The Use of Copyrighted Works by AI Systems: Art Works in the Data Mill

Mirko DEGLI ESPOSTI*, Francesca LAGIOIA** and Giovanni SARTOR***

We shall first introduce the use of artificial intelligence (AI) in producing new intellectual creations, distinguishing approaches based on knowledge representation and on machine learning. Then we shall provide an overview of some significant applications of AI to the production of intellectual creations, distinguishing the extent to which they depend on pre-existing works, and the different ways in which such pre-existing works are used in the creative process. In addition, we shall discuss some methods to automatically assess the similarity of works and styles, in the context of AI technologies for text generation. Finally, we shall discuss the legal aspects of AI-reuse of copyrighted works, focusing on the rights of the authors of such works relative to the process and the outputs of AI.

I. AI AND INTELLECTUAL CREATIONS: FROM KNOWLEDGE REPRESENTATION TO MACHINE LEARNING

In the last two decades, artificial intelligence (AI) has gone through a major and rapid development, increasing its societal impact. A number of successful AI applications have been developed which have left the laboratories entering the market place, private and public organisations, the life of everybody, in both physical and virtual environments. AI systems are now able to recognise sounds, images, and faces, to translate texts, to make forecasts, to handle dialogues, to engage in trading, etc.

This success is linked to a major change in the leading paradigm in AI research and development. Until a few years ago, the main idea was that, to develop an intelligent system, humans had to provide a formal representation of the relevant knowledge, as well as algorithms to make inferences out of such knowledge. More recently, the focus has shifted to the possibility of applying machine learning algorithms to vast masses of data: given a vast set of examples of correct or interesting behaviour, embodied in a dataset, a system may learn to address new cases according to the examples given. Besides this basic model of machine learning (so-called supervised learning, based on a pre-defined training set of examples) other models exist in which

* University of Bologna, Department of Computer Science and Engineering; email: mirko.degliesposti@unibo.it.
** European University Institute of Florence, Law Department, University of Bologna; CIRSFID; email: francesca.lagioia@unibo.it.
*** European University Institute of Florence, Law Department, University of Bologna; CIRSFID; email: giovanni.sartor@unibo.it.

European Journal of Risk Regulation, 11 (2020), pp. 51–69
doi:10.1017/err.2019.56

learning is based on the evaluation of the results provided by the system (reinforcement learning) or on the non-supervised clustering of information. A very successful example of a system based on learning is automated translation: machine learning methods, as applied to large datasets of multilingual documents, have enabled the development of effective translation systems (such as Google Translate). By finding probabilistic associations between patterns of words in different languages, according to statistical methods (or neural networks), results have been achieved that could not be obtained through analytical methods, ie through the extraction of syntactic and semantic structures from source text according to linguistic theories. It has indeed been claimed that data have an "unreasonable effectiveness", ie that "invariably, simple models and a lot of data trump more elaborate models based on less data".[1]

The division between these two models (knowledge-representation vs machine-learning) should not be taken too rigidly, since systems based on formal models can rely on machine learning to support the construction of knowledge bases, and machine learning systems can use various knowledge structures (rules, ontologies, etc) to improve their performance. However, the two kinds of systems can be viewed as useful paradigms, and the transition to systems centred towards machine-learning leads to new problems pertaining to the use of pre-existing copyrighted works.

The main issue so far addressed, relative to automated creativity, has been whether copyright protection may also be granted to AI-generated works. It has indeed been wondered whether the originality that is protected by copyright only concerns human expressions (the author's personal touch) or whether and under what conditions it may extend to new expressions created by automated tools, when such expressions embed creative inputs (by software developers, users, or machines). Here we shall not address this issue,[2] but a separate, orthogonal problem, so far neglected but equally important: whether authors' rights over their works also extend to outputs produced by AI on the basis of such works.

This problem emerges in different ways in the two approaches to AI we have described. In the context of human-produced knowledge representations, the production of new works either fully reflects human choices (pertaining to knowledge representation or algorithms) or is based on combinatorial/random variations. Issues pertaining to reuse only emerge if the elements to be combined or transformed pertain to previous copyrighted work.

In the context of machine-learning systems an additional issue arises, pertaining to the use of copyrighted works as elements of a training set, and to the connection between the training set and the automated output. Artistic works become inputs for a data-mill,[3] which amalgamates, adapts, and develops micro-elements, patterns, styles into new outcome, different from each of the input works, and possibly having some novel artistic meaning.

[1] A Halevy et al, "The Unreasonable Effectiveness of Data" (2009) IEEE Intelligent Systems 8.

[2] See A Ramalho, "Will Robots Rule the (Artistic) World? A Proposed Model for the Legal Status of Creations by Artificial Intelligence Systems" (2017) Journal of Internet Law; J Ginsburg and L Ali Budiardjo, "Authors and Machines" (2019) 34 Berkeley Technology Law Journal.

[3] M Sag, "Orphan Works As Grist for the Data Mill" (2012) 27 Berkeley Technology Law Journal 1503.

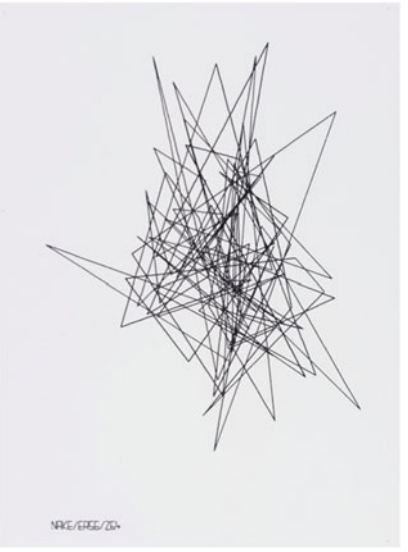

Figure 1. *Random Polygons*, Frieder Nake, 1965

II. Cases and scenarios

In this section we describe some systems intended to produce new artistic works.

Various authors have accepted the idea that AI systems can exhibit creativity, to different degrees. For instance, Boden[4] distinguishes three levels of creativity in AI systems: (1) combinatorial creativity, which produces new combinations of familiar models and ideas; (2) exploratory creativity, which selects and explores new opportunities within known models and ideas; and (3) transformational creativity, which departs from existing models and ideas.

