KANDINSKY Patterns as IQ-Test for Machine Learning

Andreas Holzinger, Michael Kickmeier-Rust, Heimo Müller

KANDINSKY Patterns as IQ-Test for machine learning

Andreas Holzinger[1][0000−0002−6786−5194]
Michael Kickmeier-Rust[2] and Heimo Müller[1]

[1] Medical University Graz, Auenbruggerplatz 2, A-8036 Graz, Austria
<andreas.holzinger><heimo.mueller>@medunigraz.at
[2] University of Teacher Education, Notkerstrasse 27, CH-9000 St.Gallen, Switzerland
michael.kickmeier@phsg.ch

Abstract. AI follows the notion of human intelligence which is unfortunately not a clearly defined term. The most common definition, as given by cognitive science as mental capability, includes, among others, the ability to think abstract, to reason, and to solve problems from the real world. A hot topic in current AI/machine learning research is to find out whether and to what extent algorithms are able to learn abstract thinking and reasoning similarly as humans can do – or whether the learning outcome remains on purely statistical correlation. In this paper we provide some background on testing intelligence, report some preliminary results from 271 participants of our online study on explainability, and propose to use our Kandinsky Patterns as an IQ-Test for machines. Kandinsky Patterns are mathematically describable, simple, self-contained hence controllable test data sets for the development, validation and training of explainability in AI. Kandinsky Patterns are at the same time easily distinguishable from human observers. Consequently, controlled patterns can be described by *both* humans and computers. The results of our study show that the majority of explanations was made based on the properties of individual elements in an image (i.e., shape, color, size) and the appearance of individual objects (number). Comparisons of elements (e.g., more, less, bigger, smaller, etc.) were significantly less likely and the location of objects, interestingly, played almost no role in the explanation of the images. The next step is to compare these explanations with machine explanations.

Keywords: Artificial Intelligence, Human Intelligence, Intelligence Testing, IQ-Test, explainable-AI, Interpretable Machine Learning

1 Introduction and Motivation

"If you can't measure it, nor assign it an exact numerical value, nor express it in numbers, then your knowledge is of a meager and unsatisfactory kind"
(attributed to William Thomson (1824–1907), aka Lord Kelvin)

Impressive successes in artificial intelligence (AI) and machine learning (ML) have been achieved in the last two decades, including: 1) IBM Deep Blue [6]

defeating the World Chess Champion Garry Kasparov in 1997, 2) the success of IBM Watson [10] in 2011 in defeating the Jeopardy players Brad Rutter and Ken Jennings, or 3) the sensation of DeepMind's Alpha Go [42] in defeating Go masters Fan Hui in 2015 and Lee Sedol in 2016.

Such successes are often seen as milestones for and "measurements" of AI. We argue that such successes are reached in very specific tasks and not appropriate for evaluating the "intelligence" of machines.

The development of intelligence, therefore, is the result of the incremental interplay between challenge/task, a conceptual change (physiological as well as mentally) of the system, and the assessment of the effects of the conceptual change. To advance AI, specifically in the direction of explainable AI, we suggest bridging the human strength and the human assessment methods with those of AI. In other words, we suggest introducing principles of human intelligence testing as an innovative benchmark for artificial systems.

The ML community is becoming now aware that human IQ-tests are a more robust approach to machine intelligence evaluation than such very specific tasks [9]. In this paper we provide 1) some background on testing intelligence, 2) report on some preliminary results from 271 participants of our online study on explainability [3], and 3) propose to use our Kandinsky Patterns [32] [4] as an IQ-Test for machines.

2 Background

A fundamental problem for AI are often the vague and widely different definitions of the notion of intelligence and this is particularly acute when considering artificial systems which are significantly different to humans [28]. Consequently, intelligence testing for AI in general and ML in particular has generally not been in the focus of extensive research in the AI community. The evaluation of approaches and algorithms primarily occurred along certain benchmarks (cf. [33], [34]).

