Impacts of new Internet applications and artificial intelligence on global energy demand – an issue of concern?

Richard Alexander Roehrl, Un Department of Economic and Social Affairs, roehrl@un.org

Abstract

The energy demand of Internet applications and artificial intelligence – while relatively small in the past – has already become significant and continues to increase unabatedly. These technologies are key to "smart" energy systems and overall energy efficiency. However, they also continue to lead to entirely new services, most of which are not geared toward increasing efficiencies of the socio-technical system – hence further increasing global energy demand. Currently, the energy efficiency of current silicon-based computing is at least 10,000 to 100,000 times lower than human brains. Against this background, it is a matter of concern that the energy efficiency of the computers, the Internet and deep neural network-based AI applications (which will replace and complement human cognitive work) has reached fundamental limits, while overall computing performance and usage increases unabated. The most likely overall result will be accelerated, increased energy demand for the Internet and AI in the coming decades, unless sufficiency considerations fundamentally change the current direction.

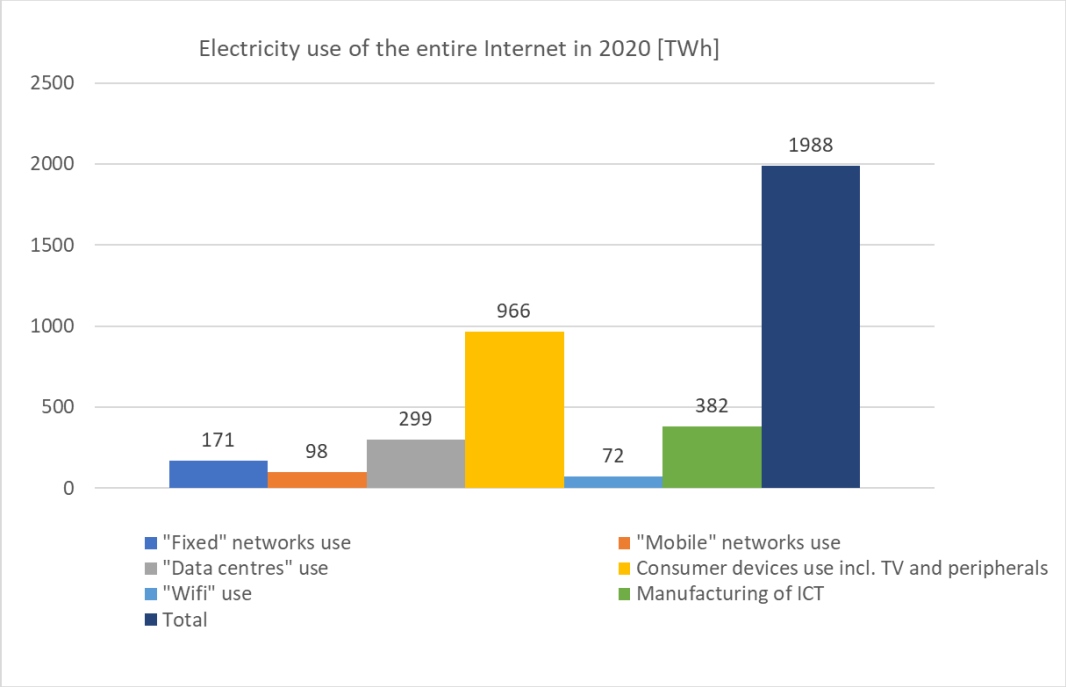
The fast pace of technological change in recent years in robotics, artificial intelligence (AI), biotechnology, nanotechnology and related areas such as "big data" are having broad impacts on economy, society and environment. At the heart of these trends are telecommunications and information and communication technologies (ICT). On the one hand, these emerging technologies hold great promise for a range of high-efficiency energy and water systems that could be deployed in all countries and catalyze the global move towards sustainability. On the other hand, despite efficiency increases, these technologies and especially AI will require ever-increasing electricity and mineral resources with its associated pollution and wastes (e.g., e-waste, nano-waste, and chemical wastes). This might be especially the case, since these new technologies make entirely new services possible, many of which are not geared toward increasing the efficiency of the existing socio-technical system. When fundamental limits to increased energy efficiency of silicon-based computing are also considered, it is evident that additional non-efficiency enhancing applications will continue increasing energy demand, unless strict sufficiency considerations or energy use limits are introduced.

