

# Too Many Butterflies from One Chrysalis

## *Continual Learning, Continual Forgetting and the Harmonic Flow of Information*

Elio GRANDE <sup>a,1</sup> Luigi QUARANTIELLO, <sup>a,2</sup>

<sup>a</sup> *University of Pisa*

ORCID ID: Elio Grande <https://orcid.org/0009-0008-2896-5900>, Luigi Quarantiello  
<https://orcid.org/0009-0005-5428-156X>

**Abstract.** Despite addressing dynamic learning scenarios, the Continual Learning paradigm is still an evolving field, with no consensus on a definitive methodology among the numerous approaches proposed. In this study, we reflect upon possible novel perspectives about the learning process itself, posing a few questions: how does information get structured in models' parameters? what if memory and oblivion were two faces of the same coin? therefore, could we conceive a network capable of both learning and forgetting continuously? We put forward that information be distributed as a harmony, meaning that there should be some degree of consonance in the data for the continuous learning process to succeed. Provided that, Continual Learning might be possible, say, as a *variation on the theme*, possibly deeming optimization as a kind of orchestration, even among various agents. We encourage the enhancement of this framework, where current brute-force monolithic models would be surpassed in favor of more efficient agents, capable of evolving dynamically from their interactions.

**Keywords.** Continual Learning, Continual Forgetting, Information.

## 1. Introduction

*«To breed an animal with the right to make promises-is not this the paradoxical task that nature has set itself in the case of man? [...] That this problem has been solved to a large extent must seem all the more remarkable to anyone who appreciates the strength of the opposing force, that of forgetfulness. Forgetting is no mere vis inertiae as the superficial imagine [...].»*

(Friedrich Nietzsche, *Genealogy of Morals*) [1]

To breed artificial intelligence with the right to generalize in the long run - is not this, paraphrasing, the paradoxical task that we dream of, making it incrementally acquire new skills and jump like a giant towards artificial general intelligence? Unfortunately, a cursed phenomenon threatens this eager ambition: *catastrophic forgetting*, that is, the drastic performance disruption on previously learned items after training on a new set

---

<sup>1</sup>Corresponding Author: Elio Grande, [elio.grande@phd.unipi.it](mailto:elio.grande@phd.unipi.it)

<sup>2</sup>Corresponding Author: Luigi Quarantiello, [luigi.quarantiello@phd.unipi.it](mailto:luigi.quarantiello@phd.unipi.it)

of items [2]. Not necessarily «finding solutions that work in the real world, but rather finding stable algorithms that can *learn* in the real world»[3], the Continual Learning (CL) paradigm mainly addresses the abovementioned issue, longing to carry out actual autonomous agents, mostly similar to living beings.

More formally, CL can be defined as follows [3]:

**Definition 1.1.** Given  $\mathcal{D} = \{D_1, D_2, \dots, D_N\}$  a potentially infinite sequence of unknown data distributions, where at each time step  $i$  a training set  $Tr_i = \{X_i, Y_i\}$  is drawn from  $D_i \in \mathcal{D}$ , a Continual Learning algorithm  $A_i^{CL}$  is defined as:

$$\forall D_i \in \mathcal{D}, \quad A_i^{CL} : \langle h_{i-1}, Tr_i, M_{i-1}, t_i \rangle \rightarrow \langle h_i, M_i \rangle$$

where  $h_i$  is the parametric model learned after seeing all the training sets up to  $Tr_i$ ,  $M_i$  is an external memory that can be used to store samples from previous training steps, and  $t_i$  is an optional task label. In other words, CL heavily focuses on creating flexible agents that can adapt to an infinite number of tasks, enabling also to reuse and transfer knowledge. This marks a distinct separation from standard Machine Learning (ML) approaches, where the objective is to achieve optimal performance on a single task.