Here we shall focus on the extent to which computer creations rely on existing works, and on the extent to which the new creations converge towards or depart from such works. In the following we shall first present some systems whose "creativity" is not based on pre-existing works, and then move onto those which combine discrete elements of previous works, and finally discuss those which are trained to produce new works according to pre-existing examples.

As an example of "creativity" that is not based on pre-existing works, being directed by computer generated inputs, consider the *Random Polygons* by Frieder Nake, displayed in Stuttgart in 1965, a graphic composition of broken lines randomly oriented to form open or closed polygons (see Figure 1). The computer had not been explicitly programmed with the coordinates of the polygons' vertexes. These coordinates were calculated on the basis of complex mathematical functions and parameterised according to values provided by the computer itself, such as the time of the computer clock.

Let us now turn to computer creations based on previous works. In some systems, the new works result from new combinations of fragments of previous works. In other

[4] MA Boden, "Computer Models of Creativity" (2009) 30 AI Magazine 23.

systems, those based on contemporary machine learning approaches, discrete fragments from previous works can no longer be detected, since the system seamlessly merges its inputs to generate new outputs.

An example of combination of pre-existing discrete elements is provided by WASP[5] (Wishful Automatic Spanish Poet), which composes formal poetry in Spanish. WASP applies a set of construction heuristics to its initial data, according to certain parameters, such as vocabulary and user-selected words and verse patterns. On this basis it generates either an unrestricted set of verses, or a poem fitting user-selected literary structures.

An example of the application of machine learning to artistic creation is provided by Xiaoice, a Microsoft system that generates literary works, based on a training set consisting of 2,027 modern Chinese poems. Xiaoice has become popular on various social platforms[6] and has already generated more than 12 million poems since its release in July 2017, among which 139 have been published in the poetry collection "Sunshine Misses Windows". Xiaoice poems are based on images:[7] a set of keywords representing objects and sentiments are extracted from an image and then associated with keywords linked to the poems in the training set. A neural network uses the latter keywords to generate verses. For instance, the following poem was generated by A Xiaoice on the basis of Figure 2:

> Wings hold rocks and water tightly
> In the loneliness
> Stroll the empty
> The land becomes soft.

A striking result of machine learning in the visual arts is provided by *The Next Rembrandt*, an application developed by Microsoft in collaboration with the Rembrandt House Museum, aimed at bringing the great Dutch painter back to life, ie to have "him" create one more portrait, as reported on the homepage of the dedicated website.[8] The project team defined the subject of the painting to be produced: a Caucasian male with facial hair, aged 30–40, wearing black clothes with a white collar and a hat, facing to the right (a most typical Rembrandt subject). Then the system was trained on a database including all portraits by Rembrandt. It learned to mimic the painter's style based on the most significant aspects of his paintings: composition, geometry and proportion patterns, brushstrokes, light, as well as painting materials, texture patterns of canvas surfaces and layers of paints. The entire process has resulted in a 3D painting which looks exactly like a work of Rembrandt (see Figure 3).

[5] P Gervás, "Wasp: Evaluation of Different Strategies for the Automatic Generation of Spanish Verse" (2000) Proceedings of the AISB-00 symposium on creative & cultural aspects of AI.

[6] Heung-yeung Shum et al, "From Eliza to XiaoIce: Challenges and Opportunities with Social Chatbots" (2018) 19 Frontiers of Information Technology & Electronic Engineering 10.

[7] Wen-Feng Cheng W-F et al, (2018) Image inspired poetry generation in xiaoice <arxiv.org/pdf/1808.03090.pdf> accessed 26 September 2019.

[8] See <www.nextrembrandt.com> accessed 26 September 2019.

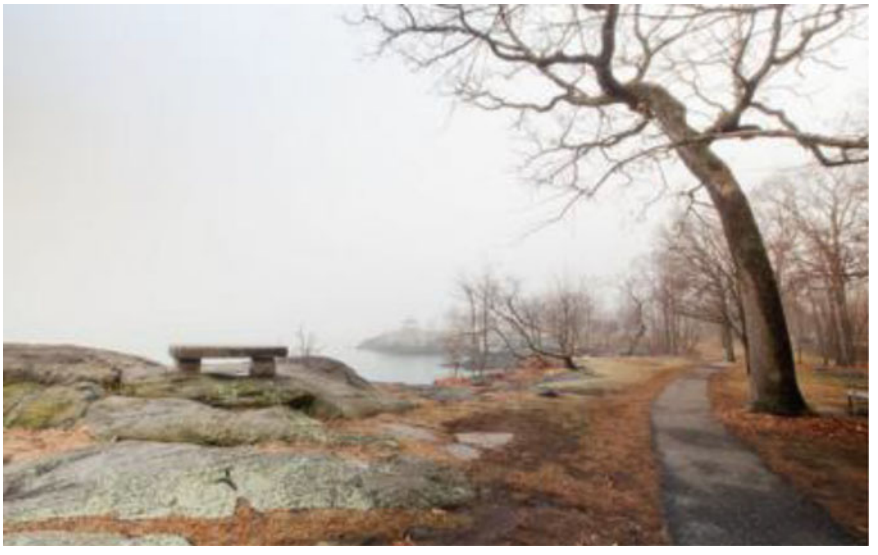

Figure 2. Image example upon which a poem is generated, from Cheng et al 2018 (see note 7)