Significant and increasing overall energy demand of Internet and AI

Digital technologies hold great potential for creating environmental benefits in many sectors, but they themselves are also rapidly emerging as important drivers of overall energy demand, energy mix and environmental pollution[i], especially when a life-cycle perspective is taken to ICT. The following key categories of electricity use of ICT need to be taken into account: (a) consumer devices, including personal computers, mobile phones, TVs and home entertainment systems; (b) network infrastructure; (c) data center computation and storage; and (d) production of the above categories.[ii]

The Internet has become truly pervasive. As of Jan. 2021, of the 7.8 billion people in the world, there were 4.8 billion Internet users, 2.7 billion Facebook users, 1.8 billion Websites, and 370 million active Twitter users.[iii] In a typical day in Jan. 2021, there were 265 bill. emails, 794 mill. tweets, 7.5 bill. youtube videos watched, 155 mill. Tumblr posts, 453 mill. skype video calls, 89 mill. videos uploaded. Internet traffic reached an incredible 9.4 bill. GB/day. On that day, 4.3 mill. smart phones and 0.7 mill. Computers were sold.

How much energy did this take? The best guess, peer-reviewed estimate of energy used by the entire global Internet was estimated to be about 1,988 TWh or 7.2 EJ for the year 2020 (see figure), which was equivalent to about 9 per cent of total global electricity use. Roughly half of the total, or 966TWh, was due to consumer devices, such as computers, mobile phones, laptops and TVs. The remainder (1,022 TWh) was due to local, fixed and mobile networks, data centers, and the manufacturing of the various components. Excluding consumer devices, the remainder alone caused emissions of about 949 $MtCO_2$ in 2019. The mobile networks component, in particular, is expected to rapidly increase with the advent of 5G and mobile video streaming services. It is important to note the estimates for 2020 presented in the figure were made at the end of 2019 and that actual numbers may actually have been around 40% higher due to digital response strategies to the pandemic!



Source: Roehrl (2019) [iv]*, based on data reported* by Andrae[v,vi,vii] . *Note*: This estimate for 2020 was made at the end of 2019.

The production of these devices and other ICT components are highly energy- and resource-intensive. For example, the energy production footprint of all smartphones in the world was about 30 per cent larger than that of all passenger cars and is expected to continue to an increase in line with more rapidly increasing numbers of smartphones than cars. It is estimated that it takes more energy to create a computer than it takes to run the computer for its entire working lifetime. [viii] The short product life cycle of electronic products is the prime culprit.

Online video streaming has shown to have a noticeable impact on carbon emissions, as a significant level of electricity is needed to sustain the data flows associated with the streaming process. It is estimated that globally video streaming accounts for annual carbon emissions equivalent to that of Spain. [ix] Much of this is due to mobile data use, a component that is expected to increase rapidly with the deployment of 5G networks.

Looking towards the next ten years, artificial intelligence and especially deep learning neural networks are expected to become an important driving factor for increasing energy demand and GHG emissions. Already in 2019, the overall energy use and CO_2 emissions of training every single state-of-the-art deep learning neural network (DNN) consumed an estimated 656 MWh. It was also responsible for 313 t of CO_2 emissions, about as much as five-passenger cars emit over their entire lifetimes.[x] These algorithms rely on vast amounts of data that are stored in data centres.

Bottom-up estimates (based on the number of computations) for data centers' energy use in 2030 range from a five-fold increase (from 200 to 1,000 TWh) up to a fourteen-fold increase to roughly 4,900 TWh. Much if not most of this increase will be due to AI applications.

A recent expert survey showed that a majority of experts and scenario analysts expect an increase of global energy demand over and above the dynamics-as-usual trends until 2030. [xi] A minority of experts (20%) expect a decrease and almost one third (30%) of respondents highlight the uncertainty factors.

