern and Fig. 7 Kandinsky Figures which are generated with a simple but not valid hypothesis.

- **Question 1**: Which machine learning algorithm can classify Kandinsky Figures of challenge 1.
- **Question 2:** Identify layers and regions in the network, which correspond to "small" and "big" shapes and the restrictions on object membership and color.

Download the data set for challenge 1 here: https://tinyurl.com/Kandinsky-C1[4]

[2] https://github.com/human-centered-ai-lab/dat-kandinsky-patterns.
[3] https://github.com/human-centered-ai-lab/app-kandinsky-pattern-generator.
[4] https://github.com/human-centered-ai-lab/dat-kandinsky-patterns/tree/master/challenge-nr-1.

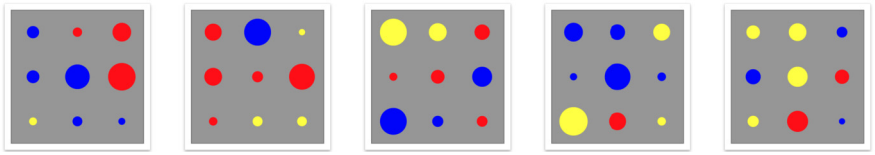


Fig. 9. Kandinsky Figures not belonging to the Kandinsky Pattern of challenge 2.

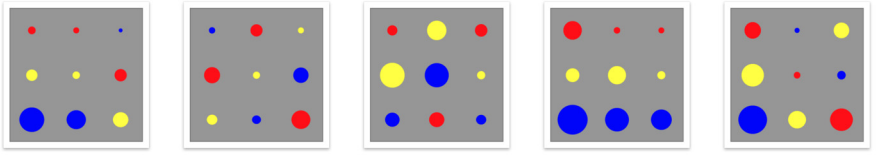


Fig. 10. Kandinsky Figures which falsify a simple hypothesis for challenge 2.

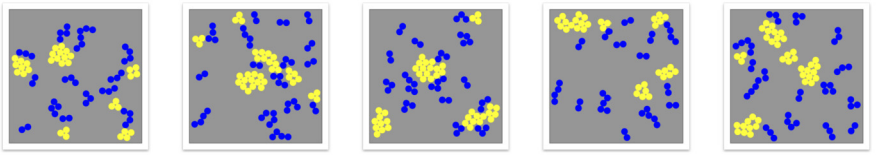


Fig. 11. Kandinsky Figures according to ground truth of challenge 3.

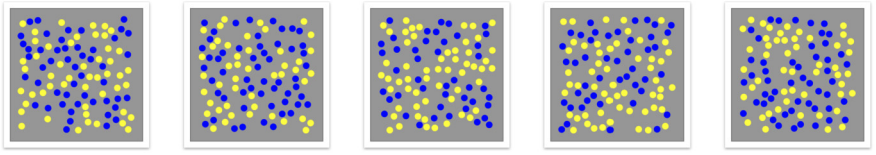


Fig. 12. Kandinsky Figures not belonging to the Kandinsky Pattern of challenge 3.

5.2. Challenge 2 - nine circles

In the challenge **Nine Circles** the set of Kandinsky Figures consists of 9 circles arranged in a regular grid. Fig. 8 shows Kandinsky Figures according to ground truth. Fig. 9 shows Kandinsky Figures not belonging to the Kandinsky Pattern and Fig. 10 shows Kandinsky Figures which are "almost true", i.e. they fulfill a hypothesis similar to ground truth, but are counter factual.

- **Question 1**: Explain the Kandinsky Pattern in an algorithmic way, i.e. train a network which classifies Kandinsky Figures according to ground truth of challenge 2.
- **Question 2:** Explain the Kandinsky Pattern in natural language.

Download the data set for challenge 2 here: https://tinyurl.com/Kandinsky-C2[5]

5.3. Challenge 3 - blue and yellow circles

In the challenge **Blue and Yellow Circles** the set of all possible Kandinsky Figures consists of equal size blue and yellow circles. Fig. 11 shows Kandinsky Figures according to ground truth. Fig. 12 shows Kandinsky Figures with approximately the same number of objects not belonging to the Kandinsky Pattern and Fig. 13 Kandinsky Figures which are "almost true", i.e. they fulfill a hypothesis similar to the ground truth.

- **Question 1**: Explain the Kandinsky Pattern in an algorithmic way, i.e. train a network which classifies Kandinsky Figures according to ground truth of challenge 3.
- **Question 2:** Explain the Kandinsky Pattern in natural language.

[5] https://github.com/human-centered-ai-lab/dat-kandinsky-patterns/tree/master/challenge-nr-2.

