A View from Outside the Loop

Anders Hedman[1]

[1] KTH The Royal Institute of Technology, Stockholm, Sweden
ahedman@kth.se

Abstract. There is a growing interest today in how to combine human intelligence with artificial intelligence in the best possible ways. One reason for this interest is that in this territory of combined intelligence it is, in many cases, unclear how the total system of human and machine will behave, and unless we know that, how could we know what the perils and opportunities might be? It is clear then that we need a body of research to investigate the nature of humans in the loop in order to design wisely. Such wise design would prima facie appear to be achievable through analysis of existing and possible human in the loop systems from a vantage point outside of such systems. But, is such a vantage point achievable today and if so, will such a vantage point be available in the future? This paper considers the possibility of humans as always being in the loop and what it might mean for our understanding of humans in the loop.

Keywords: human in the loop, AI, HCI.

1 A Perspicuous Loop

1.1 The Pervasive Loop

French renaissance enlightenment philosopher Denis Diderot [1] noticed a peculiarity about our human mental life, it is essentially recursive and not merely reflexive. In reflecting on mental life Diderot came to the conclusion that the brain is a book that reads itself. We have to use the mind to understand the mind and this situation is very different from using the eye to study the eye. This is because to some extent what we mean by the word mind is a construct. The term "mind" didn't come into widespread English usage until around the time when John Locke started to use it systematically in his 1690 *An Essay Concerning Human Understanding* [2]. We still lack direct cognates for the word mind in other languages and must as Locke had to, still improvise when translating the word "mind" to French, German and other languages.

As we try to understand how human intelligence might be combined with artificial intelligence in a loop of interaction do we run into an extended version of Diderot's book that reads itself? Has the book that reads itself become a hybrid of human and machine? Let us look at some examples. Whenever I type text, I get auto suggestions for words I can use. Let us suppose I am a novice and a poor writer; I pick words that I would not have chosen without this algorithmic help. I search for something online and search words are suggested to me there too and again, I might not have thought of them

myself. My searches are automatically limited in accordance with my search history. If I ask myself what exactly I do online or with interactive technology that does not involve being inside a loop with AI I have to admit that I don't know. What I do know is that whatever I do could potentially be part of an AI loop and many of my actions are.

To have a point of view outside of the loop would mean that one would be certain that one was not affected cognitively by a persuasive or otherwise "helpful" AI system within the activity performed. Within our modern society such states are as we have seen through our mundane examples often not available. It seems rather that the normal state is to be in the loop. The question then of how to analyze and understand the human in the loop is therefore one that naturally is tackled from within the loop. What does this mean for our understanding of the human in the loop then? To better answer this question let us imagine a thought experiment.

Imagine that human-computer interaction advances to the point where suggestions for words that I type come not on my screen but are inserted into my visual field through some direct brain interface. I see the alternative word and can pick it automatically. Since the word appears so transparently, directly in my visual field, would I still be aware of it? What if the word instead simply entered my thought stream? Perhaps I could sense that the word came from outside of me or perhaps I couldn't. If I couldn't sense that the word came from outside of me then I would have no way of knowing if I was in a loop or not. The example here is a simple one of a very mundane task, but we can also imagine variations of it so that the thought insertions come in a broad variety of situations.

1.2 Lost in the Loop

We can imagine a future were we humans cannot step out of the loop because we no longer know where our thoughts come from. At this point, the philosopher who brought the word mind into common parlance would say that we have lost ourselves. For Locke the essential feature of a person was a unified consciousness that owned its mental contents throughout time.

In 2003 B. J. Fogg published his seminal work on persuasive technologies [3] on the basis of an idea that he had during a course on mind control with the mind control expert Philip Zimbardo. In that work [3] it appears that Fogg has a clear understanding of how persuasive technologies could be used for mind control, for he explicitly says that persuasive technologies are to be used for persuasion and not for coercion. Yet, it is unclear how AI driven persuasive technologies in, e.g., social media have been used for persuasion on any significant scale. It is clear instead that they are routinely used for coercion on a large scale and this has led some scholars to ask for a redefinition [4].

In a world of AI driven persuasive and other "helpful" technologies we may well ask how far we are from losing ourselves in the loop? We may also ask what the future may hold? Is one of the biggest threats to humanity really malevolent AI in the shape of a transcendent race of machine intelligence as the late Stephen Hawking suggested [5] or is it another, more subversive kind of AI, the kind which leaves us with no view outside of the loop?