Nonetheless, CL remains an evolving field, with no consensus yet established on a definitive methodology among the numerous approaches proposed. As highlighted in the definition above, most current *state-of-the-art* CL algorithms rely on the use of a memory buffer, with examples from previous tasks. In other words, to make neural networks retain knowledge, they need to be constantly prompted with the same information over and over. Such rehearsal approaches [4,5], despite being functional and easy to implement, represent a simplistic solution to the problem of forgetting. Rather than enabling neural networks to retain knowledge effectively, these methods essentially require the systems to *re-learn* information over time. Moreover, they diverge significantly from the human way of learning and can be computationally intensive, posing scalability issues *by design*.

With these limitations, it seems that a novel vision about the learning paradigm is needed. Indeed, in this work, we want to reflect on some questions concerning the deep nature of continuously learning from different data. In the first place, how does information get structured in a model's parameters? Secondly, the CL community strongly assumes memory to be an incremental process. What if, instead, memory and oblivion were two faces of the same coin? As a consequence, could we think of a model capable of both learning and forgetting continuously?

## 2. To Learn *secundum naturam*

Why to imitate something else's nature when one's own is given? However, if not the nature of an animal, and perhaps more specifically of a mammal — towards which CL seems instead running — which intrinsic nature is an artificial neural network given?

In *The Human Use of Human Beings*, Norbert Wiener wrote that «cybernetics takes the view that the structure of the machine or of the organism is an index of the performance that may be expected from it»[6]. The ant unavoidably behaves rigidly, our brain being instead neotenus and plastic. Now, if ML itself might be deemed as a family of meta-algorithms, since optimization and back-propagation mean programming program-

ming<sup>3</sup>, CL contemplates in fact a sort of second-order meta-algorithms trained, so to say, on sequences of tasks rather than of data. However, how many butterflies can sprout from the same chrysalis? Metamorphosis will be hampered by oblivion, at least up until models' parameters — the footprints of correlations among data — and data itself will be deemed as inert objects, little blocks, or little bricks. Inertia is the main property of hard things, say, at the opposite of consciousness, often meant as a continuous stream.

All is not lost, though. Neural networks do work — obscurely, especially when they are deep — but they do. Now, let us linger for a moment not so much on data and parameters (indeed, every activation of a previous layer might be seen as new data by the next layer within a neural network) but more precisely on information, *i.e.* something much more unworldly. Rephrasing an old expression taken from the *De Contemptu Mundi* («On Contempt for the World») by Bernard de Cluny, *stat rosa pristina numero, numeros nudos tenemus*<sup>4</sup>. We can observe parameters, save them, wind the training process back, but they remain parameters — millions of numbers that are indecipherable to the human eye. Nonetheless, the mathematical function learned by a model *has a sense*, since it both represents a portion of reality *and* serves a purpose.

Therefore, two observations might be made. Firstly, information flows forward and backward *wildly* and it is difficult both to localize "meanings" throughout the network and to forecast how they will be distributed (this seems particularly evident in the case of the *de facto* standard backpropagation algorithm [7], where all the neurons in the network are updated jointly, spreading the information associated to the current learning iteration). Secondly, we find traces of plasticity in a neural network, but as previously mentioned catastrophic forgetting teaches us that such kind of spontaneity is, as an oxymoron, inertial.

Some disagreement might arise, concerning the missed localization of information. In the CL literature there are methods that assign each specific task to a separate group of neurons — the architectural approaches [8]— implementing, at first glance, a clear distinction among the different representations. Nonetheless, this family of approaches simply shifts the issue forward: in such methods, it is not straightforward how to combine these sets of neurons into a single, autonomous network. This is confirmed by the fact that a task label is often required, specifying which network to use for each sample in input. Once again, this for sure represents an easy solution, but it does not solve the problem of *really* learning incrementally.

We observe a sort of informational hostility to change, a somewhat incorporeal hardness while the universal approximation theorem holds and, with sufficient parameters and time, every continuous function defined on a closed interval can be arbitrarily approximated by a combination of simpler functions [9]. Memory, here meant both as data and stored parameters, seems on the other hand to be *fluid*, *i.e.* both being ontologically a flux instead of a bunch of things, and bearing that property of fluids, according to which they distribute themselves, even if not with homogeneous density, all along their container. Change the shape of the glass, and you will also reshape the water. It might be objected that, strictly speaking, there is *no* container, otherwise said that «the medium is the message»[10]. That's a point or, better, that's *the* point, going way far and deeply lying

<sup>3</sup>not a typo!

