A human-machine co-authoring experiment

by Rodolfo Ocampo

On February 2019, OpenAI made headlines by publishing a lengthy post to explain the reasons behind the decision to not release their new language model. The text generated by it was so human-like that the risk of potential malicious uses was too great. Instead, they released a lightweight version which still produces unprecedented quality synthetic text. The code for the released model is available on GitHub.

I used the model to write the essay on the next page. I provided a first paragraph as prompt, and the model generated a second. I then generated a third, and fed it again as a prompt. I did this several times. I published the code used to do this experiment in a Jupyter (Google Colab) notebook so anyone can try their hand at human-machine co-authoring right in the browser. It can also be used to replicate my results.

(H) = Human
(C) = Computer

A Techno-anthropological Essay

(H)
This essay is an exploration of the relationship between humans and technology in two senses. The first one is literal; the second one is a meta-exploration. The text itself was written through a collaboration between a text generating algorithm, in particular, the GPT-2 model created by OpenAI and not released in its entirety due to the societal risks associated with it. The fact that this model is used in this essay is paradoxical since our main concern is answering the question: can the development of artificial intelligence, and cyber physical systems be regulated? If so, how?

(C)
The first dimension of the relationship is the one that is most important. We are discussing the relationship between humans and technology in the most fundamental sense, that is, the relationship between human beings, which is, at the same time, technology.

(H)
If we look at the history of our civilization, we see that our very first tool is cooperation facilitated by language. This technology explains the rise of our species and serves as the ground from which all other technological creations have stemmed out. All modern technology is merely an extension of social interaction, the most fundamental tool of all.

(C)
To understand this, we must look at the history of the evolution of language. The development of language as we know it started in the Stone Age. The first humans, the hunter-gatherers, used to hunt and gather their food. They were not very good at this task, and they needed something better.

(H)

By using vocal sounds, which later transformed into language, they were able to transmit useful information for hunting. But what is interesting for us is that this new technology was influenced by the underlying dynamics of the human groups that developed it. So for example, when we look at the dangers posed by an algorithm like the one co-authoring this essay, we see that they emerge from the underlying context of the society that develops it. One way in which GPT-2 could be misused is the generation of fake news at scale or the deployment of armies of social media bots to influence elections. These scenarios are plausible because there are social flaws that generate incentives to attack democracy. Another example: language models usually perpetuate biases from the training data, which are mere reflections of real biases in society. Thus, the dangers of cyber-physical systems arise from social phenomena. In order to prevent them, we need to understand them, and a technical lens is not enough: we need social science.

(C)
So, what are some of the most important aspects of understanding human behaviour? As a first step, we need to understand how people behave. We need to understand how people act in a given situation. For instance, how are men and women different, and how do they behave in different circumstances? How do people relate to other people? How do they perceive the world? These questions are central to understanding the social order and how to manipulate it. Then we need to understand how people are related to each other. What do we find when we compare men and women, how do they behave when they interact with each other, how do relationships change when people are older than 50, how do people relate to children? It turns out that the dynamics of human interaction are not limited to the individual. It is not just individuals that experience these relations, but also groups, organizations, and nations.

(H)
Once we model these interactions, we can begin to study which forces will be at play and what sort of intentions can be embedded in artificially intelligent systems. With that, we can

then build a set of risks and dangers presented by the intersection of new technological tools and social incentives. But what about the actions to prevent them? Who will be in charge of enforcing regulations, if anyone?

(C)
And what will the consequences be? The answer, in this case, is that each of these technologies will have to be regulated individually, with each new system designed according to an individualized risk assessment that considers the potential harms and benefits associated with each technology. In other words, regulation by autonomous systems will be based on a mix of human judgment and algorithmic determinations.

(H)
This is why we need a new applied science for enabling safe and ethical design of cyber-physical systems. There is not a one-size-fits-all formula for this task. The breadth of technological tools that will emerge in the near future will require a structured approach to the problem and frameworks that enable a scalable method of answering the questions that will enter the public discussion. Moreover, a scientific approach will guarantee that the answers provided are based on the best available evidence.

(C)
The application of this approach will not only be useful for the design of safe and ethical devices, it will also help to identify the most beneficial applications of cyber-physical systems. In the past, the process of application of science has been a process of evaluation and refinement. A scientific approach will be both more transparent and more effective.

There is another fundamental question that confronts society. What is the ideal state of cyber-physical systems? The ideal state is the state where, in the absence of any other means, all persons can be reasonably assured of personal privacy and security. The ideal state of cyber-physical systems is one in

which privacy and security are guaranteed both in the physical world and in the virtual environment.

(H)
It is important that we bring this topic up before we end, since the concept of an ideal state should serve as guide to the new science. As with any other endeavor that seeks to push humanity towards a more desirable state, it is important to internalize the fact that the ideal, which serves as a north star, is itself unattainable. The cyber-physical ideal of privacy and security both in the virtual and physical worlds is no exception. There exists a compromise between the two and one of the fundamental choices that a society has to make is where to stand between them. More security or more privacy? Are modern societies equipped to make this choice in a world where the exponential rate of technological progress will make it almost impossible to estimate the impact of any given decision? The answer is no.

We need a new science.

Notes:

The model usually outputs several lines of text. To stick to my idea of intercalated paragraph generation, I defined how many sentences of each computer output I was going to keep each time.

I also had to make small edits to my own text afterwards since I realized a few typos. In order to recreate my results (that is: the computer generated texts), the human texts need to be entered with typos (which I omit here but provide in the notebook).

I added some line breaks and added some spaces to the computer text for the sake of readability.