Utopian or Dystopian?: using a ML-assisted image generation game to empower the general public to envision the future

JANET F. RAFNER
Aarhus University, Denmark
janetrafner@mgmt.au.dk

LOTTE PHILIPSEN
Aarhus University, Denmark
lottephilipsen@cc.au.dk

STEVEN LANGSFORD
Aarhus University, Denmark
slangsford@mgmt.au.dk

SEBASTIAN RISI
ITU Copenhagen, Denmark
sebr@itu.dk

ARTHUR HJORTH
Aarhus University, Denmark
arthur@phys.au.dk

JOEL SIMON
Morphogen
joelsimon6@gmail.com

MIROSLAV GAJDACZ
Aarhus University, Denmark
gajdacz@mgmt.au.dk

JACOB SHERSON[1]
Aarhus University, Denmark
sherson@mgmt.au.dk

The rise of digital technologies and Machine Learning (ML)-tools for creative expression brings about novel opportunities for studying creativity and cognition at scale. In this paper, we present a pilot study of crea.blender SDG - an online GAN based image generation game. We designed crea.blender SDG with two goals in mind: The first, to let people create images relating to the United Nations Sustainable Development Goals (SDGs) and through them, engage in large-scale conversations on complex socioscientific problems. The second, as a fun and inspiring gateway for public participation in research, generating data for the creativity and cognition research and design community. Specifically in this pilot, we study and affirm that the design of crea.blender SDG is flexible enough to allow users to create images that express both anxiety and hope for the future; affirm that user generated images express these ideas in ways that are meaningful to people other than the original creator; and begin to investigate which specific features of images are more closely related to dystopian or utopian ideas of the future. Finally, we discuss implications for future design and research with ML-based creativity tools.

CCS CONCEPTS • General reference → Cross-computing tools and techniques → Design • Human-centered-computing → Collaborative and social computing → Collaborative and social computing devices• Applied computing → Arts and humanities → Media arts

Additional Keywords and Phrases: GAN, crowdsourcing, creativity, sustainability, SDGs

---

1 Corresponding author

1 INTRODUCTION

Artistic expression from both professional artists and the general public is a key method for raising awareness of and facilitating discussions around the Sustainable Development Goals (SDGs) [31, 27, 32, 33, 35]. These 17 goals were set into place by the United Nations in 2015 as a blueprint to achieve a better and more sustainable future for all [37]. Generative Adversarial Networks (GANs) [7] are Machine Learning (ML) models that can be used as creative tools by both professional and non-professional users in a variety of contexts. Professional artists use GANs as a new medium [8, 13, 14, 11], as a means of inspiring new ideas [8, 4, 15, 9, 10, 21] or as a helping hand for laborious tasks [6, 16, 17, 19, 20, 22]. By lowering the threshold for technical skills or capabilities, some GANs also allow 'non-artists' or those with physical impediments to more easily express themselves creatively [1, 18, 2, 23]. GAN images are ideal for envisioning the future as they are genuine phantom images – a AI + human imagination of what images of the world could look like.

Recently, platforms like Artbreeder have successfully given the general public access to collaborative image generation [30]. Artbreeder is a massive online tool for creating images based on interactive latent variable evolution [3]. Images are 'bred' by selecting the generated offspring of parent images in addition to direct 'gene' editing. Artbreeder operates as a hybrid of a tool and a social network, allowing users to share what they create and edit what they see. This community driven innovation allows certain images to go viral, spreading their 'DNA' throughout the image repository. As a recent adaptation of Artbreeder, crea.blender [12] supports systematic, quantitative investigation of creativity by letting players "blend" a restricted and carefully curated set of background free, hard coded source images into new images. A pilot study found that crea.blender provides a playful experience, affords players a sense of control over the interface, and elicits different types of player behavior. More generally, the study indicated the potential of ML-assisted image generation for use in a scalable, playful, creativity assessment [12]. Building on the work of [12, 30], we here present crea.bender SDG. This game retains the structured, goal-oriented setting of crea.blender, this time focusing on creating utopian and dystopian images of the future, but extends the open-ended creative potential by using the Landscape GAN of Artbreeder and allowing for free substitution of source images.