<sup>4</sup>«The original and first rose lies in the number. We only possess naked numbers». *Stat rosa pristina nomine, nomina nuda tenemus*, was the initial phrase («The original and first rose lies in the name. We only possess naked names»).

in the very concept of the enforced mathematics. After all, information is a *log-arithm*, literally, *a word becoming a number*.

Another counterargument might be the sole existence of Convolutional Neural Networks (CNNs) [11]. Striding, padding, and repeatedly filtering data, CNNs surely put in evidence particular properties — like the shape of a wolf's tail, or the colour of a beautiful pair of eyes. Here, even if not a "grandmother cell" [12], one might nonetheless argue in favor of meanings' localization — what once would have been called *universalialia in re* — truly to the point that algorithms like GradCAM [13] or GradCAM++ [14] extract the gradients precisely from the last convolutional layers in order to find some local interpretation. However, beware: information might still be in the eyes of the beholder.

Just bringing up a *very* trivial case, take these two phrases: «Today is a beautiful day»; «To be or not to be?». A banal everyday sentence might even require more memory, more bytes than the nihilist Hamletic doubt, but nobody could even dream of deeming the Shakesperian expression less significant. What is meant here, is that the only fact that some features get (even with some precision) extracted does not necessarily entail the extraction of meanings. It unpredictably depends both on data and on the model's structure: was it *so* manifest, for example, up until the success of the attention mechanism, that a phrase is not simply a sequence of token?

### 3. Remember me, forget me, write me again

Recollect your own experience, taking a quiet pause. How much have you forgotten, how much do you still remember? When you were a child, for example, you learned a specific way of walking, a *technique du corps* [15] you cannot alter anymore. However, undressing from your own skin due to loss or bewilderment, you sometimes crossed the time and forgot the past. Indeed, despite often being lived with anguish, oblivion represents, within a certain margin, a necessary experience for the proper working of memory [16]. Why, if — let us say, *natural* — memory is not an incremental and cumulative process, should its imitation be so? Not to mention consciousness, our nervous system is structurally far more complex than a large artificial neural network, being multilayer in quite a different sense. Neural networks forget too, but clumsily: from memorizing everything, they no longer remember anything (relevant). A bit of forgetting might mean no catastrophic forgetting. However, how to teach them to forget properly?

Especially having to do with large, nonlinear models might induce us to believe that information be shared among parameters according to some harmonic entirety, rather than to some assemblage of meanings. Strangely enough, it resembles what ancient Stoics called *to pneûma*, «the blow». Provided that, by «fluid memory» we mean the emerging of (purportedly, let alone unexpected patterns) *one sense*. Indeed, all in all, one is the network where every neuron seems to find its place. Differently from meaning, sense can be touched, *i. e.* loss can be overall reduced through optimization, but not explicitly described. Could this viewpoint help us to interpret the phenomenon of catastrophic forgetting?

Suppose you have a piece of polymeric matter. It has some chemical and physical properties, scientifically observable. Yet, strictly speaking, except for the possibility to take almost infinite shapes, it is useless. Now imagine that, since you need to fabricate a hammer out of it, you fuse the polymer and, while fused, you give it your preferred

shape. After waiting for it to cool down, you finally get a hammer. Now you grasp the grip of your hammer and drive a nail. Not surprisingly, you did that by your hand and not, say, by your foot. Now it is useful, *now* it has a sense.

Unfortunately, you find out you need *also* a washbowl (yes, a washbowl!). Since nothing else is available, you give a look at the hammer and think: «I need to make *also* a washbowl out of that». So as to spare time and energy, you decide *not* to fuse your polymer again, but just to warm it up. You try not to heat up the entire body of the tool, to round and hollow out this and that surface... but today luck is not on your side and you just obtain a misshapen polymer, neither an excellent hammer nor a good washbowl.