The most popular approach is the one proposed by Alan Turing in 1950 [45], claiming that an algorithm can be considered intelligent (enough) for a certain kind of tasks if and only if it could finish all the possible tasks of its kind. The shortcoming of this approach, however, is that it is heavily task-centric and that it requires an a-priori knowledge of all possible tasks and the possibility to define these tasks. The latter, in turn, bears the problem of the granularity and precision of definitions. An indicative example is the evaluation, or in other terms, the "intelligence testing" for autonomously driving cars [29], or another example is CAPTCHA (completely automated public Turing test to tell computers and humans apart), which are simple for humans but hard for machines and therefore used for security applications [1]. Such CAPTCHAs use either text or images of different complexity and pose individual differences in cognitive processing [3].

[3] https://human-centered.ai/experiment-exai-patterns
[4] https://human-centered.ai/project/kandinsky-patterns

In cognitive science, the testing of human aptitude – intelligence being a form of cognitive aptitude – has a very long tradition. Basically, the idea of psychological measurement stems from the general developments in 19th century science and particularly physics, which put substantial focus on the accurate measurement of variables.

This view was the beginning of so-called *anthropometry* [36] and subsequently the psychological measurement. The beginning of intelligence testing occurred around 1900 when the French government had passed a law requiring all French children to go to school. Consequently, the government regarded it as important to find a way to identify children who would not be capable to follow school education. Alfred Binet (1857-1911) [11] started the development of assessment questions to identify such children. Remarkably, Binet not only focused on aspects which were explicitly taught in schools but also on more general and perhaps more abstract capabilities, including attention span, memory, and problem solving skills. Binet and his colleagues found out that the childrens capacity to answer the questions and solve the tasks was not necessarily a matter of physical age. Based on this observation, Binet proposed a mental age – which actually was the first intelligence measure [4]. The level of aptitude was seen relative to the average aptitude of the entire population. Charles Spearman (1863-1945) coined in 1904 [43], in this context, the term *g-factor*, a general, higher level of intelligence.

This very early example for an intelligence test already makes the fundamental difference to the task-centric evaluation of later AI very clear. Human intelligence was *not* seen as the capability to solve one particular task, such as a pure classification task, it was considered being a much wider construct. Moreover, human intelligence generally was not measured in an isolated way but always in relation to an underlying population. By the example of the self-driving cars, the question would be whether one car can drive better against all the other cars, or even whether and to what extent the car does better than human drivers. In the 1950s, the American psychologist David Wechsler (1896-1981) extended the ideas of Binet and colleagues and published the *Wechsler Adult Intelligence Scale* (WAIS), which, in its fourth revision, is a quasi standard test battery today [48]. The WAIS-IV contains essentially ten subtests and provides scores in four major areas of intelligence, that is, verbal comprehension, perceptual reasoning, working memory, and processing speed. Moreover, the test provides two broad scores that can be used as a summary of overall intelligence. The overall full-scale intelligence value (IQ was already coined by William Stern in 1912 for the German term Intelligenzquotient) uses the popular mean 100, standard deviation 15 metric.

In advancing Spearmans g-factor idea, Horn and Cattell [17] argued that intelligence is determined by about 100 interplaying factors and proposed two different levels of human intelligence, fluid and crystallized intelligence. The former includes general cognitive abilities such as pattern recognition, abstract reasoning, and problem solving. The latter is based on experience, learning, and acculturation; it includes general knowledge or the use of language. In addition

to Wechslers WAIS-IV, among the most commonly used tests, for example, is *Raven's Progressive Matrices* [37], which is a non-verbal multiple choice measures of the reasoning component of Spearmans g, more exactly, the two components (i) thinking clearly and making sense of complexity, and (ii) the ability to store and reproduce information. The test was originally developed by John Raven in 1936 [37]. The task is to continue a visual pattern (cf. Figure 1). Other tests are the *Reynolds Intellectual Assessment Scales, the Multidimensional Aptitude Battery II, the Naglieri Nonverbal Ability Test* (cf. U[46]), and in German speaking countries the *IST-2000R* [2] or the *Berlin Intelligence Structure Test* (BIS; [20]).

