Existential Risks of Co-Creative Systems

Maria Teresa Llano and Jon McCormack
Department of Human Centred Computing
Monash University, Melbourne, Australia
Teresa.Llano@monash.edu

Consider three possible futures in which co-creative systems become more creative than humans: i) human-AI interactions with co-creative systems become better than ever known, ii) heavy dependency on machine creativity diminishes our ability to think creatively, or iii) machines, working in collaboration with humans, devise a creative way to end humanity. Many consider scenario *iii* only possible in movies (at least in the near future); however, with technology increasingly evolving and playing more important roles in our lives, scenario *ii* seems like a more plausible concept. We argue here that in order to avoid these and other dystopian scenarios, we need to consider what kind of risks advanced co-creative systems will represent to human civilisation over the coming decades and formulate strategies to avoid them.

In line with the growth of AI systems as creative collaborators different ways of interacting with co-creative systems are being explored. Some theories suggest that human-machine collaboration requires a very close coupling (Licklider 1960) to the point in which the boundary between both will be blurry (Lopes et al. 2015). The focus is then shifted from interaction to integration and on the idea of a *new integrated self* (Mueller et al. 2020), where computers and the human body are combined into one in order to achieve greater collaboration – an idea explored extensively by transhuamism.

This growth in the field suggests great potential for the development of co-creative systems, but it also brings concerns that advanced approaches to co-creation with AI systems might represent a threat to humanity. The existential risks of AI, which represent threats to the ongoing existence of humanity and human culture, are being raised by a number of scholars in the field (Bostrom 2014; Ord 2020). The search for an Artificial General Intelligence (AGI), an AI whose intelligence potentially far exceeds our own, is what researchers have identified as the most plausible existential risk of AI. A 2016 survey of over 300 leading machine learning researchers estimated a 50% chance that AI systems will be "able to accomplish every task better and more cheaply than human workers" by 2061 and a 10% chance by 2025 (Ord 2020, p.141).

We believe that this discussion is ever more pertinent for the future of co-creative systems. Although there are more present threats to humanity (such as pandemics and climate change), the increasing reliance on technology and the closer coupling between people and computers being sought in the field makes it crucial to explore the potential (long-term) risks of developing co-creative systems. Speculating on possible risks before they are technologically possible is always problematic, but given the significance of the threat, highly important to consider. Additionally, thinking about possible threats *before* they occur makes it much easier to mitigate them by design.

Diminished Human Creativity

With increasing reliance on technology our ability to think creatively may be diminished by advanced co-creative AI systems. In particular, the engineering of automation may pose an implicit risk to diminishing human creativity (McCormack 2019). A simple example is the 'smile detection' function on many modern digital cameras. With this 'feature', the camera makes the decision when to take the picture based on it detecting everyone in the frame smiling. While simple and seemingly benign, subjugating creative decision making based on cultural norms reduces the capacity for individual difference and simply reinforces the norm – the opposite of inspiring creativity. As AI technologies become more powerful this automated decision making could easily become highly prevalent, to the point of not only making decisions on what images to capture and distribute, but to automatically modify them based on cultural norms (automatically removing blemishes, modifying body shape, skin colour, removal of elements from a scene, and so on).

A more speculative example is that of an *integrated painter* who has been "enhanced" with electric muscle stimulation to manipulate drawing actions in order to co-create with a human while drawing – an actual plausible scenario as developments of the underline technology are already in place (in a less creative form), see for instance the MusclePlotter (Lopes et al. 2016), a system that manipulates a user's wrist in order to write or draw from instructions given to it (e.g. formulae, graphs, etc.). Initially, the partnership works well, the system complements, challenges and even inspires the human painter; however, the better the system becomes, the more it takes over control of the partnership. The painter develops a dependency to the system that makes him/her lose its own identity as an artist.

Creative Manipulation

Co-creative AI systems could also devise creative ways to disempower or manipulate humans to their own advantage. Even without a deliberately malignant goal, super-intelligent machines may accidentally threaten humanity as a by-product of a sought optimisation or originally benign goal. A popular machine-initiated creative idea may have unforeseen consequences if adopted widely. To be successful, systems that are explicitly programmed to be creative and work closely with humans must assist with or initiate the development of new creative ideas and artifacts. By definition, this involves finding the novel and valuable. But novelty and value alone are insufficient to ensure long term benefit or be ethical.