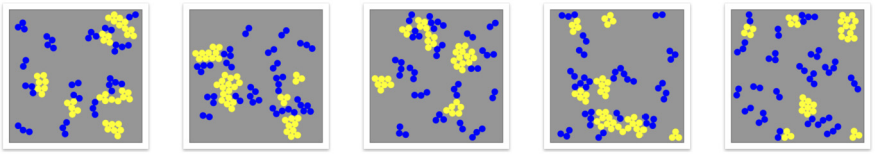


Fig. 13. Kandinsky Figures which falsify a simple hypothesis for challenge 3.

Download the data set for challenge 3 here: https://tinyurl.com/Kandinsky-C3[6]

6. Conclusion

By comparing both the strengths of machine intelligence and human intelligence it is possible to solve problems where we are currently lacking appropriate methods. One overriding question is "How can we perform a task by exploiting knowledge extracted during the solving of previous tasks?" To answer this question it is necessary to get insight into human behavior, but not with the goal of mimicking human behavior, rather to contrast human learning methods to machine learning methods. We hope that our Kandinsky Patterns challenge the international machine learning community and we are looking forward to receiving comments and results. Updated information can be found at the accompanying Web page.[7]

Declaration of competing interest

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Acknowledgements

We are grateful for the valuable comments and encouragement of the anonymous reviewers. Parts of this work have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 824087 (EOSC-Life) and by the Austrian Science Fund (FWF), Project: P-32554 explainable Artificial Intelligence. This publication reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains.

References

[1] J. Schmidhuber, Deep learning in neural networks: an overview, Neural Netw. 61 (1) (2015) 85–117, https://doi.org/10.1016/j.neunet.2014.09.003.
[2] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[3] A. Holzinger, M. Plass, K. Holzinger, G.C. Crisan, C.-M. Pintea, V. Palade, A glass-box interactive machine learning approach for solving np-hard problems with the human-in-the-loop, arXiv:1708.01104.
[4] D. Gunning, D.W. Aha, Darpa's explainable artificial intelligence program, AI Mag. 40 (2) (2019) 44–58, https://doi.org/10.1609/aimag.v40i2.2850.
[5] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 9 (4) (2019) 1–13, https://doi.org/10.1002/widm.1312.
[6] J. Pearl, Embracing causality in default reasoning, Artif. Intell. 35 (2) (1988) 259–271, https://doi.org/10.1016/0004-3702(88)90015-X.
[7] A. Holzinger, Usability engineering methods for software developers, Commun. ACM 48 (1) (2005) 71–74, https://doi.org/10.1145/1039539.1039541.
[8] A. Holzinger, A. Carrington, H. Müller, Measuring the quality of explanations: the system causability scale (scs): Comparing human and machine explanations, in: Special Issue on Interactive Machine Learning, Edited by Kristian Kersting, TU Darmstadt, Künstl. Intell. (J. Artif. Gen. Intell.) 34 (2) (2020) 193–198, https://doi.org/10.1007/s13218-020-00636-z.
[9] A. Holzinger, B. Malle, A. Saranti, B. Pfeifer, Towards multi-modal causability with graph neural networks enabling information fusion for explainable ai, Inf. Fusion 71 (7) (2021) 28–37, https://doi.org/10.1016/j.inffus.2021.01.008.
[10] W. Samek, T. Wiegand, K.-R. Müller, Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models, arXiv:1708.08296.
[11] B.M. Lake, R. Salakhutdinov, J.B. Tenenbaum, Human-level concept learning through probabilistic program induction, Science 350 (6266) (2015) 1332–1338, https://doi.org/10.1126/science.aab3050.
[12] A. Holzinger, C. Biemann, C.S. Pattichis, D.B. Kell, What do we need to build explainable ai systems for the medical domain?, arXiv:1712.09923.
[13] H. Düchting, Wassily Kandinsky, 1866-1944: A Revolution in Painting, Taschen, Köln, 2000.
[14] W. Kandinsky, Über die Formfrage, Der Blaue Reiter 3 (1912) 74–100.
[15] A. Makin, S. Würger, The iat shows no evidence for Kandinsky's color-shape associations, Front. Psychol. 4 (2013) 616, https://doi.org/10.3389/fpsyg.2013.00616.
[16] K. Popper, Die Logik der Forschung. Zur Erkenntnistheorie der modernen Naturwissenschaft, Springer-Verlag, Wien, 1935.
[17] D.H. Hubel, T.N. Wiesel, Receptive fields, binocular interaction and functional architecture in the cat's visual cortex, J. Physiol. 160 (1) (1962) 106–154, https://doi.org/10.1113/jphysiol.1962.sp006837.
[18] T.M. Mitchell, Machine Learning, McGraw Hill, New York, 1997.