There exists a large amount of classifications and sub-classifications of subfactors of intelligence, The Cattell-Horn [17] classification includes, for example:

- Quantitative knowledge (the ability to understand and work with mathematical concepts)
- Reading and writing
- Comprehension-Knowledge (the ability to understand and produce language)
- Fluid reasoning (incl. inductive and deductive reasoning and reasoning speed)
- Short term memory
- Long term storage and retrieval
- Visual processing (including closure of patterns and rotation of elements)
- Auditory processing (including musical capabilities)
- General processing speed

An - at the first sight similar - classification was introduced by Gardner [12] based on his *theory of multiple intelligences*. As opposed to prior classification, his theory includes a much broader understanding of intelligence as human aptitude. Gardners theory, therefore, was a starting point for an (often discussed as inflationary) increase of types of intelligence, for example in direction of emotional, social, and artistic intelligence [30]. Over the past 120 years, the 20th century ideas of human intelligence have been further developed and new models have been proposed. These new models tend to interpret general intelligence as an emergent construct reflecting the patterns of correlations between different test scores and not as a causal latent variable. The models aim to bridge correlational and experimental psychology and account for interindividual differences in terms of intraindividual psychological processes and, therefore, the approaches look into neuronal correlates of performance [7]. One of these new approaches is, for example, *process overlap theory*, a novel sampling account, based upon cognitive process models, specifically models of working memory [22].

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 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 [40]. On the other hand there are the corresponding human concepts and "contextual understanding" needs effective mapping of them both [24], and is among the future grand goal of human-centered AI [13].

For a detailed description of the KANDINSKY Patterns please refer to [32].

When talking about explainable AI it is important from the very beginning to differentiate between Explainability and Causability: under explainability we understand the property of the AI-system to generate machine explanations, whilst causability is the property of the human to understand the machine explanations [15]. Consequently, the key to effective human-AI interaction is an efficient mapping of explainability with causability. Compared to the map metaphor, this is about establishing connections and relations - not drawing a new map. It is about identifying the *same areas in two completely different maps.*

3 Related Work

Within the machine learning community there is an intensive debate if e.g. neural networks can learn abstract reasoning or whether they merely rely on pure correlation. In a recent paper the authors [41] 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 Ravens Progressive Matrices (RPM) and Mill Hill Vocabulary Scales, which were developed 1936 for use in fundamental research into both the genetic and the environmental determinants of "intelligence" [37]. 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 (see Figure 1). RPMs are strongly diagnostic of abstract verbal, spatial and mathematical reasoning ability. To succeed at the challenge, models must cope with various generalisation 'regimes' in which the training and test data differ in clearly-defined ways.

The amazingly advancing field of AI and ML technologies adds another dimension to the discourse of intelligence testing, that is, the evaluation of artificial intelligence as opposed to human intelligence. Human intelligence tends to focus on adapting to the environment based on various cognitive, neuronal processes. The field of AI, in turn, very much focuses on designing algorithms that can mimic human behavior (weak or narrow AI). This is specifically true in applied genres such as autonomously driving cars, robotics, or games. This also leads to distinct differences in what we consider intelligent. Humans have a consciousness, they can improvise, and the human physiology exhibits plasticity that leads to real learning by altering the brain itself. Although humans tend to make more errors, human intelligence as such is usually more reliable and robust against catastrophic errors, whereas AI is vulnerable against software, hardware and energy failures. Human intelligence develops based on infinite interactions with an infinite environment, while AI is limited to the small world of a particular task.

The development of intelligence, therefore, is the result of the incremental interplay between challenge/task, a conceptual change (physiological as well as mentally) of the system, and the assessment of the effects of the conceptual change. To advance AI, specifically in the direction of explainable AI, we suggest

