Kant and Digitalization. A Perspective.

Talk by Daniel Elon, Ruhr University Bochum, November 14, 2024

Abstract: This talk explores the relevance of Immanuel Kant's philosophy in the age of digitalization, providing a foundation for critical reflection and discussion. The presentation is divided into three main parts. *First*, it examines Kant's theory of objectivity, rooted in his transcendental philosophy, and its analogies to object-oriented programming paradigms in digital systems. The relationship between Kant's notions of cognition, object formation, and digital structures highlights the historical continuity between philosophical concepts and technological advancements. *Second*, the talk considers Kant's ethics in the context of digitalization. The discussion focuses on the implications of Kant's categorical imperative for ethical challenges posed by artificial intelligence (AI) and digital technologies. It addresses whether AI, as a potentially rational but non-autonomous entity, could fulfill Kantian moral principles or serve as an ethical decision-making agent. *Finally*, the talk emphasizes the enduring importance of autonomy, as central to Kantian morality, in navigating AI's role in society. While AI presents opportunities for innovation, Kant's call to "dare to know" reminds us that human reason and autonomy must remain the ultimate guides in moral and technological endeavors. By bridging 18th century philosophy with contemporary issues, the talk demonstrates the enduring relevance of Kantian thought in addressing the theoretical and ethical dimensions of digitalization.

My aim for today is to provide you with some various perspectives concerning the significance of Kant's philosophy in the age of digitalization. These perspectives may be partly subjective and not completely systematic, but I nevertheless hope that I can set some thoughts in motion and that I can provide a basis for a vivid discussion afterwards.

My talk will consist of three main parts: At first, I am going to take a look at Kant's significance for the recent processes of digitalization in a primarily *theoretical* sense: In this context, the main focus will be on Kant's definition of *objectivity* in his so-called "transcendental philosophy," and on how this may have shaped the way in which digital structures are designed. In the second part, I am going to consider some *ethical* aspects of Kant's philosophy in the context of the digital age. In this sense, I will briefly discuss the question on how the general historical changes caused by digitalization may, or may not, affect the validity of Kant's ethics, which centrally is an ethics of duty. In the third part, I will try to develop some thoughts

concerning the recent unfolding of AI, artificial intelligence, and in how far a Kantian perspective may be relevant and helpful with regard to future challenges in this area of technology.

But before, one important question that may come up has to be addressed: What does an 18th century philosopher have to do which digitalization, which apparently is a development of the twentieth and the 21st century? Isn't this talk just a fruitless attempt to renew an obsolete historical perspective in a completely different age?

So, let me get this straight: Digitalization is based on an extensive scientific and technological process in modern history and is not just a phenomenon of the last decades. It is now a generally accepted view that the first actual computer scientist was Gottfried Wilhelm Leibniz, the important philosopher and mathematician of the 17th and early 18th century: Foremost, Leibniz developed the binary system in the very version that is now the basis for every digital system. So, binary digits – "bits" – in our understanding as combinations of ones and zeroes are an invention by Leibniz. Furthermore, he built a relatively complex calculator – a mechanical and not an electronical one, though – for basic mathematical operations, and already had the idea of a purely logical formal language for universal scientific purposes. It could dive deeper into this topic, but let us just say that the beginning of modern computer science dates back to early modern philosophy and mathematics in their historical development. The idea of a formalization, an externalization, and an automation of mental and cognitive processes, like calculating or the general processing of data in a broader sense, has its roots in early modern rationalism.

It is thus neither improper nor fallacious to consider philosophical perspectives from past centuries when it comes to the big subject of digitalization. This brings me to Kant, so to say one of the big successors of Leibniz in the trail of German philosophy, and to the relevance of his theory of cognition and knowledge in this context. Although Kant followed Leibniz' rationalist philosophy in his early academic years, he later developed his very own, and groundbreaking, approach to philosophical questions. This is what Kant calls "critical" and "transcendental" philosophy: Instead of studying "objects," or "things" in a most general and realistic sense, we should first study our *way of perceiving* things by our senses, our understanding, and our reason. And, as a crucial premise, we should consider this insofar as our way of

perceiving and cognizing *precedes* actual experience. Hence, we have to examine your faculties of cognition in an *a priori* way.

This is what has also been called the *Copernican Revolution* of modern philosophy: Kant himself stated that philosophy has to undergo a fundamental shift of perspective. We should not assume that our cognition is determined by the things, but the other way round. This does not mean that our mind "creates" things like illusions – this assumption is far away from Kant's actual position. However, the very structure of the appearing world of objects is determined by the structure of our cognitive faculties. This also means that we do not perceive the things as they are in themselves, independent from our cognition, but only as they appear to us. And this form of appearance relies, as explained, on the structure of cognition.

On the one hand, this especially applies to time and space: According to Kant, they are no properties of the things in themselves, but subjective structures of the perception by our senses. This is what Kant calls "transcendental idealism," in the sense that the reality in its tempo-spatial form doesn't exist independently from a subject of perception. On the other hand, even the very basic concepts of the understanding don't apply to things in themselves, but are also our cognitive way of establishing and structuring a world of objects. This especially includes basic concepts like unity, reality, causality, and actuality, which all belong to Kant's famous "categories," or, interchangeably, pure concepts of the understanding.