Suppose that, at time  $t$ , you initialize a neural network randomly — *i. e.*, keeping up with the metaphor, a fused polymer which in addition is partially capable of self-resaping — and train it in a supervised manner on a task  $a$ , say, the classification of pictures portraying dogs and cats. Random weights *do* implement a mathematical function, which in itself does not need any adjustment and does not make any difference from what will be learned up to the last training epoch. However, what is missing at the beginning of the training procedure is something ulterior, say, the core of supervision: an ambiguous mixture of a qualitatively significant representation of the world inscribed in the data distribution, and a precise practical goal to realize.

At time  $t + 1$ , you would like the model to experience a different data distribution according to a task  $b$ , say, *in addition to task  $a$* , distinguish pictures of airplanes and cars. You carefully select the learning rate, choose an appropriate optimizer along with its hyperparameters, and employ your favorite CL algorithm. Yet, despite your efforts, a performance drop ultimately comes up. Is not it, perhaps, that it has been attempted to violate some at this stage well established overall and purposive balance? Is not it that you are hoping for two butterflies to emerge from a single chrysalis?

#### 4. Music for the neurons

*«Now the thirty-oared ship, in which Theseus sailed with the youths, and came back safe, was kept by the Athenians up to the time of Demetrius Phalereus. They constantly removed the decayed part of her timbers, and renewed them with sound wood, so that the ship became an illustration to philosophers of the doctrine of growth and change, as some argued that it remained the same, and others, that it did not remain the same.»*

(Plutarch, *Parallel Lives*) [17]

By learning and forgetting simultaneously, we have come across the ancient enigma of "the ship of Theseus": is it *always the same* model which we are dealing with, when performing continuous learning? Despite us being *catastrophically* pushed to reply: «neither it is, nor it should be», this holds only as long as the components of such an entity — be it the Theseus' ship or a neural network filled with memories — were strictly considered a collection, paradoxically countering the brilliant idea of reusability. We might indeed find clues to the opposite direction: among others, the wonder — still studied and from someone considered illusory [18] — of emergent abilities [19]. Resting the eyes on in-context learning, a fascinating hypothesis of implicitly learned meta-gradients has been put forward [20], while some attention has been placed on the semantic texture of data [21].

From the point of view of a model that is accustomed to taking photographs of data according to a uniform reality called «ground truth», CL will resemble beholding the landscape from a moving train. Time will go by bringing with it itself a sequence of experiences shifting in the data distribution but suddenly, continuously running away, leaving just a halo. Not only, then, would it be hard for it to *accumulate* memories, but it also will not strictly manage sequences of *frames*, even with brutally changing distributions. Tasks being similar to notes in a melody, a musical analogy appears to arise. What if CL were meant as a variation on a theme and optimization alluded to a consonant blend of chords and voices?

The very difficulty lies in optimizing models elegantly serving a fluctuating purpose, *i. e.*, neither simply juxtaposing memories nor orienting learning in an univocal way. Such a dynamic and "permanently unbalanced" AI, intrinsically multivocal, where training and inference phases would be superimposed, would perhaps be even in the long run a chimera — or, more simply, a brain. Yet, until then, an alternative solution should be found to brute-force algorithms which devour tons of data. Say, besides feeding forward and back-propagating, there should be also space for moving left and right.

A similar reasoning is expressed as the *collectionless principle* [22]. As the name suggests, it represents an alternative point of view on subsymbolic AI, which in essence revolves around the idea of disregarding large data collections — the ones needed by every current ML algorithm — and let instead the agents process the environmental stimuli and learn from them in a truly online fashion.

Now, this properly describes a possible way to give artificial agents the instability we mentioned before, the one they intrinsically require to learn dynamically. In a scenario with strong relations, both human-to-agent and agent-to-agent, one way to get out of the riddle might lie in the accurate choice of the data with which to feed the model, so as not to go "off-theme". It would be possible to "*tune*" the models — either metaphorically and not — providing the adequate data points for each instant in time.