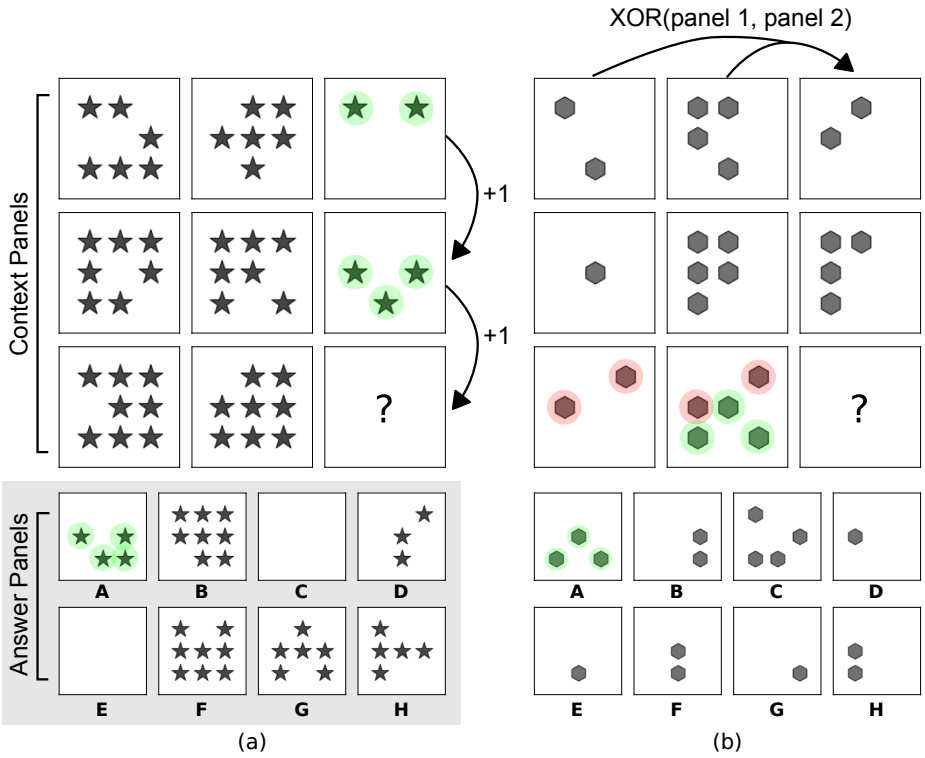

Fig. 1. Raven-style Progressive Matrices. In (a) the underlying abstract rule is an arithmetic progression on the number of shapes along the columns. In (b) there is an `XOR` relation on the shape positions along the rows (panel 3 = `XOR`(panel 1, panel 2)). Other features such as shape type do not factor in. **A** is the correct choice for both, Figure taken from [41].

bridging the human strength and the human assessment methods with those of AI. In other words, we suggest introducing principles of human intelligence testing as an innovative benchmark for artificial systems.

We want to exemplify this idea by the challenge of the identification and interpretation/explanation of visual patterns. In essence, this refers to the human ability to make sense of the world (e.g., by identifying the nature of a series of visual patterns that need to be continued). Sensemaking is an active processing of sensations to achieve an understanding of the outside world and involves the acquisition of information, learning about new domains, solving problems, acquiring situation awareness, and participating in social exchanges of knowledge [35]. The ability can be applied to concrete domains such as various HCI acts [35] but also to abstract domains such as pattern recognition.

This topic was specifically in the focus of medical research. Kundel and Nodine [23], for example, investigated gaze paths in medical images (a sonogram, a tomogram, and two standard radiographic images). They were asked to summa-

rize each of the images in one sentence. The results of this study revealed that correct interpretations of the images were related to attending the relevant areas of the images as opposed to attending visually dominant areas of the images. The authors also found a strong relation of explanations to experiences with images.

A fundamental principle in the perception and interpretation of visual patterns is the likelihood principle, originally formulated by Helmholtz, which states that the preferred perceptual organization of an abstract visual pattern is based on the likelihood of specific objects [27]. A, to a certain degree competing, explanation is the minimum principle, proposed by Gestalt psychology, which claims that humans perceive a visual pattern according the simplest possible interpretation. The role of experience is also reflected in studies in the context of the perception of abstract versus representative visual art; [47] demonstrated distinct differences in art experts and laymen in the perception and their preferences of visual art. Psychological research could demonstrate that the nature of perceiving and interpreting visual patterns, therefore, is a function of expectations [50]. On the one hand, this often leads to misinterpretations or premature interpretations, on the other hand, it increases the explainability of interpretations since the visual perception is determined by existing conceptualizations.