In this sense, according to Kant, "objectivity" doesn't imply independence from the cognition by a subject. Objectivity is rather the very structure of reality that is constituted by the act of perceiving and understanding. This does *not* mean that our surrounding world is just subjective or imaginary. The opposite is the case: This world is empirically real, as Kant states. However, the content of the conception of objectivity is determined by the structure of our cognition.

So, this description should give us an overview of Kant's theory of cognition. If we now take a big leap and move on to the processes of digitalization in the last decades, we can become aware of striking connections and analogies: One basic concept of recent digital structures, and especially a common paradigm of programming these structures, is "object-orientation": This paradigm means that in establishing and developing digital environments, we should imitate our specifically *human* way

of structuring and arranging reality as such, that means as an interconnected, network-like system of *objects*, with specific properties, states, and behaviors. It is not hard to see that this directly corresponds to Kant's theory of cognition, in the sense that the formation of a reality of objects relies on the application and activation of the *a priori* concepts of our understanding. This is what Kant calls "transcendental logic": Simply put, the structure of the objective world is based on the core concepts of our understanding, this means of our *thinking*.

Programming and thus creating digital environments in an object-oriented way in fact conforms with Kant's theoretical philosophy concerning the basic structure of our understanding of reality. I would indeed suggest that the paradigm of object-orientation in the digital world can directly be derived from Kant's transcendental logic. This does not necessarily mean that the software developers of the 20th century actually read Kant's *Critique of Pure Reason* and took this book as an advice. I guess this is not the case. But the core concepts, as explained, clearly coincide, which still has to be explored and developed in future research in philosophy and computer science.

But for now, I will take another leap, this time into Kant's *practical* philosophy, which is of interest to perspectives of digitalization as well. This is now the second main part of my talk. Similar to his theoretical philosophy, the establishment of an *a priori* view, as explained, is crucial in moral philosophy as well. This is what Kant calls the *Groundwork of the Metaphysics of Morals*, a book from 1785. Here, Kant works on the establishment of a basic moral law, which ought to be valid independently from any external or personal circumstances, that means from any empirical influences. According to Kant, this moral law is the famous categorical imperative, which is supposed to be unconditionally valid for all rational beings. This law states that in every action, the *subjective maxim* of this act should, at the same time, be wanted to become a universally valid and thus *objective law*. Since we, as humans, are no purely rational beings, "vernünftige Wesen," but are also affected by emotions, feelings, subjective interests, and bodily needs, the categorical imperative only states how we *should* act, even if we generally don't do so. In terms of Kant's moral philosophy, it is our *duty* to accept the validity of the law anyway.

However, hypothetical beings without any affectivity, thus purely rational beings, would always follow the moral law without any exception. This would simply be their inner *will*, in contrast to the somehow corrupted human will. And this brings me to the crucial aspect in this context: It has repeatedly been suggested that an advanced and complex artificial intelligence, designed to decide on ethical issues, may eventually be considered to be such a purely rational being void of any "corruption" by subjective interests, emotions etc. One may even consider whether such an instance of artificial morality would have been welcomed by Kant as a means to enforce the validity *and* the actuality of the categorical imperative in the imperfect and somehow defective reality of human society. In this scenario, an advanced state of digitalization could provide humanity with a new, consequently *rational* morality that would finally be able to fulfill the Kantian demands of a "Metaphysics of Morals."

However, this somehow dystopian scenario misses one important fact: According to Kant, another crucial aspect of moral philosophy is the significance of *autonomy*: In the original sense, autonomy is moral legislation by oneself, meaning that the subject for which the law is valid is, at the same time, its own legislator. There can be no genuine morality without autonomy in this ethical meaning. Or, in other words, the categorical imperative is only valid insofar as it's *us*, rational human beings, who establish and deploy this law for ourselves, *as* human beings. Everything else would be heteronomy and thus no genuine morality.

At this point, it becomes clear that in Kantian terms, AI can never serve as a true moral instance of decision. It may be able to consult or to provide us with various perspectives concerning moral problems or dilemmas, maybe to get a better, more comprehensive picture of the morally relevant situation as a whole. But it can neither establish valid moral laws nor make actual decisions in critical situations.

This already brings me to the last short part of this talk, the possible relevance of a Kantian perspective for recent and future challenges in the field of AI. I also see the great chances and opportunities in the development of AI in so many different areas, in science, medicine, economics, education, and so on. Of course, caution is necessary, but we should also regard the development of AI as a huge technological accomplishment with a multitude of new perspectives in digitalization.

However, we always have to consider the boundaries and limitations of these recent and future developments, and in this respect, a Kantian perspective is extremely helpful. In his famous essay on the question “What is enlightenment” from 1784, Kant borrowed the words “Sapere aude!” – “Dare to know!”, from the Roman author Horace, and stated that, in the end, it’s all about using one’s own intellect and reason, in order to think and to act in true autonomy. So, to put it simply, according to a Kantian view, AI may help, but can never replace.