Furthermore, if we aim at distinguishing — still, *not* rigidly localizing — the sub-functions that compose a labyrinthine global function just like we discern trumpet and piano throughout a rhapsody, optimization should be meant as a sort of orchestration. As a result, we imagine that this kind of *collectionless* agents could easily communicate and cooperate over time, benefiting from the unique experiences the other entities had, overcoming the single monolithic models supremacy.

## 5. Conclusion

What can Continual Learning become? In this work, we advocate the development of a new ML paradigm, that could effectively support the continuous adaptation of an intelligent agent in a dynamic environment. We imagine an net of endless interactions among agents and humans, all harmonized on the same theme. At the same time, we hope to surpass the need for gigantic and centralized models, the ones that, besides a superficial smart behavior, are indeed a great example of inefficiency and stupidity.

Being prompted with the same data *over and over and over*, practically having the entire human knowledge at disposal, to finally say something reasonable; is this *intelligence*?

## 6. Acknowledgments

Work supported by PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", funded by the European Commission under the NextGeneration EU programme.

## References

- [1] Nietzsche F. On the genealogy of morals. Translated by Walter Kaufmann and R.J. Hollingdale. Ecce Homo. Translated by Walter Kaufmann. Vintage Books: New York; 1967.
- [2] McCloskey M, Cohen NJ. Catastrophic interference in connectionist networks: The sequential learning problem. In: Psychology of learning and motivation. vol. 24. Elsevier; 1989. p. 109-65.
- [3] Lesort T, Lomonaco V, Stoian A, Maltoni D, Filliat D, Díaz-Rodríguez N. Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges. *Information fusion*. 2020;58:52-68.
- [4] Lesort T, Caselles-Dupré H, Garcia-Ortiz M, Stoian A, Filliat D. Generative models from the perspective of continual learning. In: 2019 International Joint Conference on Neural Networks (IJCNN). IEEE; 2019. p. 1-8.
- [5] Rebuffi SA, Kolesnikov A, Sperl G, Lampert CH. icarl: Incremental classifier and representation learning. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition; 2017. p. 2001-10.
- [6] Wiener N. The human use of human beings: Cybernetics and society. 320. Da capo press; 1988.
- [7] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *nature*. 1986;323(6088):533-6.
- [8] Rusu AA, Rabinowitz NC, Desjardins G, Soyer H, Kirkpatrick J, Kavukcuoglu K, et al. Progressive neural networks. *arXiv preprint arXiv:160604671*. 2016.
- [9] Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. *Neural networks*. 1989;2(5):359-66.
- [10] McLuhan M. Understanding media : the extensions of man. New York: New American Library; 1988. Available from: [http://www.worldcat.org/search?qt=worldcat\\_org\\_all&q=0451624963](http://www.worldcat.org/search?qt=worldcat_org_all&q=0451624963).
- [11] LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*. 1998;86(11):2278-324.
- [12] O'Shea M. The brain: a very short introduction. vol. 144. Oxford University Press, USA; 2005.
- [13] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 618-26.
- [14] Chattopadhyay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE winter conference on applications of computer vision (WACV). IEEE; 2018. p. 839-47.
- [15] Mauss M. Les techniques du corps. *Journal de Psychologie*. 1936;XXXII:3-4.
- [16] Galanti MA. Smarrimenti del Sé. *Educazione e perdita tra normalità e patologia*. ETS, Pisa; 2012.
- [17] Plutarch. *Parallel Lives - Complete*. Start Publishing LLC; 2012.
- [18] Schaeffer R, Miranda B, Koyejo S. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*. 2024;36.
- [19] Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, et al. Emergent abilities of large language models. *arXiv preprint arXiv:220607682*. 2022.
- [20] Dai D, Sun Y, Dong L, Hao Y, Ma S, Sui Z, et al. Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers. *arXiv preprint arXiv:221210559*. 2022.
- [21] Perri EF, Grande E. Ghosts in the AI. 2024.
- [22] Gori M, Melacci S. Collectionless Artificial Intelligence. *arXiv preprint arXiv:230906938*. 2023.