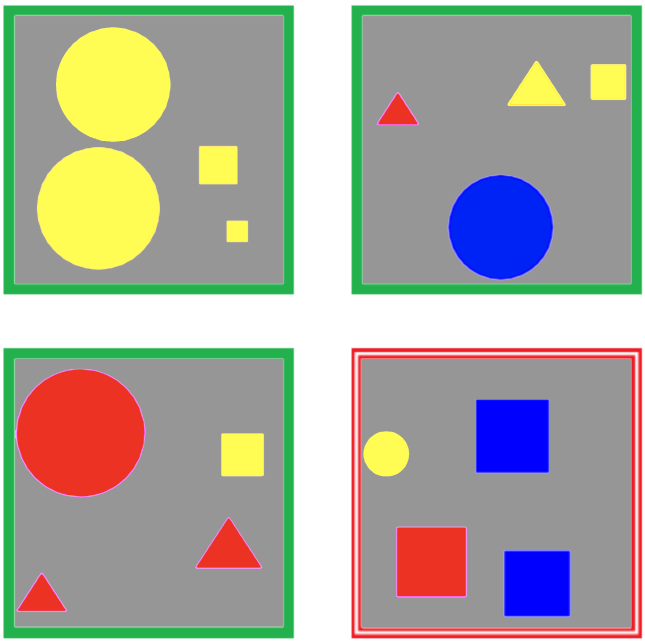

Fig. 2. Visual patterns to be explained by humans.

4 How do humans explain? How do machines explain?

In a recent online study [14], we asked (human) participants to explain random visual patterns (Figure 2). We recorded and classified the free verbal explanations of in total 271 participants. Figure 3 summarizes the results. The results show that the majority of explanations was made based on the properties of individual elements in an image (i.e., shape, color, size) and the appearance of individual objects (number). Comparisons of elements (e.g., more, less, bigger, smaller, etc.) were significantly less likely and the location of objects, interestingly, played almost no role in the explanation of the images.

	Explanatory Element			
	1st	2nd	3rd	Total
Location	0	0	3	3
Number comparisons	1	2	0	3
Color comparisons	9	19	6	34
Size comparisons	16	14	18	48
Shape comparisons	17	13	18	48
Size	32	42	18	92
Color	46	41	45	132
Number	127	124	116	367
Shape	133	129	91	353

Fig. 3. Visual patterns to be explained by humans.

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*, see Figure 4 (A) for color and (B) for shapes.
- 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"* , see Figure 4 (C).
- 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"*, see Figure 4 (D).
- Gestalt concepts (see below) e.g. *closure*, *symmetry*, *continuity*, *proximity*, *similarity*, e.g. *"in a Kandinsky Figure objects are grouped in a circular manner"*, see Figure 4 (E).
- Domain concepts, e.g. *"a group of objects is perceived as a "flower""*, , see Figure 4 (F).

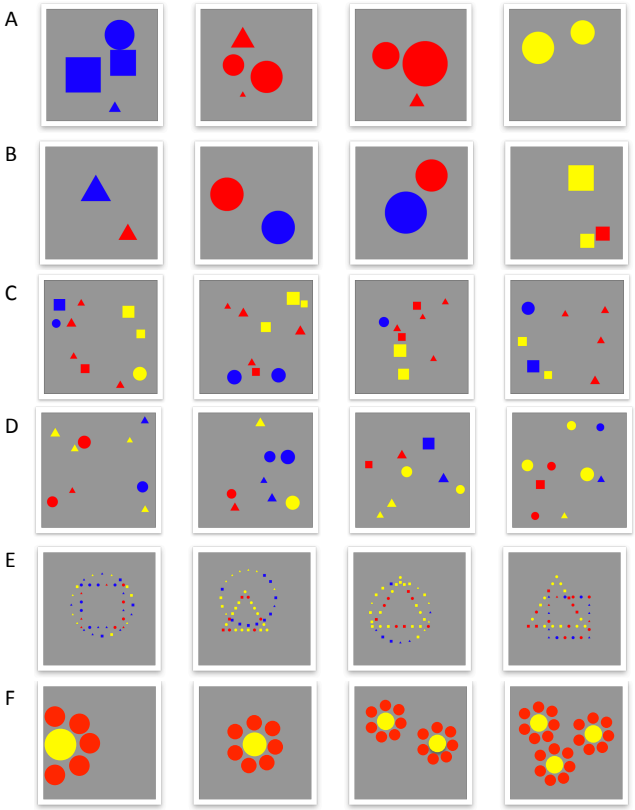

Fig. 4. Kandinsky Pattern showing concepts as color (A), shape (B), numeric relations (C), spatial relations (D), Gestalt concepts (E) and domain concepts (F)

These basic concepts can be used to select groups of objects, e.g. 'all red circles in the upper left corner', and to further combine single objects and groups in a statement with logic operator, e.g. 'if there is a red circle in the upper left corner, there exists no blue object', or with complex domain specific rules, e.g. 'if the size of a red circle is smaller then the size of a yellow circle, red circles are arranged circular around yellow circles'.