"Smart" systems increasing efficiency

ICT in general and AI in particular will have applications and impacts in almost all aspects of the global energy system, supply (mining and production), power plants and utilities, final distribution, and end-user devices. It is also increasingly used for modelling of the energy system and even climate modelling, promising significant improvements in accuracy and flexibility of approaches and data acquisition.

Wilson et al. (2018) [xii] report on the perceived disruptiveness and emissions reduction potentials for 99 low carbon innovations, in mobility, food, buildings & cities, as well as energy supply and distribution. He also provides estimates of the resulting CO_2 emissions

reductions that could be realized through these innovations in the UK in the coming decade. For example, smart heating controls in buildings could reduce emissions by 1.2 to 2.3%, and smart appliances (fridges) could reduce emissions by 0.1%. Car clubs could reduce emissions by 0.8 to 0.9% and e-bikes and e-bike sharing by less than 0.1%.

AI applications are ideally suited for resource discovery and increasing the efficiency of *mining operations*, in view of large amounts of structured data (good example: fracking).[xiii] "*Machine learning systems can improve the ability to map and understand the size and value of underground deposits of oil and gas—in turn, making it easier to tap those resources at lower cost*."[xiii]

Deep learning neural networks can support optimizing the design and operation of wind and solar farms, making them much more efficient. Solar and wind forecasting has the potential to increase efficiencies of these systems.[xiii]

There are several high impact areas to use AI in utilities and power plants, in order to optimize generation, improve resilience and reduce operating costs. [xiv] "*Machine learning, sensors and hybrid energy storage… maximize generation efficiency*" [xiv] by permitting near real-time adjustments. Machine-learning enabled forecasting with hybrid energy storage maximize the use of intermittent renewables by anticipating and reacting to supply and demand peaks.[xiv] "*Smart wires combined with machine learning to enable real-time power dispatching, and optimize it to current grid load and to buildings' asset portfolios. Drones and insect-sized robots identify defects, predict failures, and inspect assets [including power lines] without interrupting production.*" [xiv]

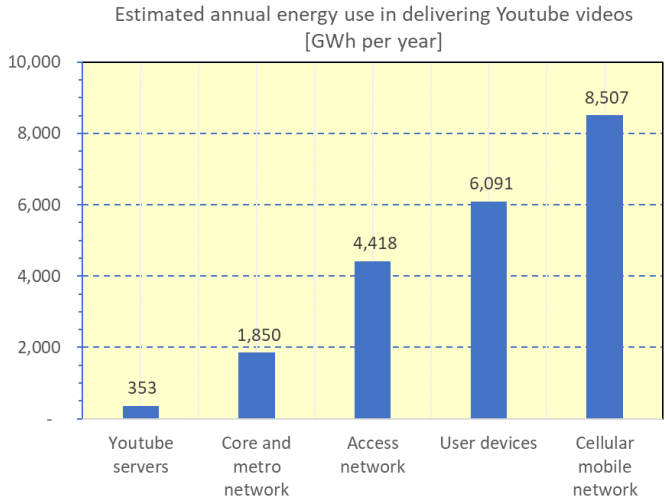
AI has been heralded by energy technology companies as key element in smart power systems and networks, by supporting the seamless integration of intermittent modern renewables, such as wind and solar, and reducing energy storage needs. However, for this study no quantitative estimates in this regard have been found from a reliable source.

The potential impact of AI is most likely biggest on the demand/consumer side, simply because of the existing large inefficiencies especially from final to useful energy. Hence, there is a need for new energy technology modelling approaches that take a consumer and demand-side perspective and that include explicitly energy flows of prime importance for understanding new ICT and AI development constraints.