Collaborative image generation and reflections is only possible if the interface affords users enough expressivity that they can generate images that are not simply idiosyncratically labeled, but that are recognizable by others as being either utopian or dystopian. Previous work [12], demonstrated that the crea.blender interface afforded the necessary control of the GAN for *deliberate* expression. However, it did not test whether people felt they could express specific ideas, thoughts, or concepts, or *how* this would happen. Addressing this is the primary purpose of this study, and we ask the following research questions:

1. Does crea.blender SDG offer enough flexibility to produce a variety of both utopian and dystopian images?
2. Are these images recognizable as being utopian or dystopian?
3. Which particular components of the images are recognizable as utopian and dystopian?

2 DESIGN OF CREA.BLENDER SDG

Crea.blender SDG allows users to blend two different components, 'style' and 'content' of a set of source images into new images. This is done using the generator of a pre-trained StyleGAN2 [36], trained to produce images based on style and content. After completing a tutorial, users blend images by changing the slider values for the style and content of each of the images (Fig 1, left), taking content (large scale features) from some images, and blending them with the style (small scale features and texture) of other images. Users are presented with 4 source images that can be freely 'swapped' with alternative images. When users save an image, they label it as being either utopian, or dystopian. The image and the label are then stored in a database. There is also a publicly available gallery of all generated images (Fig 1, right). Users can download their own saved images and any images in the public gallery.

In crea.blender SDG there are 102 source images, selected from the landscape category in Artbreeder. We deliberately chose source images with as great a variety as possible in terms of different motivic elements (water, buildings, mountains, etc.) as well as different colors. We used the landscape GAN to avoid portraits and people due to concerns about relating utopian/dystopian ratings to people's outer appearance [5].

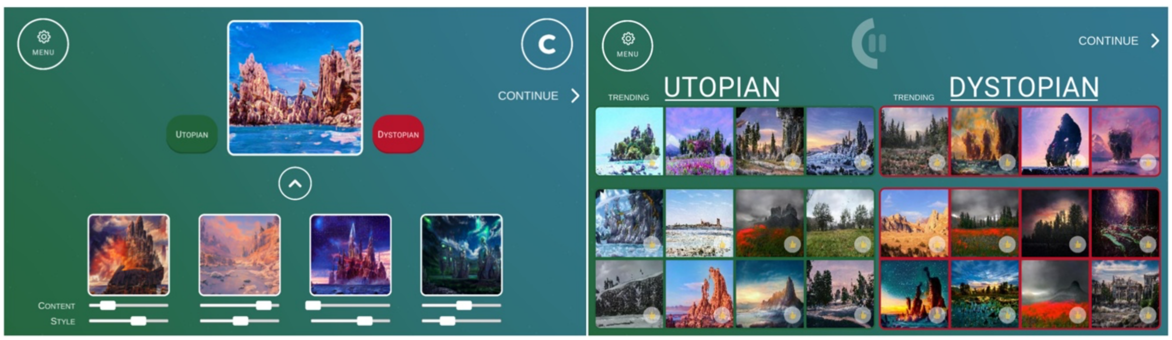

Figure 1: *Left:* Task screenshot. *Right:* Publicly available gallery of images.

3 EXPERIMENTAL STUDY: TESTING RECOGNIZABILITY AND FEATURES

In December 2020 we launched the initial version of crea.blender SDG in conjunction with Artbreeder and the United Nations platform AI4Good [26, 39]. Between December 2020 and March 2021 there were 580 user sessions, producing 8475 images of which 1624 were saved to the utopian/dystopian galleries and 196 downloaded by users. To test whether created images were recognizable as a utopia or a dystopia we collected classification judgments. We examined the degree of agreement between raters and the degree of agreement with the original creator's label. This is critical as appropriateness as well as novelty [40] are key components of creativity and thus the generated images must be at a minimum recognizable as appropriate to the prompt. We also asked raters to give an indication for each image of which aspects were most important in characterizing it as utopian or dystopian. A total of 24 raters were recruited, but only data from eight raters who rated more than 100 images was considered for analysis. Rater agreement statistics were calculated only for the three raters who rated all images. All included raters self-reported normal vision. Raters were familiar with the project, but had not seen these images before, and were blind to each other's ratings and to the label given by the original user.