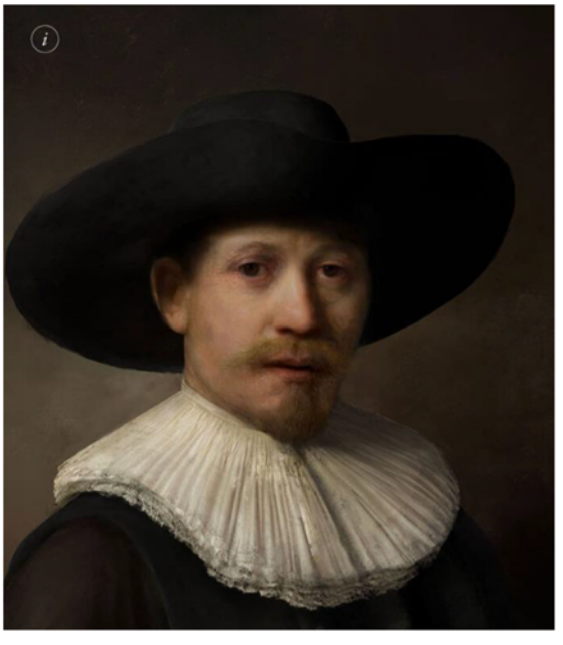

Figure 3. *The Next Rembrandt*, by Microsoft and the Rembrandt Museum

A similar application, publicly available, is Deep Dream Generator,[9] which uses a convolutional neural network to find and enhance patterns in images.[10] The application is trained on a vast data set of images to identify patterns (eg faces)

[9] <deepdreamgenerator.com> accessed 26 September 2019.

[10] A Mordvintsev et al, "Deepdream-a Code Example for Visualizing Neural Networks" (2015) 2 Google Research 5.

Figure 4. Examples of images generated by Deep Dream Generator according to different authorial styles (Van Gogh, Renoir, and a mix of various styles).

and painting styles. Based on the acquired knowledge, it transfers painting styles to images uploaded by users on the web platform. Figure 4 shows the results produced by Deep Dream Generator, when applying different artistic styles to the same input image.

So far, we have only considered what we may call "convergent AI-creativity", namely, cases in which an AI system emulates the authors of pre-existing works. In such cases, the AI system aims at producing what the authors of pre-existing works would have created had they addressed a certain topic or issue.

Some recent applications show how AI can also be used to deliver "divergent AI-creativity", ie to create new works that to some extent depart from pre-existing ones, providing a result that differs from what the authors of the latter works would have created had they addressed the same topic or issue. This approach is exemplified by the CAN (Creative Adversarial Networks) system, which combines convergent and divergent AI-creativity, based on machine learning:[11] it learns art styles by looking at examples and increases the arousal potential of the generated art by deviating from the learned styles. This is obtained through the interaction of two sub-networks, ie a generator and a discriminator, which act as two counterpoised forces. The discriminator is trained on a large set of artworks associated with style-labels (eg Renaissance, Expressionism, Impressionism, Abstractionism), while the generator does not have access to any kind of artworks. The generator proposes sketches to the discriminator, which evaluates such sketches based on two distinct criteria: the extent to which a sketch fits within one of the styles and the extent to which the same sketch exhibits aspects that do not fit within any of the given styles. In this way the generator learns to create works that show innovative elements while preserving a connection to given styles, so that an observer can relate to the new work, while appreciating its originality. Figure 5 shows examples of images generated by CAN.

Systems exist that provide artistic performances, in the domain of music and the performing arts, rather than generating new artistic creations. Such systems raise

[11] A Elgammal et al, "Can: Creative adversarial networks, generating art by learning about styles and deviating from style norms" (2017) <arxiv.org/abs/1706.07068> accessed 26 September 2019.

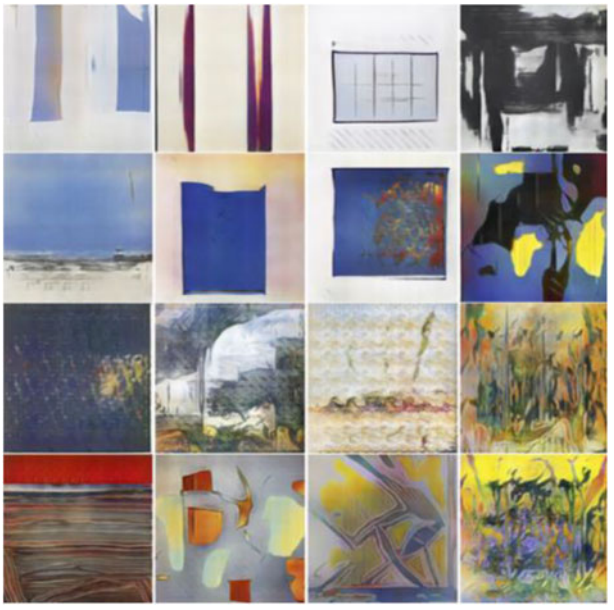

Figure 5. Examples of images generated by CAN, from Elgammal et al 2017 (see note 11)

issues that are similar to those addressed so far, to the extent that they reproduce the characteristic voice or movements of particular artists (eg the Vocaloid system by Yamaha).

III. Measuring style and similarities

Any legal framework of AI-reuse of copyrighted works would greatly benefit from quantitative measures or specific indicators to measure correlations between training sets and outputs generated by computational processes. Such measures should enable us to determine to what extent a new work can be viewed as derivative of pre-existing works, creative, or embodying a particular style.

To put it differently, if we want to successfully address the legal issues concerning AI-reuse, we need quantitative methods to measure the extent to which algorithms grasp, reuse, and reproduce the works or the "style" of an author, and the extent to which they are "creative". In this context, *creativity* means the *generation* of novel, original, and hopefully coherent *structures* using elementary elements according to old or newly-created rules. This very general and debatable definition applies not only to literature, but also to music, painting, and any other form of art.

In any creative process, nothing is really generated from scratch. This also holds for human creations, which result from complex interactions of multiple factors: individual experience and skills, previously produced works, awareness of contemporary and past works by others, the subject matter of the new work, the author's direction towards innovation, and much more. Because of this, creative works usually include *patterns*

characterising the author's identity, and in particular his or her individual style. These patterns are often sufficient to identify and distinguish the author or, more important for us, to quantify how much an artificially generated work owes to the previous works of a given author used in the learning phases.

It is true that the concept of "style" itself is a very ambiguous and questionable one, vaguely defined and also strongly dependent on the medium (written text, visual art, video, music, etc). However, precisely the generality and abstractness of this definition enables a mathematical approach, even if concrete generic applications present a further level of complexity, not yet fully overcome.