In their experiments [18] 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 back-propagation 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 [26]. 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 [31] in other words it is training an algorithm to distinguish between examples and non-examples (we call the latter counterfactuals).

Concept learning has been a relevant research area in machine learning for a long time and had it origins in cognitive science, defined as search for attributes which can be used to distinguish exemplars from non exemplars of various categories [5]. 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 [19].

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 a 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 else-where. Given an image, obtaining an optimal or even likely interpretation W, or associated SW, can be difficult. For example, in edge detection previous work [8] asked what is the probability of a given location in a given image belonging to the component of interest.

[44] presented a model of concept learning that is both computationally grounded and able to fit to human behaviour. 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 so-called 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

[39]. Gestalt-principles [21] can be seen as rules, i.e. they discriminate competing segmentations only when everything else is equal, therefore we speak more generally as Gestalt-laws and one particular group of Gestalt-laws are the Gestalt-laws of grouping, called Prägnanz [49], 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.

Unfortunately, the currently best performing machine learning methods have a number of disadvantages, and one is of particular relevance: Neural networks ("deep learning") are difficult to interpret due to their complexity and are therefore considered as "black-box" models [16]. 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). This makes their inner workings difficult to interpret and understand. However, the "why" would often be much more useful than the simple classification result.

5 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 grand general question is "How can we perform a task by exploiting knowledge extracted during solving previous tasks?" To answer such questions 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 many comments and results. Updated information can be found at the accompanying Web page[5]. A single Kandinsky pattern may serve as an "intelligence (IQ) test" for an AI system. To make the step towards a more human-like and probably in-depth assessment of an AI system, we propose to apply the principles of human intelligence tests, as outlined in this paper. In relation to the Kandinsky patterns we suggest applying the principle of Raven's progressive matrices. This test is strongly related to the identification of a "meaning" in the complex visual patterns [38]. The underlying complex pattern, however, is not based on a single image, the meaning only arises from the sequential combination of multiple images. To assess AI, a set of Kandinsky patterns, each of which complex in itself, can be used. A "real" intelligent achievement would be identifying the concepts - and therefore the meaning ! - of sequences of multiple Kandinsky patterns. At the same time, the approach solves one key problem of testing "strong AI", the language component. With this approach it is not necessary to verbalize the insights of the AI system. Per definition, the identification of the right visual pattern that "traverses" the Kandinsky patterns (analogous to Raven's

[5] `https://human-centered.ai/kandinksy-challenge`

matrices) indicates the identification of an underlying meaning. Much further experimental and theoretical work is needed here.

6 Acknowledgements

We are grateful for interesting discussions with our local and international colleagues and their encouragement. Parts of this project have been funded by the EU projects FeatureCloud, EOSC-Life, EJP-RD and the Austrian FWF Project "explainable AI", Grant No. P-32554.