Entirely new services that are not enhancing energy efficiency

Human ingenuity and perceived needs will drive the supply and demand for new services, many of which will have not beneficial effect on the global energy efficiency. Let us look at just one example: watching YouTube videos on a mobile phone. Preist et al. (2019) provide a detailed analysis of the energy and CO_2 emissions from watching YouTube videos. [xv] Youtube servers themselves only consumed 353 GWh per year,[xvi] whereas the use of core and metro networks (1,850 GWh per year), access networks (4,418 GWh per year), user devices (6,091 GWh per year), and especially cellular networks (8,507 GWh per year) are using many times more energy to watch those Youtube videos (see figure). Hence, streaming the video on the mobile phone is vastly more energy consuming and emissions creating than watching it on LAN-connected computer, especially close to backbone network. In fact, moving to the next generation 5G mobile networks[xvii] are expected to greatly increase the energy and climate footprint of online video streaming, due to much higher bandwidth available.

An even higher impact could be new video gaming streaming. According to the New Scientist Magazine, Google launched its Stadia streaming service which allows video gaming with a wifi controller, instead of a computer or a game console. Due to the streaming and especially if this will be streamed on mobile phone networks, this new service like many others that are being planned are poised to greatly increase energy use and GHG emissions.



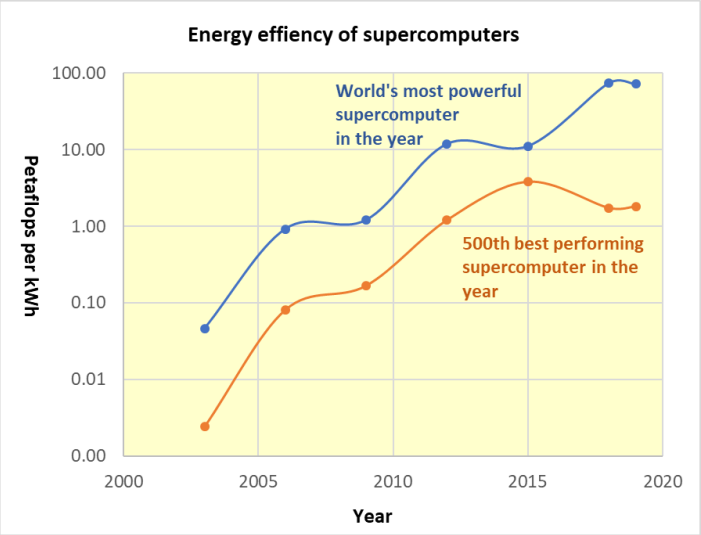
Source: Roehrl (2019) [xviii], based on data reported in Preist et al. (2019)

Fundamental limits to energy efficiency of computing

According to Moore's law, the number of transistors in a dense integrated circuit (and hence computing power) has doubled about every two years ever since 1970. According to Moore's second law (or Rock's law) the capital cost of a semiconductor fab also increases exponentially over time. Ray Kurzweil quantitatively illustrated the continued exponential growth of computing ever since 1900, well before the advent of digital technology.[xix] The question is whether or for how long this exponential growth will continue. In fact, almost technologies appear to progress at exponential rates, at least at certain points in time – this is nothing extraordinary. However, many of the technologies ultimately reach saturation and thus follow an S-shaped progress over time, similar to growth of living species in general. It is unclear whether there are exponential technologies that are fundamentally different, essentially leading to a singularity, most likely in the form of a major structural shift, or whether its exponential growth phase in these cases is simply extended over many more orders of magnitude.

Since around 2012, a slowdown of Moore's Law and Dennard Scaling has been observed. Transistors are not getting more efficient, and as a result, general purpose microprocessors are not getting faster nor more efficient (Sze, V., 2019). [xx] On the other hand, supercomputers have continued to increase their performance along an exponential law. We are only now entering artificial intelligence territory, whereas before computing power was vastly inadequate. By 2014, the top supercomputer for the first time ever reached 20 Petaflops is roughly the hardware-equivalent of the human brain, according to Kurzweil (1999) [xxi]. By the end of 2019 (today), the top supercomputer has a peak performance of about 201 Petaflops which is equivalent to 10 brains, the top500 supercomputers combined roughly as much as 50 brains. By 2025, the top supercomputer will be equivalent to 500 brains and by 2030 maybe around 10,000 brains, by 2040 similar to 700,000 brains, etc.