Raters were presented with one image at a time via a web app that prompted them to label each image as 'utopia' or 'dystopia' - described in the instructions as "a future paradise where society is doing great" and "a nightmarish future where society has gone horribly wrong" respectively. Raters were asked to "pick the category the image is most similar to, even if it's not a particularly good example". For each response, raters were asked if the main factor in each decision was the colors in the image, the style of the image (described as "level of details, realism or abstraction, or other stylistic aspects such as blurry or sharp and well-defined shapes"), or a motif in the image (described as "a specific identifiable element such as trees, a river, or a building"). These factors were chosen because: First, the two sliders used when generating the images were assigned to content and style (covering colors) respectively. Second, whereas color can be completely abstract – thus not referring to specific meaning bearing objects – motif is always inscribed in the rater's existing cultural frame of reference (otherwise it would not be considered a motif). Hence, color, style and motif provided the raters the choice of taxonomically different motivations. A set of 1516 images were available for rating, 620 labeled dystopian and 860 utopian by their original creators, with 36 unclassified due to a technical issue. Order of presentation was randomized for each rater.

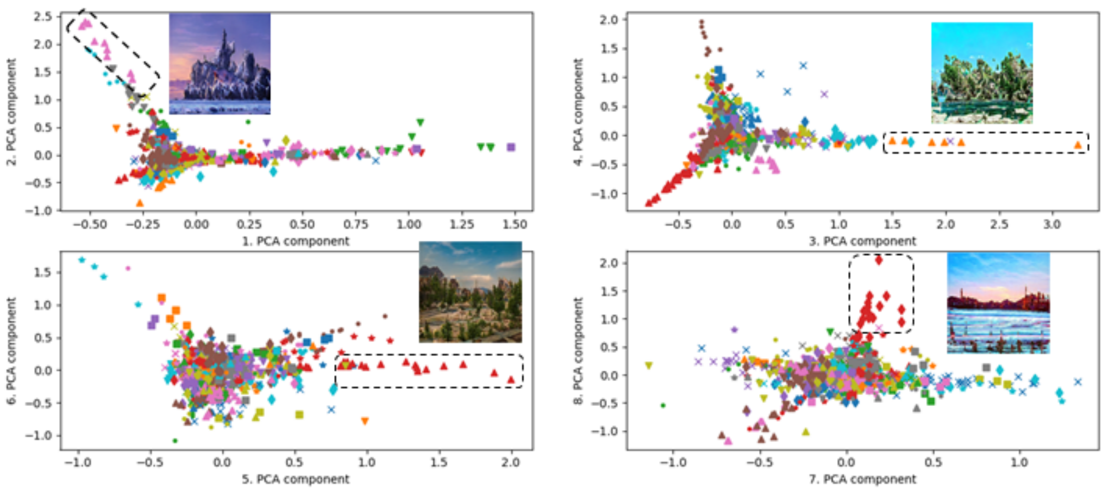

Figure 2: Representation of 1117 images after k-means clustering in the Principal Component Analysis basis (first 8 components). Shown are 77 non-trivial clusters (each identified by a combination of color and symbol). The insets show examples of images from the clusters marked by the dashed outline.

4 ANALYSIS

4.1 Does crea.blender SDG offer enough flexibility to produce both utopian and dystopian images?

It is relevant to understand the total number of distinct images that could be produced in principle by the full input space of the GAN. Images were produced by blending up to 4 images chosen from a pool of 102 source images, which gives about $4*10^6$ different starting sets of source images. The sliders were quasi-continuous, but if we conservatively assume that only 10 different values for each slider give reasonably distinguishable changes in the resulting image, the lower bound is about 10^8 unique slider settings per image set and $4*10^{14}$ possible output images. To quantify the extent to which people *actually* explored this space, we look at both the diversity of the 4 source images chosen, and the range of slider values used for them. Frequency of selecting a source image was over-dispersed by factor of two relative to the binomial distribution expected under chance, indicating that some source images were more popular than others. However, each source image was used in at least 23 generated images and no source image was used in more than 99. Style and content slider values were correlated at r=.49 (i.e. somewhat reducing the effective size of the space explored), but there was no association between source image and slider value.

