At the Boundary of Artificial Intelligence and Expert Work - Why Machines Cannot (and Should Not) Do What We Do

Authors and affiliations:
Charlotta Kronblad – Chalmers University of Technology
Johanna Pregmark - Chalmers University of Technology
Mikael Hansson – Dreamler, Mikael Hanson AB

Abstract:
In this paper we explore the boundaries of artificial intelligence (AI) and discuss what type of expert work that can and should be replaced by machines, and what type of work that should not be replaced. There are immense opportunities to use AI in court settings, not only in order to increase efficiency but also to increase quality and reduce human bias in judging. However, to switch over to machine judging would potentially entail large risks and come with some specific challenges. Moreover, we argue that due to specifics of the human brain, certain expert work (demanding ethical and/ or emotional considerations) should ultimately reside within human capacity. For example, while AI can be powerful in assisting human judges in making unbiased decisions in for instance cases where the judgement is based on how much (weight) of a drug posession, or how fast (speed) a car was going, we do not believe that judges should be replaced by machines in more complex matters involving for instance assessment of intent. We hold that there is a human element in judging, that involve parts and processes of the brain that cannot be replaced by AI that merely mimics processes of the frontal lobe. Combining the theoretical fields of organization, digitalization and law with psychological theories of the mind and the geography of the brain, we reach a novel understanding of the current boundary of AI and a humanistic approach to its limits.

Introduction

Digitalization is currently affecting all areas of society, it is considered an industrial revolution comparable to the introduction of the of the steam engine, the lightbulb and the computer (Schwab, 2017). The implementation of digital technologies and new work process has a particularly potential in the public arena where of public spending, and increased access to services, in a society where the need and demand for public services are increasing. Here technological development and innovative services impose promising , including the court system it represent a promise of efficient use and increased access to justice (Susskind & Susskind, 2015). In effect; digitalization of the court system is an area on the rise, that is subject to increased attention (Susskind, 2019). Across the globe, courts, and judiciary systems, are implementing digital technology as never before. We have recently experienced the establishment of virtual courts and a discussion has emerged on the topic of artificial judges (Susskind, 2019). This development has gained speed and become increasingly relevant due to the pandemic times of Covid-19 (Kronblad and Pregmark, forthcoming).

Several researchers have pointed out the benefits in the implementation of new technology and that applying more digital technology in courts would not only tackle efficiency issues (saving time and cutting costs) but also help solving the general problem of human bias effecting court verdicts (Kleinberg et al., 2017). Where research for instance has shown that the likelihood for a positive decision increases substantially if the judge has just been on a lunch break (Danziger et al., 2011). AI would help solve this problem as it would not deliver judgements influenced by human bias, or be affected by whether or not the judge had lunch. Instead AI and algorithms would create unbiased and objective verdicts and decisions. However, although most of us probably agree that a machine judge would probably be more objective than a human, most of us would still feel unease being judged by a machine. Jargo (2019) argues that there is an element of authenticity to human work, and that people in general believe that algorithmic work is less authentic than human work and that moral decisions made by algorithms are relatively less ethical than identical human decisions. This implies a challenge for digital judges and algorithmic verdicts.

In this paper we depart from this inate challenge for artificial replicants of human experts and take it one step further; into a normative discussion for the context of the court system. Thus we ask how judges today make use of digital technologies, and where they see the limit for the

use of AI and furthermore, if court verdicts delivered by machines is something that we should strive for, or if there could be elements inherent to human cognition that simply cannot be replicated by a machine?

To answer these questions we seek to combine some theoretical areas that are not often understood together. Thus we are combining theories from organization and digitalization of work with legal theories and psychological studies of the brain. The research is built on the cooperative work of two management scholars (one with a legal background and extensive experience of working in the court system) and one scholar from the field of psychology, having 30 years of experience from working as a psychologist. This mix of theoretical, and professional, backgrounds we believe resonate well with the current need to combine different fields and experiences to solve the increasingly complex problems of the future. To obtain a mix of competences in research groups is particularly relevant in the fast changing process of the curernt digital transformation.

Theoretical framework

As a point of departure we use the concept of the second machine age (coined by Brynjolfsson & McAfee, 2014) - where creative, and expert, work is increasingly being replaced by machines (following the first machine age replacing workers in manufacturing and agriculture). Thus there is an ongoing dicussion in expert communities if digitalization will impact their work, and in particular if their work will be replaced by AI. For the completion of service work, Huang and Rust (2018) state that four different types of intelligences are needed. They present a model of mechanical, analytical, intuitive and empathetic intelligences and argues that machines will (or have already) replaced humans for tasks demanding mechanical and analytical intelligence whereas tasks demanding intuitive or empathetic skills are harder to replace, but will be replaced further on. Thus there will be a time of continuous transition where machines and humans work together while the jobs that the machines can do, and do, are increasing.