A somewhat naive way of tackling the problem is to imagine a mathematical space where each author (or each work of each author) is represented by a point, possibly, but not necessarily, identified by a set of coordinates. We can then assume that it is also possible to define a distance between any pair of points in this space, so that such points are closer the more the corresponding works are "similar". This distance can be used to quantify how much an artificially generated work owes to other works in the training set and, in particular, to the works of a given author. It is interesting to note that the foundations of information theory and the related theory of algorithmic complexity ensure the existence, at least theoretically, of such "similarity distances". Thus, they indicate how to deal with real cases in concrete and operational ways.[12] This fundamental approach to the measurement of similarities between digital objects is not recent. On the contrary, it constitutes one of the fundamental theoretical bases that have stimulated the development of new and modern methods of classification, nowadays based on machine learning or deep learning techniques. Powerful methods are available today for processing naked data, without (almost) any additional representation of knowledge. A direct approach to unstructured data is greatly emphasised by the most modern algorithms. Computer science has in fact addressed the general problem of clustering and discriminating objects of different natures for a very long time. An approach has become dominant in the last decades: define and extract suitable features of the objects being considered and then, in the spirit of machine learning, train and use a suitable algorithm/machine (neural nets, supported vector machine, Bayesian tools, etc) to discriminate, classify, and clusterise such objects, placing them in several known or unknown classes.[13] Thus, in parallel with the constant and impetuous growth of artificial systems capable of mimicking human creativity,[14] methods and algorithms are being developed for measuring creativity.

In the following we shall focus on Natural Language Generation, which encompasses the capability of machines to synthesise text resembling spoken or written language typically employed by humans.[15] We will examine how the creativity of generated text and its relations to the corresponding training set can be measured.

[12] See, for example, M Li and P Vitanyi, *An Introduction to Kolmogorov Complexity and Its Applications* (Springer-Verlag, New York 1997) and R Cilibrisi and PMB Vitányi, "Clustering by Compression" (2005) 51 IEEE Transaction on Information Theory 11523.

[13] IH Witten et al, *Managing Gigabytes* (Morgan Kaufmann Publishers 1999).

[14] MA Boden, "Creativity and Artificial Intelligence" (1999) 103 (1-2) Artif Intell 347.

[15] E Reiter et al, *Building Natural Language Generation Systems* (MIT Press 2000) p 33.

Natural Language Generation has recently gone through a period of exciting change, mostly due to the huge development in the area of deep learning,[16] whose methods and algorithms have contributed to significant steps forward in the generation of creative artificial textual documents, in human languages. Stunning results have been already achieved even though there is still a long way to go before generating long texts with a coherent semantic structure. Here we like to recall just two different methods that allow the definition of similarity distances between natural language texts, and which can be used to address the legal issues related to originality, such as similarity and plagiarism. Neither method, following the spirit of machine learning, requires any additional representation of knowledge, but considers texts as a simple sequence of characters. For both methods an input text is "just" a sequence of symbols: neither the content of the text nor its grammatical aspects are considered: letters of the alphabet, punctuation marks, blank spaces between words are just abstract symbols, without a hierarchy. Moreover, the word as basic constituent of the text has no more meaning than other aggregates of symbols: it is just viewed as an *n-gram*, ie as an arbitrary sequence of *n* consecutive characters.

We might say that this approach toward texts started in 1948, when Shannon[17] determined that the quantity of information within a message is the minimum number of bits needed to codify it, and defined *entropy* as the minimum number of bits per character.

There are programs which attempt to codify a message using the least possible number of bits: they are the data compression programs. Data compression is nowadays a very well established field in information theory, thanks to the seminal papers published by Ziv, Lempel, and their colleagues in the 1970s.[18] They proposed a variety of compression algorithms (the family of) *LZ algorithms*, based on the idea of a clever *parsing* (subdivision) of a symbolic sequence, ie the splitting of a text string into pieces that can be used to produce a shorter and equivalent version of the string itself. The compression rate (obtained by comparing the dimension of the compressed text with that of the original text) allows the entropy of a text to be estimated.[19] Essentially, we can consider the compression algorithm as a machine that learns from the past (the part of the sequence already processed) to effectively forecast the future (the characters that will appear next). For example, just to give an idea, the *LZ77*

[16] Y LeCun et al, "Deep Learning" (2015) 521 Nature 436.

[17] Information theory was born in 1948 with CE Shannon's, "A Mathematical Theory of Communication" (1948) 27 Bell System Technical Journal 379, which poses and solves the problem of defining the amount of information contained in a "message", for example a text or more generally any sequence of symbols. For a more extensive account see JR Pierce, *An Introduction to Information Theory: Symbols, Signals and Noise* (Dover, New York 1980).

[18] *Cf*, among others, A Lempel and J Ziv, "On the Complexity of Finite Sequences" (1976) 1 IEEE Transactions on Information Theory 75; A Lempel and J Ziv, "A Universal Algorithm for Sequential Data Compression" (1977) 3 IEEE Transactions on Information Theory 337; A Lempel and J Ziv, "Compression of Individual Sequences via Variable-Rate Coding" (1978) 5 IEEE Transactions on Information Theory 530, and the review paper A Lempel and J Ziv, "Compression of Individual Sequences via Variable-Rate Coding" (1998) 44 IEEE Transactions on Information Theory 2045.V

[19] AD Wyner, "Typical Sequences and All That: Entropy, Pattern Matching and Data Compression" (1995) IEEE Information Theory Society Newsletter.

algorithm compresses a string sequentially, starting from the beginning and following this procedure:

- if a character has not appeared before, it rewrites it as is;
- if a character has already come up, it finds the longer substring seen until that moment which matches a substring starting with the character under consideration; it writes the length of that substring and the number of characters setting it apart from the considered character.

The string is therefore globally codified by a sequence of characters and pairs of numbers. An example is useful: let us consider the following sentence, where the blank space has been replaced by the underscore "_":

hey_diddle_diddle_the_cat_and_the_fiddle

Starting from the left, there are no repeated characters until the second "d", so the compressor writes "hey_di"; the second "d" is an already seen character, but the couple "dd" is not, thus the second "d" is coded as (1,2). Then another "d" follows which is coded as (1,1). Then an "l" follows, which never appeared before, and is therefore coded by "l", and so on.