The top performing supercomputer shows exponential performance improvement in terms of energy efficiency over the past two decades (see figure). It reached 72 Petaflops per kWh in November 2019. It also shows that the lower performing supercomputers tended to be less efficient by about an order of magnitude. The total annual electricity consumption of the top supercomputers in each year increased from 12.6 GWh in 2006 to 88.4 GWh in 2019. Hence, even though the energy efficiency improved by a factor of 10 ever 5 years or so, the absolute electricity consumption continued to rapidly increase. Hence, the rapid emergence of supercomputers as significant contribution to world energy consumption.



Source: Roehrl (2019) [xxii], based on data in TOP500 list. [xxiii]

Deep learning neural networks (DNN) are the one artificial intelligence (AI) technology that has led to the current hype about AI in the private sector and even the general public. This highly data- and computing-intensive technology has only recently become practical, due to much larger amounts of data, orders of magnitude increased supercomputing power and new hardware.

A state-of-the-art DNN model for facial recognition, as described for example by Strubell[xxiv], requires 656 MWh for the training phase, leading to 313 tonnes of CO_2 emissions. It has a 1.52 KW power requirement for 274,120 hours parallel computing costing an incredible US$1 to 3 million in cloud computing expenses. The electricity cost for training accounts around 10% of the overall expenses. In other words, deep learning is a highly energy-intensive business. In fact, as running more and bigger models improves accuracy but greatly increases energy consumption, energy has increasingly become an important limiting factor. The reason for the above estimate is that it aims at human-level accuracy and DNN's are currently vastly less efficient than human brains which consume a mere 25W and would solve the above problem in seconds.

Hence, the human brain is many orders of magnitude more energy efficient than state of the art DNN. Replacing human cognition with current DNN technology would thus quickly run into serious global energy constraints. This would even be the case when at the expense of precision, a less energy-intensive DNN were used. In any case, the energy efficiency of current silicon-

based computing is at least an estimated four to five orders of magnitude (i.e., a factor of 10,000 to 100,000) lower than human brains. Organic computers and DNN chips might emerge by the mid-2020s that are slower but with much higher energy efficiencies, which is highly speculative, though.

It is important to note that in DNN there is a training phase that is followed by an inference phase in the actual use of the DNN. Hence, this becomes a problem in sofar as there will be many different types of tailored AI applications, unless the energy efficiency of DNN is drastically improved. It is even more serious, as further improvements in energy efficiency of DNN will require co-design of specialized hardware and software for different applications, due to the break-down of Moore's law. Previously, further minituarization led not only to higher computing performance but also higher efficiencies, but this development has reached an end, as transistors are on the order of merely a few atoms today.

To take one example, today's self-driving car prototypes use an enormous amount of power for the AI-based navigation system. For example, in 2018, self-driving car prototypes typically used an estimated 2,500 Watts of computing power. Cameras and radar alone generate about 6 GBytes of data every 30 seconds. Some prototypes even need water cooling, since the generated heat is so large.[xxv] This is not practical.

Most recently, the Navion test chip can achieve localization (i.e., the AI can localize objects in 3D) at under 25mW. This is 65 nm CMOS test chip (4 by 5 mm) with 250 configurable parameters to adapt to different sensors and environments. [xxvi] It is the first chip that performs complete visual inertial odometry. It consumes 684 times and 1,582 times less energy than mobile and desktop CPUs, respectively. Navion enables a class of low energy robotics that uses less than one 1 Watt to interact with the real world. Examples include applications for air quality monitoring; miniature satellites for deep space exploration, and origami robots for medical applications. In all these cases the actuation and computation power is low. These are ingenious innovations, but further improvements will be limited by fundamental Moore's law breakdown.