Although these results suggest saved images came from diverse regions of the space, it is important to check for the appearance of possible clusters of similar images as people might have been attracted to particular images. To test this, we clustered the images using the k-means algorithm with 100 clusters (30 repetitions) in the 2*102 dimensional space corresponding to the two slider values for each possible source image (see Fig. 2). There is one large cluster with about 195 images, and the remaining clusters have a mean image count of 13±8. There are about 24 clusters generated by single users which contain on average 5 images. 15 of those clusters contain purely utopian or dystopian images (by creator label). There are about 59 clusters with images from at least 5 users. The fraction of utopian images in those clusters correlates at r = 0.61 with the mean utopian fraction assigned by the 8 main raters (part of RQ2). We conclude that people created images using a wide range of source images and slider values, and that while it is possible to identify some clusters of similar generated images users did not converged on a small number of prototypical utopias or dystopias.

4.2 RQ2: Are images recognizable by others as utopian or dystopian?

We assessed agreement between raters, and between raters and creators, using Fleiss' Kappa [38]. The three raters who rated all images agreed with each other with kappa of 0.253, (z = 16.8, p<.001). Raw proportion of matching ratings between all pairs of raters ranged between 0.44 and 0.83. Roughly a third of the images (539) received unanimous ratings, a subsample of these can be seen in Fig 3.

Figure 3: Example utopias (left) and dystopias (right) with unanimous ratings, including creator intention. These are a subset of all 539 images with unanimous ratings, representing the most recognizable examples of utopias and dystopias created by users of the system.

Including creator intention as another 'rater' gave a Fleiss kappa of 0.2 (z=18.9, p<.001). The raw proportion of ratings that matched creator intention ranged between 0.52 and 0.65 over the eight raters. For 367 of the images all raters and the creator agreed unanimously on the image classification. The average number of raters per image was 4.3. Even conservatively assuming only 3 raters, pure chance would have led to roughly $2*(1/2)^{(3+1)}*1516=189$ unanimous images. Despite the somewhat low average rater agreement, this estimate as well as Fig 3 clearly demonstrate that there exist subsets of the created images for which the categorization (as a utopian or dystopian set) is clear.

4.3 RQ3: Which particular components of the images are recognizable as utopian and dystopian?

We collected preliminary data about the features raters attended to with the three-option 'reason' question following each image classification. The proportions of each reason given for the ratings were broadly similar across utopia and dystopia decisions: overall ratings, 50% were reported as motivated by color, 29% by motif, and 21% by style. Agreement on reason decisions gave a Fleiss kappa 0.12 ($z=11$, $p<.001$).

The categorization of utopias and dystopias cuts across the physical features of the landscapes as shown by the high variation in the selection of source images and their blend proportions. Despite the apparent lack of a set of necessary and sufficient criteria for making these categorization decisions, we found that blind ratings of category membership agreed with each other and with creator intention at levels above chance. It is not clear what features creators are attending to when deciding which images are distinctively utopian or dystopian. Our results suggest that color might be particularly salient, but also that rater agreement on the reasons for a particular classification is low. The fact that raters were more motivated by color (50%) than by motif (29%) could be caused by the fact that all source images were selected from enquiry into ‘landscape’ on Artbreeder, which may have led to a reduced spread in perceived motifs. Alternatively, no participants had training in image analysis, and color may simply be more salient to the novice eye than other features