The final result is the following:

hey_di	d	d	l	e	_diddle	_	t	he	_
hey_di	(1,2)	(1,1)	l	(1,8)	(7,7)	(1,7)	t	(2,19)	(1,4)
ca	t	_	a	n	d	_the_	f	iddle	
ca	(1,6)	(1,4)	(1,3)	n	(1,14)	(5,12)	f	(5,23)	

The second sequence (code) has exactly the same content as the first one (text): during the decoding the compressor interprets the numbers (x,y) as the instruction "go back y characters and from there copy here x characters". The dimension of the compressed file is smaller to the extent that the strings the compressor finds are longer. In other words, *LZ77* uses the redundancies (repetitions) inside the text to write a shorter version of it.

By developing these ideas, one can obtain effective instruments for classification and clustering and also for the authorship attribution problem, ie to estimate how much a text owes to the style of another: the concept of *relative entropy*. Relative entropy is a very powerful tool with which to quantify the difference among sequences, and therefore among authors or styles (see, for example, the survey by Stamatatos).[20]

Again, everything started a while ago, in 1993, when Ziv and Merhav[21] proposed a method to estimate the relative entropy (or Kullback-Leibler divergence) between pairs of information sources. Relative entropy is basically a measure of the similarity

[20] E Stamatatos, "A Survey of Modern Authorship Attribution Methods" (2009) 60 Journal of the American Society for information Science and Technology 538.

[21] J Ziv and N Merhav, "A Measure of Relative Entropy Between Individual Sequences with Application to Universal Classification" (1993) 39 IEEE Transactions on Information Theory 1270.

between the information emitted by the sources. They proved that a modified version of an LZ algorithm, where the subsequences for a sequence are searched within another sequence, can approximate the relative entropy between the two sources that generated such sequences. This important result was used in various subsequent studies, among which were those conducted by Benedetto et al[22] and Basile et al,[23] to deal with text classification and clustering.

This method can be implemented through compression algorithms, to effectively measure the similarity distance in different domains, not only textual ones. Algorithm *LZ77* suggests in fact a method for estimating relative entropy, and hence the "proximity" between texts:[24] a method that is elementary, easy to implement, and efficient on large amounts of data. The approach can be considered as a precursor of much more modern and complex methods, and for this reason we give here a very brief description of it, mainly to underline its characteristic of working on abstract symbolic sequences, without any further knowledge attached to it.

Suppose you compress text A+X, that is, the text obtained attaching text *X* to text A. The compression algorithm, being sequential, will code first all the characters in A and then will start coding those in X, looking for the strings it has already read, ie those in text A. The more similar the two texts are, the longer the strings in X which are found in A will be, and therefore the more effective the compression of the whole file will be. In fact the compressor can in this case use not only the redundancy within the single texts, but also the redundancy between the two texts, improving the compression rate. The difference between the lengths of the compressed versions of A+X and of A, divided by the length of X, is a measure of the relative entropy of X with respect to A. Such a number is smaller when more parts of X are found in A or, more evocatively, when knowing A makes it easier to express the content of X.[25]

Let us now turn to a second method to determine the similarity distance between texts, which is again very simple and easy to implement, but still quite effective in measuring stylistic similarity between textual pieces. It is based on n-grams, and it has a relatively short history in published literature.[26] Here we present the definition of similarity distance as introduced and discussed by Basile et al.[27] We call ω an arbitrary n-gram, and we denote by $f_X(\omega)$ (respectively $f_Y(\omega)$) the relative frequency with which ω occurs in text *X* (respectively *Y*). $D_n(X)$ is the *n-gram dictionary* of text *X*, that is, the set of all

[22] D Benedetto et al, "Language Trees and Zipping" (2002) 88 Physical Review Letters 48702.

[23] C Basile et al, "An Example of Mathematical Authorship Attribution" (2008) 41 Journal of Mathematical Physics.

[24] Benedetto et al, supra, note 22.

[25] If you are interested in a detailed analysis of the compression of attached files, see A Puglisi et al, "Data compression and learning in time sequences analysis" (2003) 180 Physica D 92.

[26] After a first experiment based on bigram frequencies presented in 1976 by WR Bennett, *Scientific and Engineering Problem-Solving with the Computer* (Prentice-Hall, Englewood Cliffs, NJ 1976), V Keselj and collaborators published in 2003 a paper in which n-gram frequencies were used to define a similarity distance between texts: V Kešelj et al, "N-gram-based author profiles for authorship attribution" (2003) Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING'03, pp 255–264. See also R Clement and D Sharp, "Ngram and Bayesian classification of documents for topic and authorship" (2003) 18(4) Literary and Linguistic Computing 423.

[27] Basile et al, supra, note 23.

n-grams which have non-zero frequency in X (similarly for Y) and we define what we will call the n-gram distance between text X and text Y as:

$$d_n(X,Y) := \frac{1}{|D_n(X)| + |D_n(Y)|} \sum_{\omega \in D_n(X) \cup D_n(Y)} \left(\frac{f_X(\omega) - f_Y(\omega)}{f_X(\omega) + f_Y(\omega)} \right)^2 \quad (1)$$

Here, $|D_n(X)|$ and $|D_n(Y)|$ are the numbers of different n-grams in the two dictionaries, respectively, and the sum is taken over all the n-grams occurring in the two texts.[28] Thus, the distance increases to the extent that the frequency of n-grams in one text differs from the frequency of the same n-grams in the other text.

In this section we have discussed some simple tools in the field of text generation that can be easily implemented, and are general enough to be effective in non-textual contexts (images or music, for example) and fast enough to work on large amounts of data. The application scenarios are very varied and complex and we will certainly see in the future a constant growth of more and more refined tools for the study of creativity and copyright. There is no doubt that along with an impetuous development of algorithms capable of producing artistic objects of various kinds, there will be a similar growth in methods and algorithms to estimate, if not measure, the degree of creativity and reuse of the author's material contained in a learning set. These tools will be extremely useful for anyone who has to make political, economic, or legal decisions in the field of artistic creativity and copyright.