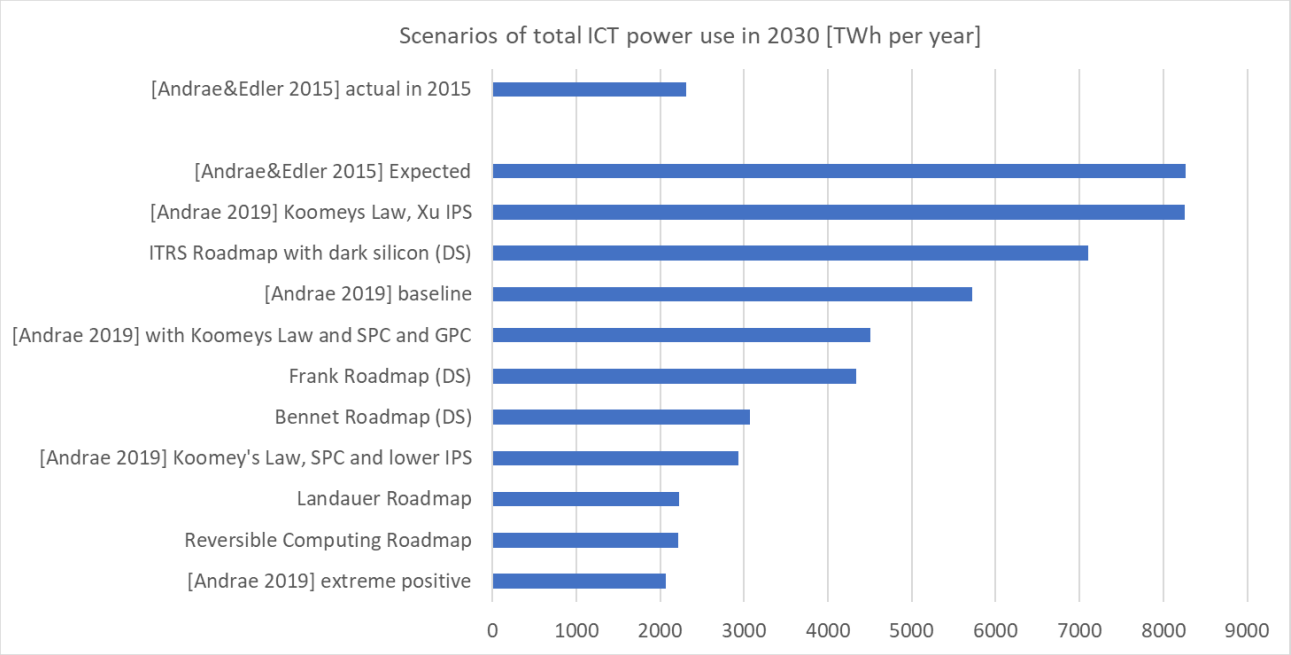
Future energy demand and the need for sufficiency considerations

Grubler et al. (2019)[xxvii] describe a pioneering, global low energy demand (LED) scenario which meets the 1.5 °C climate target as well as many sustainable development goals, without relying on negative emission technologies, such as bioenergy with carbon capture and storage. By 2050, their fully quantitative scenario reaches a global final energy demand 245 EJ (i.e., 40% lower than today), despite rises in population, income and activity. This much lower global energy system dramatically improves the feasibility of a low-carbon supply-side transformation. The LED scenario explores new social, behavioral and technological innovations.

Many if not most of these efficiency increases presumes AI or similar technologies that might have the potential to accelerate technological progress. Yet again, the AI energy consumption itself is not being considered in the scenario. However, the achieved energy demand reductions in all sectors are so large they would likely pale in comparison to even most high AI energy demand scenarios. For example, in the scenario shared and 'on-demand' fleets of more energy efficient electric vehicles with increased occupancy could reduce global energy demand for transport by 60% by 2050. Intelligent smartphones could nudge preferences towards services and against ownership. Energy performance standards of buildings could reduce energy demand from heating and cooling by 75% by 2050. Low meat diets could reduce agricultural emissions w*hile increasing forest cover*. In the scenario there is a strong emphasis on electrification and current rates of renewable energy deployment would suffice to meet future energy needs.[xxvii]

Anders Andrae and colleagues have carried out pioneering work on the future electricity demand of various ICT system components until 2030, starting with a much-cited 2015 paper on pathways between best and expected for the whole ICT and entertainment[xxviii] and this year's bottom-up estimates on semiconductors[xxix], and on high-level estimates with simple metrics.[xxx] Some of these estimates were also discussed in the response to the survey question.