Much of the discussions of today regarding the human in the loop appear to deal with ethical risks and challenges of using AI systems that are not part of the coercive artillery that quietly bombards us in our daily lives through technologies labeled as persuasive. This AI artillery poses ethical challenges that ought to be further discussed. But what exactly is the problem? It is not only that we may be coerced by AI and that we may with time develop cognitive problems. It goes deeper than that.

The philosopher Nick Nozick once designed a thought experiment called the *experience machine* [6]. Nozick images a future when we can choose to hook ourselves up to a machine that will give us all the experiences we would like to have. We can imagine that this machine gave us a perfect virtual reality world in which all of our dreams were fulfilled. If we would be bothered by the knowledge of being inside such a machine, then we can imagine that the memory of having hooked ourselves up to the machine was simply erased. We can also imagine other tweaks to our liking so that the machine becomes seemingly one without faults. Now the question is would you plug in to the experience machine? Intuitions differ. For those who don't want to plug in what matters most is that we live our lives authentically. In the light of Nozick's experience machine, the main problem with AI driven loops of persuasive technologies then is not necessarily that they will not help us get what we want, but that they pose an existential threat of losing control over our lives as authentically lived.

1.3 Making the Loop Perspicuous

What can we do then to retain a view from outside the loop or at least one that makes it perspicuous to us? It seems to me that a first step is to stop calling AI driven technologies for coercion, persuasive. It ought to be clear by now that if indeed Fogg believed his work would lead to a world of persuasion that dream failed. We live more in a world of coercion and not surprisingly perhaps, the project of persuasive technologies that came out of a course on mind control, has led, in the mains, to mind control and that is how we ought to refer to persuasive technologies that coerce: as technologies of mind control. We ought to reserve the term persuasive technologies for those technologies that, like humans offer us arguments that we can consciously entertain and may or may not be persuaded by. It is part of the notion of persuasion that we are offered such arguments. If we continue on the path of today, we may well find ourselves one day interacting in coercive loops that we cannot detect and loose touch of who we are in the Lockean sense of being conscious continual owners of our mental contents.

A second step would be to realize that any discussion of AI and the human in the loop depends in turn on how we define ourselves and our mind. In Diderot's time it made sense to refer to the brain as a book that reads itself (Diderot didn't use the term mind as it was unavailable to him in French). Today, we have since the inception of cognitive science a perspective on the human mind with roots in cybernetics and information processing. If we say that our best construction of the human mind is as an information processing machine as in cognitive science or in more general terms, a functionalist machine, as in the philosophy of mind, that is nevertheless a construction undergoing a cultural evolution. In a cognitive science and philosophy where no one knows what a thought is or what consciousness is, we are continually reinventing the

mind as at least partly constructed. If we see ourselves as essentially different from AI systems along the lines of the critics of AI such as, e.g., John Searle and the late Hubert Dreyfus then this will naturally color the debate over the human in the loop with humanist and existentialist overtones. If on the other hand, we construct AI as part and parcel of an accepted cognitive science and as being near a point of transcending human intelligence then we will naturally color the debate very differently. If we believe, for example, that there is in principle no difference between human intelligence and machine intelligence then our world becomes much simpler than if we acknowledge with the critics of AI that human intelligence is essentially different. This debate is not solely over the question of intelligence but also about existential questions about how we wish to live our lives.

References

1. Garrett, A. (2017). The Routledge companion to eighteenth century philosophy, p104. London: Routledge Taylor & Francis Group.
2. Locke, J. (1970). An essay concerning human understanding: 1690. Menston: Scolar Press.
3. Fogg, B. J. (2003). Persuasive computing: Technologies designed to change attitudes and behaviors. San Francisco, Calif: Morgan Kaufmann.
4. Kampik, T., Nieves, J. C., Lindgren, H., & 20th International Trust Workshop, TRUST 2018. (January 01, 2018). Coercion and deception in persuasive technologies. Ceur Workshop Proceedings, 2154, 38-49.
5. Hawking, S., & Whishaw, B. (2018). Brief answers to the big questions.
6. Nozick, R., & Nagel, T. (2013). Anarchy, state, and utopia. New York: Basic Books.