In particular they may be used to determine whether the production of AI-generated works may amount to plagiarism, or whether they may constitute derivative creations, the unauthorised distribution of which may amount to a violation of intellectual property. They may also be used to determine whether, as we shall argue in the following, the reproduction of the style of an individual author is so vast that it may involve a violation of the author's personality rights.

IV. Assessing the originality of texts generated by neural networks

The distances introduced in the previous section have been used by Lippi et al[29] to provide an extensive empirical evaluation of texts generated with Long-Short Term Memory (LSTM) networks, one of the most widely used deep learning models for Natural Language Generation. The authors trained an LSTM network with a corpus consisting of novels by Charles Dickens. The network was trained to predict the next character in a given text, and thus it could be employed to iteratively generate textual documents of any desired length. Then the distances were used to measure the similarities between the LSTM-generated texts and texts commonly produced by

[28] To be more precise, d_n is a pseudo-distance, since it does not satisfy the triangular inequality and it is not even positive definite: two texts X, Y can be at distance $d_n(X,Y) = 0$ without being the same, but this has basically no effect on concrete applications. Note that in the previous formula, in contrast with what happens for the Euclidian distance, each term of the sum is weighted with the inverse of the square of the sum of the frequencies of that particular n-gran. In this way rare words, ie n-grams with lower frequencies, give a larger contribution to the sum.

[29] M Lippi et al, "Natural language statistical features of lstm-generated texts" (2019) IEEE Transactions on Neural Networks and Learning Systems 12.

humans. The same approach can be used to measure similarities between a training corpus of human-generated works and the output resulting from the AI reuse of that corpus.

LSTMs are a kind of recurrent neural network, developed at the end of the 1990s[30] to address the issue of so-called vanishing gradients, which until recently has greatly limited the performance of standard recurrent neural networks.[31] Recurrent neural networks allow input sequences of arbitrary lengths to be processed, by exploiting a number of hidden layers. A cell's output is the function not only of its input layer at the present moment, but also of the state of the relevant hidden layer at the previous moment.[32] Recurrent neural networks are typically trained with Backpropagation Through Time,[33] suffering from the well-known problem of vanishing or exploding gradients[34] (which lead to uncontrolled errors). Thus, plain recurrent neural networks are scarcely used in practice. LSTMs solve this issue by a *memory cell*, a hidden cell that controls the information flow: the LSTM cells are then capable of maintaining their state over time, of forgetting what they have learned, and also of allowing novel information in. The most widely employed LSTM architecture for text generation is based on character-level sentence modelling. Basically, the input of the network consists in a certain number M of characters, which correspond to a fixed-size portion of text, whereas the number of output neurons is the total number S of possible character symbols in the text, each neuron corresponding to one such symbol. In Lippi et al,[35] such a model is trained in a classic supervised learning setting, where the input training corpus (novels by Charles Dickens) is fed to the network, using as target the true (known) next character, as it appears in the corpus. Long-range dependencies can be captured by the model by maintaining the previous cell-states as subsequent inputs are presented to the network. At the end of the learning process, a probability is associated to each output symbol, and this probability is used to generate an artificial text of arbitrary length: characters are generated by sampling the probability distribution. The research described in Lippi et al[36] accomplishes an initial step for providing a quantitative method to measure the creativity of algorithms. It assesses two distinct yet strictly intertwined aspects: to what extent the algorithm is capable of capturing the stylistic traits of a given author, and to what extent it avoids plagiarism. The first assessment is obtained through the use of the similarity distances just discussed. Such distances provide a quantitative estimate of the extent to which the stylistic traits of the deep-learning generated text are similar or superimposable on those of the texts in the training set (the works of Dickens, in this case). The second assessment is obtained by measuring the so-called Longest Common Subsequence

[30] S Hochreiter and J Schmidhuber, "Long Short-Term Memory" (1997) 9 Neural Computation 1735.

[31] Y Bengio et al, "Learning Long-Term Dependencies with Gradient Descent Is Difficult" (1994) 5 Neural Networks, IEEE Transactions on 157.

[32] Lippi et al, supra, note 29.

[33] PJ Werbos, "Backpropagation Through Time: What It Does and How to Do It" (1990) 78 Proceedings of the IEEE 1550.

[34] Hochreiter and Schmidhuber, supra, note 30; Bengio et al, supra, note 31.

[35] Lippi et al, supra, note 29.

[36] ibid.

(LCS), as one of the simplest ways to quantitatively measure plagiarism: given the k-th character, x_k, of the artificial text, we denote by L_k the length of the longest subsequence of consecutive characters starting at x_k that is also contained in the training corpus (and thus could possibly be plagiarised from it). By combining the distance measures above with statistics over the set of all L_k, we can quantify the extent to which various algorithms are able to reproduce some stylistic traits of a given corpus while generating new texts.

From an artificial creativity perspective, a good algorithm is one that can reproduce an author's style while minimising self plagiarism relative to the corpus used in the training phase. In more quantitative terms, an excellent algorithm to generate artificial texts should meet two desiderata. On the one hand there should be a small similarity distance relative to the works of the author in question. On the other hand, the statistical distribution of CLS L_k should not be too different from the "natural" one of that author (each author has a certain degree of autoplagiarism, measurable by calculating the statistics of the CLS of a given work of the author with respect to all the other works by the same author). For example, if one computes the maximum of the L_k of the novel *Oliver Twist* with respect to other Dickens novels (*David Copperfield*, *A Tale of Two Cities*, *Bleak House*, *Great Expectations*, *Hard Times*), one gets a maximum of 45 characters. One can use this simple indicator to assess the degree of plagiarism of an AI-generated text trained on a corpus by this author. This approach can already deliver significant results, even though more sophisticated uses of LCD L_k statistics are desirable and will have to be developed in the future to give quantitative support to copyright laws.

V. Law and policy of the reuse of original creations through AI

Here we sum up the main points to take away from the previous sections and provide some tentative ideas for a legal framework.

In Section I we introduced the trend from human knowledge representation to machine learning: AI increasingly relies on algorithms that learn from examples, rather than on formally specified knowledge provided by humans.