A classic, non-computational example is that of American chemist, Thomas Midgley, Jr., a man often referred to as the individual who "had more impact on the atmosphere than any other single organism in Earth's history." (Gilbert. 2019). In the early days of the automotive industry, knocking in engines was a major problem. Midgley's creative solution was to introduce tetraethyl lead (TEL) as an additive in gasoline production, eventually leading to hundreds of thousands of metric tonnes of lead being used annually in petroleum production and released into the atmosphere and environment (Nriagu 1990). The impact of this creative discovery is still being felt today, with lead implicated in numerous adverse health conditions, including cognitive development in children. Midgley later went on to introduce Chlorofluorocarbons (CFCs) for refrigeration, which were found to deplete the Ozone layer of the Earth's upper atmosphere, increasing the amount of ultraviolet radiation exposure on the planet's surface.

Co-creative systems could devise general creative solutions that humans couldn't imagine or anticipate (this is already the case in, for example, engineering design (Keane and Brown 1996; Layzell 2001; Hornby, Lohn, and Linden 2011)) and indeed it seems beneficial to encourage creative thinking that is beyond or outside the scope of human creativity, or for human-AI collaborations to innovate in areas that other forms of machine collaboration have been less successful. Current research into co-creative AI systems lacks an explicit ethical framework (the ethics are assumed to derive directly from the programmers and builders of such systems). Without this framework being made explicit to the the co-creative system, it has no knowledge of the ethical implications of its creative discoveries. Yet, as the Midgley example illustrates, even with an ethical framework some discoveries may have highly dangerous unforeseen consequences.

A Growing Partnership

Some researchers argue that new models of human-AI collaborations can provide exciting opportunities for transforming how people experience the world. Take for instance the emerging field of casual creators (Compton and Mateas 2015), a new range of creative technologies that allow users with different levels of expertise to engage in creative tasks by focusing on an enjoyable experience rather than in achieving a high quality final product. Or human-compatible technologies for interactive systems such as the Muscle-Plotter mentioned above, which allows the system to manipulate the user's wrist in order to write or draw. These type of models innovate not only in the way human and machines interact with each other, but also allow people to experience domains that have may be previously out of their reach (due to of lack of expertise, or because the user has some kind of disability that prevents him/her from pursuing more experience in the domain, etc.).

Nonetheless, most (if not all) of these systems are designed without an ethical framework driving the process. Take for instance co-creational systems designed to optimise for enjoyment, which carry the risk of dependency or, in the extreme, addiction. Video streaming services, for example, try to optimise screen time using machine learning algorithms to present content that the user wants to watch. However the algorithm does not differentiate on the ethical considerations of constant exposure to such content, or the personal situation, including the risk of addiction, of the individual viewer.

Advances in Co-creative AI can bring real opportunities for machines and people to work together, to build and grow human-machine partnerships. But they also bring an obligation for researchers to think carefully about the ethical foundations of these new technologies. Efforts towards more clear ethical guidelines are already being carried out by AI practitioners and others, in order to understand the ethical needs and requirements for AI systems. Recent surveys of AI ethics guidelines have tried to identify the main ethical principles discussed by researchers and practitioners (Jobin, Ienca, and Vayena 2019; Hagendorff 2020). Some of the most common topics include: justice and fairness, non-maleficence, responsibility and privacy, accountability, explainability, awareness and inclusion. Although these principles are relevant for AI in general, they lack a focus on what they entail for the context of creative collaborations. It is important then to extend these guidelines to the context of co-creative systems.

To illustrate, take the principle of explainability, which is often referred as the ability of an AI system to be able to explain its process and decisions. In a co-creative environment, explainability has also connotations of argumentation, where the aim of the explanation is not only to increase understanding, but also a mechanism to convince the listener of the acceptability of a standpoint (Llano et al. 2020). This extension in itself brings some ethical issues; for instance, how far should a co-creative system push for its ideas? How truthful should these arguments be? What constitutes a right or wrong argument?.

