Review of: "Self-Replication, Spontaneous Mutations, and Exponential Genetic Drift in Neural Cellular Automata"

Ettore Randazzo[1]

[1] Google Inc.

Potential competing interests: No potential competing interests to declare.

Full disclosure: I am one of the authors of Growing NCA and several NCA-related papers.

I found this paper very interesting with some novel insights in open-endedness and self replication. Seeing NCA model rules as the physics of the world and the internal states as their DNA opens up an exciting amount op possible ‘worlds’. The author discusses appropriately the quirk of Deep learning to tilt the space of possible worlds towards one with many attractors but also discusses methods that can be used to soften this trend.

I have a few concerns that can all be addressed which I think would really help in making the reading experience better for the reader.

There is one major mistake in the explanation on how growing NCA work which needs to be addressed. See this section: “The Alpha channel determines whether a cell is alive (Alpha>>0.1) or dead (Alpha≤≤0.1). If the cell is alive, the update rules apply: the cell's state is recalculated by applying the neural network to the neighborhood of the cell. If the cell is dead, the state vector is reset to 16 zeros and no update is applied.” This explanation is slightly incorrect, but if NCAs worked this way, no cell would ever switch from being dead to being alive, since their update rules would always be masked. What actually happens is that we also have a ‘growing’ state for otherwise ‘dead’ cells: if your alpha is <0.1 but you are nearby an ‘alive’ cell, you *can* update yourself. This makes it so a cell can transition from dead to alive.

One issue I found was with following the storyline of this paper. I think that rearranging topics would drastically improve on the clarity of the text. At the end of the introduction I believe that it would be very helpful to spend a few sentences hinting at what kinds of experiments will be done. This is particularly important because it is not implicitly understood at all by the reader about what kinds of self-replication, and mutation, experiments *can* even be done with NCAs. Speaking of which, the discussion about genetic drift appears *before* we know what kind of self-replication we are going to observe, and what is a DNA in this context. I strongly suggest to discuss the setup of self-replication that we will observe first, explaining what DNA means there, and then talk about genetic drift. In the way it is described now, only in the results section we start to really understand what we mean by self-replication and mutation. I would also suggest stressing out the difference between your first experiment setup and the Growing NCA setup: in growing NCA, a ‘seed’ is a single pixel cell, and the target is one image. In your first experiment, a ‘seed’ is a fully grown image, and the target is two of these images. Likewise, it appears that the seed of the following experiment is more than one pixel wide, which are all important details for understanding this paper.

There is one minor inaccuracy in the explanation of the training regime of growing NCA: "NCA only require the experimenter to define an initial state, a target state and a maximal number of computation steps". In truth, we always use some pooling mechanisms (whereas you perform batch substitution for similar effects) that unrestrict the maximum number of steps that a pattern can survive in, and we tune the replacement rates and pool sizes to make sure that a true stability of the pattern is observed.

Several of the topics you discuss have significant overlap with a paper of mine (disclaimer)Recursively fertile self-replicating neural agents, which I think may help in grounding the observed results with some more background. In that paper, I show how neural networks quines (self replicators) are inherently divergent and there is very little you can do about it unless you have some extra fixed structure that allows to create some basins of attraction, and performing some sort of noisy training. Then, I also show that to have some variation in some images that are capable of self-reproduction, we need multiple images, similarly to what you perform in one of your experiments. In the NCA context, the ‘fixed structure’ are the NCA rules (the physics of this world), and the self-replicators are working on the internal states. Even here, the fact that during training you replace seeds and continue training have the effect of adding noise and creating a basin of attraction, where otherwise models would inevitably diverge. This paper has some interesting discussion that informs my comment on the next paragraph about open endedness.

You state: "While our experiments satisfy the definition of unbounded innovation and unbounded evolution by Adams et al. (2017) and arguably manages to implement changing biological rules as a subset of fixed physical laws", and then also fairly describe how the system is very limited in what it can achieve. I think that unbounded innovation and evolution are explicitly vague phrases so that these kinds of experiments would still *not* fall into this bucket. After all, a white noise generator would have the maximal unbounded innovation if we were just interested in genetic drift. What we observe in this paper and on the ‘recursively fertile self-replicating neural agents’ paper is that we are able to generate something with some small variations on a manifold (in my paper I could create flowers with arguably infinite different colors, but still always flowers), which makes our results very much ‘bounded’. Still, I agree with the conclusion that having several hundreds of target patterns or similar methods could unlock true open-endedness.

I saw some typos. The ones I observed: "Ran- dazzo et al. (2021)" has a formatting error; "in the process tThe training" I suspect was a parenthesis gone wrong.