[6] https://github.com/human-centered-ai-lab/dat-kandinsky-patterns/tree/master/challenge-nr-3.
[7] https://human-centered.ai/kandinksy-challenge.

[19] J.S. Bruner, On attributes and concepts, Chapter 2, in: J.S. Bruner, J.J. Goodnow, G.A. Austin (Eds.), A Study of Thinking, John Wiley and Sons, Inc, 1956, pp. 25–49.
[20] E.B. Hunt, Concept Learning: An Information Processing Problem, Wiley, Hoboken (NJ), 1962.
[21] M.M. Bongard, The Problem of Recognition (in Russian), Nauka, Moscow, 1967.
[22] Goedel Hofstadter, D.R. Escher, Bach: An Eternal Golden Braid, Basic Books, New York, 1979.
[23] P. Dollar, Z. Tu, S. Belongie, Supervised learning of edges and object boundaries, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), IEEE, 2006, pp. 1964–1971.
[24] J.B. Tenenbaum, Bayesian modeling of human concept learning, in: S.A. Solla, T.K. Leen, K.-R. Müller (Eds.), Advances in Neural Information Processing Systems (NIPS 1999), NIPS Foundation, 1999, pp. 59–68.
[25] P.-S. Laplace, Mémoire sur les probabilités, Mém. Acad. R. Sci. Paris 1778 (1781) 227–332, http://www.cs.xu.edu/math/Sources/Laplace/memoir_probabilities.pdf.
[26] X. Ren, J. Malik, Learning a classification model for segmentation, in: Ninth IEEE International Conference on Computer Vision (ICCV), IEEE, 2003, pp. 10–17.
[27] K. Koffka, Principles of Gestalt Psychology, Harcourt, New York, 1935.
[28] M. Wertheimer, Laws of organization in perceptual forms, in: W.D. Ellis (Ed.), A Source Book of Gestalt Psychology, Paul Kegan, London, 1938, pp. 71–88.
[29] D.L. Poole, A.K. Mackworth, R. Goebel, Computational Intelligence: A Logical Approach, Oxford University Press, New York, 1998.
[30] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. Lawrence Zitnick, R. Girshick, CLEVR: a diagnostic dataset for compositional language and elementary visual reasoning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 2901–2910.
[31] K. Yi, C. Gan, Y. Li, P. Kohli, J. Wu, A. Torralba, J.B. Tenenbaum, Clevrer: collision events for video representation and reasoning, arXiv:1910.01442.
[32] D. Bahdanau, H. de Vries, T.J. O'Donnell, S. Murty, P. Beaudoin, Y. Bengio, A. Courville, Closure: assessing systematic generalization of clevr models, arXiv:1912.05783.
[33] R. Vedantam, A. Szlam, M. Nickel, A. Morcos, B. Lake, Curi: a benchmark for productive concept learning under uncertainty, arXiv:2010.02855.
[34] W. Nie, Z. Yu, L. Mao, A.B. Patel, Y. Zhu, A. Anandkumar, Bongard-LOGO: A New Benchmark for Human-Level Concept Learning and Reasoning, arXiv:2010.00763, 2020.
[35] D. Teney, P. Wang, J. Cao, L. Liu, C. Shen, A. van den Hengel, V-PROM: A Benchmark for Visual Reasoning Using Visual Progressive Matrices, arXiv:1907.12271, 2019.
[36] A. Holzinger, A. Saranti, H. Müller, Kandinsky Patterns - an experimental exploration environment for pattern analysis and machine intelligence, arXiv:2103.00519.
[37] A. Santoro, F. Hill, D. Barrett, A. Morcos, T. Lillicrap, Measuring abstract reasoning in neural networks, in: 35th International Conference on Machine Learning, PMLR, 2018, pp. 4477–4486.
[38] J. Raven, The Raven's progressive matrices: change and stability over culture and time, Cogn. Psychol. 41 (1) (2000) 1–48, https://doi.org/10.1006/cogp.1999.0735.
[39] L.A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, T. Darrell, Generating visual explanations, arXiv:1603.08507.
[40] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: visual explanations from deep networks via gradient-based localization, in: ICCV, 2017, pp. 618–626.
[41] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, R. Sayres, Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV), arXiv:1711.11279.
[42] J. Mao, X. Wei, Y. Yang, J. Wang, Z. Huang, A.L. Yuille, Learning like a child: fast novel visual concept learning from sentence descriptions of images, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV 2015), 2015, pp. 2533–2541.