5 CONCLUSION, LIMITATIONS, AND FUTURE WORK

The current study affirms the potential of human-GAN interactions with a suitably designed interface to afford the expression of recognizable ideas, thoughts, or concepts. Given the immense expressive space of GANs, the range of interpretation of resulting images is expected. Consequently, we did not expect that *every* user would be able to generate images that evoke a common interpretation, but that *many* would. The significant overrepresentation of unanimous ratings in our pool of images is an early indication that this is in fact the case. The small number of raters, and consequent small number of ratings per image, limit conclusions about the status of individual images or systematic rater heterogeneity. Our raters were WEIRD, and must be assumed to share cultural references, color and motif preferences, etc. Future work with a larger number of raters could potentially examine systematic differences among individual images or creators at a finer grain. Further, we expect there to be national and cultural differences in dystopian and utopian aesthetics. This will be another subject of future research. A tool like this will make it possible to study how people think about creative expression, and provide data for studying at scale and in depth what kinds of features of images people attend to when creating or interpreting images, and how these features connect to users’ underlying concepts. These high ambitions can only be achieved with a detailed understanding of both the optimal technological support as well as the thought processes in both creators and raters and our quantitative analysis is therefore a crucial first step in this direction. The aim is to further develop a crea.blender SDG version for use in conjunction with major events such as Global Talent Summit, The World Economic Forum, and the G20 [28, 34, 29].

ACKNOWLEDGEMENTS

We thank Neil Sahota for making the collaboration with AI4Good possible. We also thank Carlsberg Foundation, Novo Nordisk Foundation and Synakos Foundation for their support of this research.

REFERENCES

[1] Guillermo Bernal, Lily Zhou, Haripriya Mehta, and Pattie Maes. 2019. Paper Dreams: An Interactive Interface for Generative Visual Expression

[2] Frederik De Bleser. 2019. GANDelve - a Visual Interface for Creative AI. In Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)

[3] Philip Bontrager, Wending Lin, Julian Togelius, and Sebastian Risi. 2018. Deep interactive evolution. In International Conference on Computational Intelligence in Music, Sound, Art and Design. Springer, 267-282.

[4] Vivien Cabannes, Thomas Kerdreux, Louis Thiry, Tina Campana, and Charly Ferrandes. 2019. Dialog on a canvas with a machine. arXiv preprint arXiv:1910.04386

[5] Kate Crawford and Trevor Paglen. 2019. Excavating AI: The Politics of Training Sets for Machine Learning. https://www.excavating.ai/

[6] Kevin Frans. 2017. Outline colorization through tandem adversarial networks. arXiv preprint arXiv:1704.08834

[7] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, DavidWarde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014.Generative adversarial networks In Advances in neural information processing systems (pp. 2672-2680).

[8] Aaron Hertzmann. 2020. Visual indeterminacy in GAN art. Leonardo 53, 4(2020), 424–428.

[9] Nikolay Jetchev, Urs Bergmann, and Gökhan Yildirim. 2019. Transform the Set: Memory Attentive Generation of Guided and Unguided Image Collages. arXiv preprint arXiv:1910.07236

[10] Koga Tatsuki, Rie Ema, Hirose Kazuko, and Seita Jun. 2019. Human and GAN collaboration to create haute couture dress In Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS) (Vancouver, Canada) (Neurips).

[11] Reiichiro Nakano. 2019. Neural painters: A learned differentiable constraint for generating brushstroke paintings. arXiv preprint arXiv:1904.08410

[12] Rafner, Janet, Arthur Hjorth, Sebastian Risi, Lotte Philipsen, Charles Dumas, Michael Mose Biskjær, Lior Noy et al. 2020. crea. blender: A Neural Network-Based Image Generation Game to Assess Creativity. In Extended Abstracts of the 2020 Annual Symposium on Computer-Human Interaction in Play, pp. 340-344.

[13] Tim Schneider and Naomi Rea. 2018. Has artificial intelligence given us the next great art movement? Experts say slow down, the 'field is in its infancy.'. Artnet News. Retrieved https://news.artnet.com/art-world/ai-art-comes-to-market-is-it-worth-the-hype-1352011 March 31, 2021

[14] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training GANs. arXiv preprint arXiv:1606.03498

[15] Othman Sbai, Mohamed Elhoseiny, Antoine Bordes, Yann LeCun, and Camille Couprie. 2018. Design: Design inspiration from generative networks. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops. 0–0

[16] Lingzhi Zhang, Tarmily Wen, and Jianbo Shi. 2020. Deep image blending. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 231–240.

[17] Fangneng Zhan, Hongyuan Zhu, and Shijian Lu. 2019. Spatial fusion gan for image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 3653–3662.