In Section II we observed that this trend involves new ways of using pre-existing works to build new creations. Not only can (portions of) pre-existing works provide elements to be combined into new creations, or suggest to human programmers algorithms for extracting such combinations; increasingly, previous works are rather used as sets of examples on which machine learning algorithms rely to deliver new creations.

Such creations will depend on the learning algorithm being adopted, but also on the nature and the quantity of examples being processed (the more examples there are, the less a new creation is likely to correspond to a single input work), the difference among the authors of the examples (the more the authors, the less a new creation is likely to fit the oeuvre of a single author), and the convergent or divergent standard adopted by the learning algorithm (more divergence leading to departures from the input works).

In Section III we presented methods for measuring similarities between texts, and discussed their application to determine the extent to which an AI-generated text

reproduces the style of pre-existing works used as training sets, and exhibits creativity rather than plagiarism.

In Section IV we addressed text generation through recurrent neural networks, and examined an application of the metrics introduced in Section III to AI-generated texts.

Given this framework, we can now address potential copyright issues pertaining to the relation between input-works and output-works in the age of machine learning. We can distinguish two kinds of issues: a process issue and an outcome issue.

The process issue pertains to whether the very fact of using a pre-existing work as an example in a training set may involve a violation of the rightholders' entitlements over those pre-existing works. This concerns both the creation of a new digital copy of the work to be included in the training set and the subsequent manipulation of that digital copy through the system's algorithms.

The outcome issue pertains to whether a violation may consist in the fact that an AI creation is too similar to some of the pre-existing works on the basis of which it was constructed. This concerns the fact that AI creations can retain relevant aspects of some works in the training set, so that such creations can be viewed as unauthorised modifications of these works.

The processing of copyrighted works for the purpose of artistic creation is similar to other computational uses of copyrighted works, which also involve the production of digital copies and the processing of such copies to produce new outputs and services. The examples are varied, going from the simple indexing of documents on the web, to the aggregations of news, to the indexing of the content of books, to the detection of plagiarism, etc.

In the US legal system, a very tolerant approach has been adopted toward the use of copyrighted works for purposes that differ from the ordinary goal of serving a human readership. Indeed, in many instances, unauthorised copying that would have been unlawful if intended for human readers has been considered lawful when only addressed to computer systems.[37] This approach has supported the creation of a number of innovative technology-based services, which facilitate access to literary and artistic works and provide information over such works, or over the matters they address.

The legal basis for this approach has been provided in the US by the idea that copyright should not prevent "transformative uses", namely, uses that are "productive" and "employ the [. . .] matter in a different manner or for a different purpose from the original",[38] and thus do not significantly affect the rightholders' exclusive rights over traditional uses. This idea has inspired the decisions of various important cases. For instance, in the *Google v Authors Guild* case (2015), the US Court of Appeals for the Second Circuit ruled that:

> "Google's making of a digital copy to provide a search function is a transformative use, which augments public knowledge by making available information about Plaintiffs' books without providing the public with a substantial substitute for matter protected by the Plaintiffs' copyright interests in the original works or derivatives of them".

[37] For a critical discussion, see J Grimmelmann, "Copyright for Literate Robots" (2015) Iowa Law Review 657.

[38] PN Leval, "Toward a Fair Use Standard" (1990) 103 Harvard Law Journal 1105.

A similar argument was earlier endorsed by the US Court of Appeals for the Fourth Circuit in *Vanderhye v iParadigm* (2009), concerning the Turnitin software. This system detects cases of plagiarism, in particular with regard to students' works, by comparing a new work with a database of pre-existing works. The judges argued that authorisation by the authors of the works in the database was not needed, since the use of digital copies of their works by Turnitin was transformative, being "completely unrelated to expressive content", and thus it did not create a market substitute.

It is true that more stringent criteria have been adopted in Europe. The EU Copyright Directive, at Article 5, only excludes from copyright "temporary acts of reproduction [...] which are transient or incidental [and] an integral and essential part of a technological process" and "which have no independent economic significance". In *Google v La Martinière* (2009), Google was ordered by the Paris Court of First Instance to pay damages to publisher La Martinière, for the unauthorised reproduction of copyrighted works in the context of the Google Books project. Subsequently, La Martinière settled the case, signing an agreement authorising the scanning of its repertoire, and the processing of that repertoire in the context of Google Books. However, also in Europe, legal approaches favourable to transformative automated processing of copyrighted works have often been adopted, by using various legal arguments (eg by assuming non-revocable implied consent when a text is made accessible over the Internet, or by understanding in a broad sense the idea of transiency).

We cannot here examine the merit of a more or less liberal approach towards transformative uses. Let us just observe that on the one hand a more liberal approach reduces transaction costs and favours the provision of new technology-based services; on the other hand, it may reinforce existing imbalances between Internet intermediaries and copyright holders and exclude the latter from the possibility of sharing in the profits resulting from transformative uses. We shall rather focus on an important difference between the transformative uses so far considered in the case law and the use of pre-existing work in AI-based artistic creation.

The typical transformative uses involve unauthorised copying in the context of services having non-expressive purposes, namely, services meant to provide information about pre-existing works; such services do not involve the generation of new works. The cases we are here addressing are different, since they involve the use of copyrighted works for expressive purposes, namely, for training AI systems to generate new works. The purpose of the latter systems is thus aligned with the artistic or communicative purpose of the authors of the pre-exiting works included in the training set.

It seems to us that the generation of new expressions based on pre-existing materials raises specific issues, in comparison with other automated uses of copyrighted works.[39] It involves extracting and automatically deploying what is connected not only to single pre-existing works, but also to the creative personality of authors. Therefore, it concerns not only the economic rights but also the personal rights of authors. It may also involve aspects that go beyond intellectual property, as strictly understood, towards the protection of privacy and personal identity.

[39] For a liberal approach to non-expressive uses, see Sag, supra, note 3.