Conclusion

We need a common understanding of the challenges that co-creative systems may bring and an agreed consensus in the way to establish control mechanisms and ethical guidelines. The way we design our systems should reflect this. Values such as inclusion, social responsibility, explainability, and explicit ethical frameworks, should drive creative AI practitioners. Developing more active partnerships can be of great

benefit to humanity, but we want to make sure that the proper control mechanisms are studied and put in place, and that ethical and risk assessment considerations are taken into account.

Acknowledgments

We thank Dr. Alon Ilsar for proof-reading the paper and the SensiLab group at Monash University for helpful discussions about this topic.

References

Bostrom, N. 2014. Is the default outcome doom? In *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press. chapter 8, 201–213.

Compton, K., and Mateas, M. 2015. Casual creators. In Toivonen, H.; Colton, S.; Cook, M.; and Ventura, D., eds., *Proceedings of the Sixth International Conference on Computational Creativity, Park City, Utah, USA, June 29 - July 2, 2015*, 228–235. computationalcreativity.net.

Gilbert., S. G. 2019. Thomas midgley, jr.: Developed tetraethyl lead for gasoline.

Hagendorff, T. 2020. The ethics of AI ethics: An evaluation of guidelines. *Minds Mach.* 30(1):99–120.

Hornby, G. S.; Lohn, J. D.; and Linden, D. S. 2011. Computer-automated evolution of an x-band antenna for nasa's space technology 5 mission. *Evolutionary Computation* 19(1):1–23.

Jobin, A.; Ienca, M.; and Vayena, E. 2019. Artificial intelligence: the global landscape of ethics guidelines. *Nature Machine Intelligence* 1:389–399.

Keane, A. J., and Brown, S. M. 1996. The design of a satellite boom with enhanced vibration performance using genetic algorithm techniques. In Parmee, I. C., ed., *Conference on Adaptive Computing in Engineering Design and Control 96*, 107–113. P.E.D.C.

Layzell, P. 2001. *Hardware Evolution: On the Nature of Artificially Evolved Electronic Circuits*. Ph.D. Dissertation.

Licklider, J. C. R. 1960. Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics* HFE-1(1):4–11.

Llano, M. T.; d'Inverno, M.; Yee-King, M.; McCormack, J.; Ilsar, A.; Pease, A.; and Colton, S. 2020. Explainable computational crativity. In *Proceedings of the Eleventh International Conference on Computational Creativity, ICCC 2020*.

Lopes, P.; Ion, A.; Müller, W.; Hoffmann, D.; Jonell, P.; and Baudisch, P. 2015. Proprioceptive interaction. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI EA '15. Association for Computing Machinery.

Lopes, P.; Yüksel, D.; Guimbretière, F.; and Baudisch, P. 2016. Muscle-plotter: An interactive system based on electrical muscle stimulation that produces spatial output. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, UIST '16, 207–217. New York, NY, USA: Association for Computing Machinery.

McCormack, J. 2019. Creative systems: A biological perspective. In Veale, T., and Cardoso, F. A., eds., *Computational Creativity: The Philosophy and Engineering of Autonomously Creative Systems*. Switzerland, AG: Springer Nature. 327–352.

Mueller, F. F.; Lopes, P.; Strohmeier, P.; Ju, W.; Seim, C.; Weigel, M.; Nanayakkara, S.; Obrist, M.; Li, Z.; Delfa, J.; Nishida, J.; Gerber, E. M.; Svanaes, D.; Grudin, J.; Greuter, S.; Kunze, K.; Erickson, T.; Greenspan, S.; Inami, M.; Marshall, J.; Reiterer, H.; Wolf, K.; Meyer, J.; Schiphorst, T.; Wang, D.; and Maes, P. 2020. Next steps for human-computer integration. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, 1–15. Association for Computing Machinery.

Nriagu, J. O. 1990. The rise and fall of leaded gasoline. *Science of The Total Environment* 92:13 – 28.

Ord, T. 2020. *The Precipice: Existential Risk and the Future of Humanity*. Bloomsbury Publishing.