References

1. von Ahn, L., Blum, M., Hopper, N.J., Langford, J.: Captcha: Using hard ai problems for security. In: Bilham, E. (ed.) Advances in Cryptology-Eurocrypt 2003, pp. 294–311. Springer-Verlag Berlin (2003). https://doi.org/10.1007/3-540-39200-9-18
2. Amthauer, R.: Intelligenz-Struktur-Test 2000 R: I-S-T 2000 R Manual (2. Auflage). Hogrefe, Gttingen (2001)
3. Belk, M., Germanakos, P., Fidas, C., Holzinger, A., Samaras, G.: Towards the personalization of captcha mechanisms based on individual differences in cognitive processing. In: Human Factors in Computing and Informatics, Lecture Notes in Computer Science, LNCS 7946, pp. 409–426. Springer (2013). https://doi.org/10.1007/978-3-642-39062-3
4. Binet, A.: L' étude expérimentale de l'intelligence. Schleicher fréres & cie, Paris (1903)
5. Bruner, J.S.: Chapter 2: On attributes and concepts. In: Bruner, J.S., Goodnow, J.J., Austin, G.A. (eds.) A study of thinking, pp. 25–49. John Wiley and Sons, Inc (1956)
6. Campbell, M., Hoane Jr, A.J., Hsu, F.h.: Deep blue. Artificial intelligence **134**(1-2), 57–83 (2002). https://doi.org/10.1016/S0004-3702(01)00129-1
7. Conway, A., Kovacs, K.: New and emerging models of human intelligence. Cognitive Science (2015). https://doi.org/10.1002/wcs.1356
8. Dollar, P., Tu, Z., Belongie, S.: Supervised learning of edges and object boundaries. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). pp. 1964–1971. IEEE (2006). https://doi.org/10.1109/CVPR.2006.298
9. Dowe, D.L., Hernndez-Orallo, J.: Iq tests are not for machines, yet. Intelligence **40**(2), 77–81 (2012). https://doi.org/10.1016/j.intell.2011.12.001
10. Ferrucci, D., Levas, A., Bagchi, S., Gondek, D., Mueller, E.T.: Watson: beyond jeopardy! Artificial Intelligence **199**, 93–105 (2013). https://doi.org/10.1016/j.artint.2012.06.009
11. Funke, J.: Handbook of Anthropometry. Springer, Berlin (2006)
12. Gardner, H.: Changing minds. The art and science of changing our own and other people's minds. Harvard Business School Press, Boston, MA (2004)
13. Holzinger, A., Biemann, C., Pattichis, C.S., Kell, D.B.: What do we need to build explainable ai systems for the medical domain? arXiv:1712.09923 (2017)
14. Holzinger, A., Kickmeier-Rust, M.D., Müller, H.: Human explanation profiles for random visual patterns as a benchmark for explainable ai. in preparation (2019)

15. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H.: Causability and explainability of ai in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery (2019). https://doi.org/10.1002/widm.1312
16. Holzinger, A., Plass, M., Holzinger, K., Crisan, G.C., Pintea, C.M., Palade, V.: A glass-box interactive machine learning approach for solving np-hard problems with the human-in-the-loop. arXiv:1708.01104 (2017)
17. Horn, J.L., Cattell, R.B.: Refinement and test of the theory of fluid and crystallized general intelligences. Journal of Educational Psychology **57**, 253–270 (1966)
18. Hubel, D.H., Wiesel, T.N.: Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of Physiology **160**(1), 106–154 (1962). https://doi.org/10.1113/jphysiol.1962.sp006837
19. Hunt, E.B.: Concept learning: An information processing problem. Wiley, Hoboken (NJ) (1962). https://doi.org/http://dx.doi.org/10.1037/13135-001
20. Jger, A.O.: Validitt von intelligenztests. Diagnostica **32**, 272–289 (1986)
21. Koffka, K.: Principles of Gestalt Psychology. Harcourt, New York (1935)
22. Kovacs, K., Conway, A.: Process overlap theory: a unified account of human intelligence. Psychological Inquiry **27**, 151–177 (2016)
23. Kundel, H.I., Nodine, C.F.: A visual concept shapes image perception. Radiology **146**(2) (1983)
24. Lake, B.M., Salakhutdinov, R., Tenenbaum, J.B.: Human-level concept learning through probabilistic program induction. Science **350**(6266), 1332–1338 (2015). https://doi.org/10.1126/science.aab3050
25. Laplace, P.S.: Mémoire sur les probabilités. Mémoires de lAcadémie Royale des sciences de Paris **1778**, 227–332 (1781)
26. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
27. Leeuwenberg, E.L., Boselie, F.: Against the likelihood principle in visual form perception. Psychological Review **95**(4), 485–491 (1988)
28. Legg, S., Hutter, M.: Universal intelligence: A definition of machine intelligence. Minds and machines **17**(4), 391–444 (2007). https://doi.org/10.1007/s11023-007-9079-x
29. Li, L., Wang, X., Wang, K., Lin, Y., Xin, J., Chen, L., Xu, L., Tian, B., Ai, Y., Wang, J.: Parallel testing of vehicle intelligence via virtual-real interaction. Science Robotics **4**(eaaw4106), 1–3 (2019)
30. Locke, E.A.: Why emotional intelligence is an invalid concept. Journal of Organizational Behavior **26**, 425–431 (2005)
31. Mitchell, T.M.: Machine learning. McGraw Hill, New York (1997)
32. Müller, H., Holzinger, A.: Kandinsky patterns. arXiv:1906.00657 (2019), `https://arxiv.org/abs/1906.00657`
33. Nambiar, R.: Towards an industry standard for benchmarking artificial intelligence systems. In: 34th International Conference on Data Engineering (ICDE 2018). p. Pages. IEEE (2018). https://doi.org/10.1109/ICDE.2018.00212
34. Nambiar, R., Ghandeharizadeh, S., Little, G., Boden, C., Dholakia, A.: Industry panel on defining industry standards for benchmarking artificial intelligence. In: Performance Evaluation and Benchmarking for the Era of Artificial Intelligence. p. Pages. Springer International Publishing (2019). https://doi.org/10.1007/978-3-030-11404-6-1
35. Pirolli, P., Russell, D.M.: Introduction to this special issue on sensemaking. Human-Computer Interaction **26**, 1–8 (2011)
36. Preed, V.R.: Handbook of Anthropometry. Springer, Berlin (2012)