[18] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. 2019. GP-GAN: Towards realistic high-resolution image blending. In Proceedings of the 27th ACM international conference on multimedia. 2487–2495

[19] Brian Quanz, Wei Sun, Ajay Deshpande, Dhruv Shah, and Jae-eun Park. 2020. Machine learning based co-creative design framework. arXiv preprint arXiv:2001.08791

[20] Guido Salimbeni, Frederic Fol Leymarie, and William Latham. 2019. Generative system to assist the artist in the choice of 3D composition for a still life painting. In Machine Learning for Creativity and Design (NeurIPS 2019 Workshop).

[21] Philipp Schmitt and Steffen Weiß. 2018. The Chair Project: A Case-Study for using Generative Machine Learning as Automatism. In Proceedings of the 32nd Conference on Neural Information Processing Systems

[22] Gao Zhengyan, Yonetsuji Taizan, Takamura Tatsuya, Matsuoka Toru, and Naradowsky Jason. 2018. Automatic Illumination Effects for 2D Characters. In Proceedings of the 32nd Conference on Neural Information Processing Systems

[23] Le Zhou, Qiu-Feng Wang, Kaizhu Huang, and Cheng-Hung Lo. 2019. An interactive and generative approach for Chinese shanshui painting document. In 2019 International Conference on Document Analysis and Recognition (ICDAR). IEEE, 819–824

[24] Andrew Brock, Jeff Donahue, and Karen Simonyan. 2018. Large scale GAN training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096

[25] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition. IEEE, 248–255

[26] <NGO>. About Us. 2021 Retrieved March 2021 from <NGO website>

[27] Francyne Harrigan and Valentina Giani. 2016. Design leaders launch collaborative platform to support SDGs - United Nations Sustainable Development. Retrived March 2021 from https://www.un.org/sustainabledevelopment/blog/2016/11/design-leaders-launch-collaborative-platform-to-support-sdgs/

[28] GTS. 2020. Global Talent Summit 2020. Retrieved March 31, 2021 from www.globaltalentsummit.org

[29] G20. 1999. G20. Retrieved March, 2021 from https://www.g20.org

[30] Joel Simon. 2020. Artbreeder. Retrieved March, 2021 from https://artbreeder.com/browse

[31] Geneve International. 2020. Revisiting the Sustainable Development Goals through Art. Retrieved March, 2021 from www.geneve-int.ch/revisiting-sustainable-development-goals-through-art

[32] United Nations. 2016. Tokyo art scene collaborates for SDGs – United Nations Sustainable Development. Retrieved March, 2021 from https://www.un.org/sustainabledevelopment/blog/2016/07/tokyo-art-scene-collaborates-for-sdgs/

[33] SDGAction30028. 2018. Translating the 17 icons of the SDGs into a famous local art form in Nepal as a part of UN's bigger goal to localize the SDGs in Nepal's seven provinces. Retrieved March, 2021 from https://sustainabledevelopment.un.org/partnership/?p=30028

[34] Davos. 2020. World Economic Forum Annual Meeting Davos. Retrieved March, 2021 from www.weforum.org/events/world-economic-forum-annual-meeting-2020

[35] Art competition for the global goals report for young people across UK. 2019 Retrieved March, 2021 from https://oppourtunities.com/2019-art-competition-for-the-global-goals-report-for-young-people-across-uk/

[36] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 8110–8119

[37] United Nations 2017 Resolution adopted by the General Assembly on 6 July 2017, Work of the Statistical Commission pertaining to the 2030 Agenda for Sustainable Development A/RES/71/313 Archived 28 November 2020 at the Wayback Machine, http://ggim.un.org/documents/a_res_71_313.pdf

[38] Fleiss, Joseph L. 1971 Measuring nominal scale agreement among many raters. Psychological bulletin, 76(5), 378

[39] AIforgood 2021. Retrieved April, 2021 from aiforgood.itu.int

[40] Mark A. Runco and Garrett J. Jaeger. 2012. The Standard Definition of Creativity, Creativity Research Journal, 24:1, 92-96, DOI: 10.1080/10400419.2012.650092