The figure below compares a number of scenarios of ICT energy use by 2030 with that of 2015. Most of these estimates are based on a framework of transistor physics and various roadmaps. It shows an enormous range of possible outcomes, but in any case the electricity use will reach anywhere between 2,067 and 8,265 TWh in the year 2030. It would make sense to revisit the LED scenario[xxvii] for 2050 above and explicitly build into the model AI and other new technology components.



Source: Roehrl (2019) [xxxi], based on data by Andrae (2019).[xxxii,xxix]

Conclusion

Energy demand is closely interrelated to anthropogenic climate change and other global environmental issues. Indeed, it can be used as an imperfect proxy for many environmental issues, even though solutions, of course, go beyond that sector.

The energy demand of AI and related ICTs and their associated GHG emissions – while relatively small in the past – has already become significant and continues to increase unabated. The pandemic further accelerated a digitalization trend that was already accelerating before.

These technologies are key to "smart" energy systems and decisive for further increasing overall energy system efficiencies.

At the same time, they will continue to enable many entirely new services, most of which are not geared toward increasing the efficiency of the existing socio-technical systems – hence further increasing global energy demand. Sheer number of users and devices and the AI revolution call for making energy efficiency GHG emissions a design element in digital techs.

The energy efficiencies of computers, Internet and deep neural network-based AI applications have reached fundamental limits, but overall computing performance and usage increases unabated.

At present, the energy efficiency of silicon-based computing is at least an estimated four to five orders of magnitude lower than human brains, which highlights the potential energy impact of AI replacing or complementing human cognitive tasks in the world economy.

The most likely result of these trends will be accelerated, increased energy demand for the Internet and AI in the coming decades, unless sufficiency considerations fundamentally change the current direction.

Environmental policy makers should consider explicitly including AI technology trends and scenarios in their decision-making.

Response strategies and design standards for AI and related digital technologies and platforms are needed which will also be important business opportunities.

---

[i] For a recent literature review and expert survey, please see Roehrl, R. (2019). Exploring the impacts of ICT, new Internet applications and artificial intelligence on the global energy system. SLP/TFM research paper, December 2019.

[ii] Andrae, A. and Edler, T.E., (2015). On Global Electricity Usage of Communication Technology: Trends to 2030. Challenges 2015, 6, 117-157; doi:10.3390/challe6010117

[iii] Data source: Internetlivetstats.com

[iv] Roehrl, R. (2019). Exploring the impacts of ICT, new Internet applications and artificial intelligence on the global energy system. SLP/TFM research paper, December 2019.

[v] Andrae, A. (2019). Comparison of Several Simplistic High-Level Approaches for Estimating the Global Energy and Electricity Use of ICT Networks and Data Centers. International Journal of Green Technology, 2019, 5, 50-63.

[vi] Andrae ASG, Edler T. (2015). On global electricity usage of communication technology: trends to 2030. Challenges 2015; 6: 117-57. https://doi.org/10.3390/challe6010117

[vii] Andrae ASG (2019b). Projecting the chiaroscuro of the electricity use of communication and computing from 2018 to 2030. [cited 2019 Sept 25]: https://www.researchgate.net/publication/331047520_Projecting_the_chiaroscuro_of_the_electricity_use_of_communication_and_computing_from_2018_to_2030

[viii] Estimates from a UNU study. Link: https://www.eurekalert.org/pub_releases/2004-03/tca-uss030204.php

[ix] Efoui-Hess, M. 2019. Climate Crisis: The unsustainable use of online video - the practical case for digital sobriety. The Shift Project.

[x] Strubell, E., Ganesh, A., McCallum, A., (2019). Energy and Policy Considerations for Deep Learning in NLP. arXiv:1906.02243v1 [cs.CL] 5 Jun 2019.