37. Raven, J.: The raven's progressive matrices: change and stability over culture and time. Cognitive psychology **41**(1), 1–48 (2000). https://doi.org/10.1006/cogp.1999.0735
38. Raven, J.C., Court, J.H.: Raven's progressive matrices and vocabulary scales. Oxford Psychologists Press, Oxford, New York (1998)
39. Ren, X., Malik, J.: Learning a classification model for segmentation. In: Ninth IEEE International Conference on Computer Vision (ICCV). pp. 10–17. IEEE (2003). https://doi.org/10.1109/ICCV.2003.1238308
40. Samek, W., Wiegand, T., Müller, K.R.: Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv:1708.08296 (2017)
41. Santoro, A., Hill, F., Barrett, D., Morcos, A., Lillicrap, T.: Measuring abstract reasoning in neural networks. In: 35th International Conference on Machine Learning. pp. 4477–4486. PMLR (2018)
42. Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., Hassabis, D.: Mastering the game of go with deep neural networks and tree search. Nature **529**(7587), 484–489 (2016). https://doi.org/10.1038/nature16961
43. Spearman, C.: General intelligence," objectively determined and measured. The American Journal of Psychology **15**(2), 201–292 (1904)
44. Tenenbaum, J.B.: Bayesian modeling of human concept learning. In: Solla, S.A., Leen, T.K., Müller, K.R. (eds.) Advances in neural information processing systems (NIPS 1999). pp. 59–68. NIPS foundation (1999)
45. Turing, A.M.: Computing machinery and intelligence. Mind **59**(236), 433–460 (1950). https://doi.org/10.1093/mind/LIX.236.433
46. Urbina, S.: Chapter 2: Tests of intelligence. In: Kaufman, S.B. (ed.) The Cambridge Handbook of Intelligence. Cambridge University Press, Cambridge (2011)
47. Uusitalo, L., Simola, J., Kuisma, J.: Perception of abstract and representative visual art. tba (01 2009)
48. Wechsler, D.: The Measurement and Appraisal of Adult Intelligence (4th Edition). Williams and Witkins, Baltimore, MD (1958)
49. Wertheimer, M.: Laws of organization in perceptual forms. In: Ellis, W.D. (ed.) A source book of Gestalt psychology, pp. 71–88. Paul Kegan, London (1938). https://doi.org/10.1037/11496-005
50. Yanagisawa, H.: How does expectation affect sensory experience? a theory of relativity in perception. In: 5th International Symposium on Affective Science and Engineering ISASE 2019. pp. 1–4. Japan Society of Kansei Engineering (2019). https://doi.org/10.5057/isase.2019-C000014