Consider a futuristic (but not so much) case in which all novels by an author were collected in a data set and were used, through an AI system, to generate works similar to those of that author, even though having different subject matter, plot, and characters. Even if no fake were produced – the AI creations being clearly presented to the public as such – this system would affect the artistic and personal life of the author. The latter would find himself being emulated by a threatening doppelgänger, and would probably experience persistent pressure to distinguish himself from, and compete with, his digital double. Not only would the system be able to mimic the author's work up to a certain point in time, but it would also be able to frustrate all author attempts to diverge from the system, by departing, in his or her new works, from the subject or style of the works used by the system. In fact, as the author's new works would also enter the system's training set, the system would adapt to the new artistic persona of the author. The problem just described may also be present, to a lesser degree, where the AI-generated works were derived from works of more than one author. The style, pattern, personality of individual authors could still be identifiable within the AI-generated works.

The possibility of replicating an author's individuality would not be limited to literature. Consider a painter whose work has been digitised to train an AI-system, so that it can produce new paintings. The case of Microsoft's Rembrandt, discussed above, shows how accurate the automated emulation of an author could be.

Automated emulation also has implications for performers. Consider, for instance, the case of a vocal avatar reproducing the voice of individual singers. Even if there were no doubts on what performances were executed by the real performers and by their avatars, still the performers would feel displaced by their automated replicants, having to compete with their doubles and to distinguish themselves from the latter.

These examples show how the reproduction of the authors' distinctive aspect (style, ideas, patterns, personal touch) through machine learning systems raises issues that go beyond the usual problem of determining whether a new work can be viewed as a total or partial copy of a pre-existing work. We cannot limit our analysis to determining whether a new AI creation qualifies as a copy or as a derivative work relative to the training set, balancing similarities and differences between any particular item in the training set and the resulting automated creation, as we would do in the case of human creations inspired by previous works. The legal assessment must extend to creations that do not reproduce a single works of an author, but rather emulate the author's artistic personality.

Indeed, some reasons that would favour new creative reuse of previous art works by humans do not apply to the same extent to AI-generated creations. Consider, for instance, the need to protect freedom of speech and literary and artistic expression, or even the need to promote human creativity.

Moreover, extended automated reuse would affect authors to a greater extent than human reuse, given that AI-generation of new creations based on a training set can be unleashed with low marginal costs, and can explore any kind of combinations and variations. Authors would lose control over their own expressive personality, as embedded in their works.

We may indeed wonder whether in this case, as long as aspects of authors' personality can be retrieved from their work, we could raise also issues concerning data protection, or identity protection. Such issues would pertain to the unwanted automated processing of personal data (aspects of authors' personality that can be traced back to the authors themselves). The rationale of restrictions over the automated processing of aspects of authors' personality seems different from the usual rationale of data protection. It concerns information produced by individuals, rather than information about individuals. However, in both personal data and personal creations, automated processing engenders risks that are both quantitatively and qualitatively different from those resulting from human engagement.

In fact, unleashing automated mimesis of human creativity might negatively affect goals that are usually associated with copyright law: protecting the personality of authors, enabling creative works to be duly rewarded, and stimulating creativity. Authors would lose control over the characteristic feature of their expressions and would have to face the competition of cheap AI-generated works, to which the public may adapt.

Given the effect of AI-based art-generation on authors, we may wonder whether the inclusion of pre-existing works in a training set should require consent not only by the legitimate rightholders (since the inclusion may affect the economic rights of the latter), but also by the authors of these works, even after such authors have transferred their economic rights to others. We could also wonder whether consent by authors should always be revocable, like data subjects' consent to the processing of their data according to the GDPR. This would mean that when an author revokes consent, his or her work would have to be removed from the training set. Thus, the author's work would no longer be usable for training an AI system in generating new creations, but the AI-creations already generated would not be affected by the consent withdrawal.[40]

Alternatively, such a personal right of authors could be viewed as an unacceptable limitation to the idea of "data freedom" or rather "computational freedom", namely, to the possibility of applying any computations to any data, to generate new knowledge, artistic creations, algorithms, and technologies. Would this right make sense in an infosphere[41] where humans cooperate and compete with AI systems? The generation of more and more powerful and interconnected information processes and structures may even appear to be the purpose towards which human evolution is directed. This prospect may fit in with the vision of those claiming that the "singularity is near",[42] namely, that an AI-driven explosion of knowledge and technological power will enable humans to overcome the limits of their condition, by merging with information technologies.

Within this "computationalist" perspective, the anthropocentric idea of giving authors control over their artistic identity makes little sense, as humans are no longer at the centre of the artistic world, being displaced by technologies that complement, but also substitute

[40] We may wonder whether the traces of the work in a trained system (eg in the data structures resulting from the training of a neural networks) should also be removed, assuming that this does not require an unreasonable effort. We thank Bert-Jaap Koops for pointing to this issue.

[41] L Floridi, *The Fourth Revolution: How the Infosphere Is Reshaping Human Reality* (Oxford 2014).

[42] R Kurzweil, *The Singularity Is Near* (Viking 2005).

and possibly overcome human creative and cognitive skills.[43] Thus, the idea of data freedom, originally understood as providing space for humans' creative engagement with existing works,[44] acquires a new, and to some extent threatening, meaning, when creation is delegated to AI technologies.

VI. Conclusion

The AI-based generation of artistic works already is a reality, and machine-generated works are already part of the artistic world, entering museums and markets. In particular, thanks to machine learning, new creative embodiments of artistic styles and ideas can be produced based on examples of pre-existing human works (training sets).

Technologies have been developed, and are constantly improving, both for generating new works and for determining the extent to which such works match pre-existing styles, while being creative.

This raises issues pertaining to the scope and implementation of copyright, the promotion of creativity and the protection of the personality of the authors.

In conclusion, we think that the AI-generation of new works, based on training sets of copyrighted works, will engender major legal issues in the near future. Finding ways of empowering authors over the reuse of their work for machine-learning purposes will be a key goal for copyright law in the AI era.

[43] JM Balkin, "Information Power: The Information Society From An Antihumanist Perspective" in E Katz and R Subramanian (eds), *The Global Flow of Information* (New York University Press 2011).

[44] L Lessig, *Remix: Making Art and Commerce Thrive in the Hybrid Economy* (Penguin 2008).