[xi] Please see Roehrl, R. (2019). Exploring the impacts of ICT, new Internet applications and artificial intelligence on the global energy system. SLP/TFM research paper, December 2019,

[xii] Wilson, C., Pettifor, H., Cassar, E., Kerr, L., Wilson., M., (2018). The potential contribution of disruptive low-carbon innovations to 1.5C climate mitigation, Energy Efficiency, https://doi.org/10.1007/s12053-018-9679-8.

[xiii] Victor, D.G., (2019). How artificial intelligence will affect the future of energy and climate. Brookings institution report. 10 January 2019. Series: A Blueprint for the Future of AI: 2018-2019. https://www.brookings.edu/research/how-artificial-intelligence-will-affect-the-future-of-energy-and-climate/

[xiv] Bilodeau, S., (2019). The 4x4 Possibilities: 4 Smart Ways to Use AI in 4 Areas of Utilities' Operation. Chairman and Chief Technology Officer at Novacab Inc., 22 July 2019. https://www.energycentral.com/c/iu/4x4-case-4-smart-ways-use-artificial-intelligence-4-areas-utility

[xv] Chris Preist, Daniel Schien, and Paul Shabajee. 2019. Evaluating Sustainable Interaction Design of Digital Services: The Case of YouTube. In Proceedings of CHI Conference on Human Factors in Computing Systems Proceedings, Glasgow, Scotland UK, May 4–9, 2019 (CHI 2019), 12 pages. https://doi.org/10.1145/3290605.3300627

[xvi] According to some estimates, 10 MtCO2 emissions are due to Youtube servers alone.

[xvii] which are currently rolled out only in China and the Republic of Korea

[xviii] Roehrl, R. (2019). Exploring the impacts of ICT, new Internet applications and artificial intelligence on the global energy system. SLP/TFM research paper, December 2019.

[xix] Ray Kurzweil (2015). The singularity is near: when humans transcend biology.

[xx] Sze, V., (2019). Efficient Computing for AI and Robotics. MIT lecture. May 2019.

[xxi] According to Flow genome Project founder Steven Kotler. See: Diamandis, P.H. and Kotler, S., (2015). Bold: How to Go Big, Create Wealth and Impact the World, Simon & Schuster, ISBN-10: 1476709564.

[xxii] Roehrl, R. (2019). Exploring the impacts of ICT, new Internet applications and artificial intelligence on the global energy system. SLP/TFM research paper, December 2019

[xxiii] https://www.top500.org/statistics/perfdevel/

[xxiv] Strubell, E., Ganesh, A., McCallum, A., (2019). Energy and Policy Considerations for Deep Learning in NLP. arXiv:1906.02243v1 [cs.CL] 5 Jun 2019.

[xxv] Wired Magazine, Feb. 2018.

[xxvi] http://navion.mit.edu

[xxvii] Grubler A , Wilson C , Bento N, Boza-Kiss B, Krey V , McCollum D, Rao N , Riahi K , et al. (2018). A low energy demand scenario for meeting the 1.5 °C target and sustainable development goals without negative emission technologies. Nature Energy 3 (6): 517-525. DOI:10.1038/s41560-018-0172-6.

[xxviii] Andrae, A.S.G., Edler, T. (2015). On Global Electricity Usage of Communication Technology: Trends to 2030, Challenges 2015, 6(1), 117-157; https://doi.org/10.3390/challe6010117.

[xxix] Andrae, A.S.G. (2019). Prediction studies of the electricity use of global computing in 2030. Int J Sci Eng Investigations, 8 (86) 27-33. http://www.ijsei.com/papers/ijsei-88619-04.pdf

[xxx] Andrae, A.S.G. (2019). Comparison of Several Simplistic High-Level Approaches for Estimating the Global Energy and Electricity Use of ICT Networks and Data Centers. Int J Green Technol 5 (1) 50-63. https://ijgtech.com/ijgtv5a6/

[xxxi] Roehrl, R. (2019). Exploring the impacts of ICT, new Internet applications and artificial intelligence on the global energy system. SLP/TFM research paper, December 2019

[xxxii] Andrae, A., (2019). Drawing the fresco of electricity use of information technology in 2030 – Part II. Preprint Feb. 2019. Doi:10.13140/RG.2.2.31813.91361.