From a psychology/neuro science (see e.g. Rock, 2008; Lieberman, 2013; Kahneman, 2011) point of view these different intelligences depart from different parts of the brain. Where the logical reasoning – needed for mechanical or analytical cognitive processes are residing in the frontal lobe, intuitive and empathetic cognition depart from other locations, and or connections between different locations, in the brain. These different intelligences are a combination of the

dynamic in several neurological areas. In short, the frontal lobes assist us in understanding the rationality of things, making sense analyzing the components logically and their interconnected functionality. It is also responsible for regulating our emotions, trying to come up with the most functional and appropriate response in a given moment. Our limbic system is activated every time we receive any stimuli, resulting in an emotional reaction. And although these emotions vary in strengths, duration and frequency they are constantly influencing our perception and subsequent decision-making (Lieberman, 2013).

Although the application of legal rules and regulations is fairly black or white – and rather suitable for digital processing, and would also overcome some current problems of human bias apparent in verdicts, the work of judges currently consists of all four intelligences. This means that human judges activate several diffrerent parts of the brain, as well as the connections between them. Replacing this cognitive capacity with AI, that merely replicates the analytical processing of the frontal lobe, would mean that the processing at several locations, and their connected practices, would not be replaced, leaving the emotional and moral reasoning behind. This is due to the profoundly human aspect of decision-making that we refer to as values, and that relate to what we deem important, regardless of whether its rationality smart or triggers the relevant emotions. Turning human thinking into algorithmic we need to focus on this aspect, as it is currently the last instance that is activated before a decision is made, hence making it a function in the human brain that could be labeled a "gatekeeper" or potential "tipping point". An example would be that judges often need to apply their personal interpretation and intuition to cases as legal texts and previous verdicts are often not clear enough. Rangel (2009) has repeatedly shown that the activity in the ventromedial prefrontal cortex reveals the subjective value of different types of reward and this value determines our decisions. To translate this thinking into machines and incorporate moral values into algorithmic decision making is a huge challenge.

Method

This as a study that is driven by a phenomenon (Schwarz & Stensaker, 2016; von Krogh et al., 2012). AI is, along with other digital technologies, being introduced into societies and the courts are struggling to figure out how to use it. Following and analyzing these endevours we seek to understand how this potential is being realized and we set this very practical issue in relation to theory and what is previously known. To conduct the study we consequently desided upon a qualitative research design (Denzin & Lincoln, 2005; Gioia et al., 2012) We decided on this method because of its appropriatness, due to the aim (setting out to study a phenomena in depth and understand responses to this phenomena). The qualitative design is particularly suitable since digitalization is inherently a complex and still ongoing phenomenon (Eisenhardt & Graebner, 2007). The qualitative design enables us to understand how AI can potentially be used within a court setting and how judges can use AI in combination with their legal insights in order to enhance their practice and delivery of justice. With this said, we have a hope that this paper could potentiallt benefit practice as well as research and advance access to justice and the future of the judiciary.

We have taken part in three workshops involving more than 100 judges in Sweden. These are judges from the regular courts as well as appellate courts and even the surpreme courts. The data was collected at the workshops held in 2019, in terms of detailed notes. Although the exersices and questions were not identical in these sessions they were similar and can be summarized in four areas of subsequent exercises that the participants were asked to participate in during the sessions.

- The first exercise asked the judges to consider the most important ways that digitalization had already affected their work in the shape of new work tools and processes.
- The second exersice was about how they perceived the possibility to use AI in their capacity as judges
- The third exersice dealt with the limit of AI in their work (their work practices in judging) and explicitly targeted “what should we not use AI for”.
- And in the fourth exersice the judges were asked to gaze into the future. Where are we in 30 years? Our thought was that to give them a perspective of 5 or 10 years would not

provide enough of a time perspective or horizon to be really creative. But to ask for 30 we would more likely end up in a discussion of what will be here in five.

We did consider filming or otherwise recording the sessions, but we believed that this could hamper the exercises and the quality of the data, more than it would benefit its collection, and thus a decision was made to keep the notetaking analogue. However pictures were taken after the sessions (of the whiteboards summarizing the expressions of the audience) as to keep record of the main insights from each of the workshops. The collected data was analysed by the authors together, and in regard to the theoretical frame. While we did not deal with any transcribed data, but rather detailed notes and our own insights, we grouped the findings together under general themes and tried to analyze what these findings implied in relation to what is previously known.

Moreover, in regard to the research setting it should be noted that we believe that the Swedish setting is particularly suitable for studies in digitalization, as Sweden is comparably mature in sence of digital adoption among its citizens. Due to this reasoning we hold that findings from this research-settings would be transferrable also to other settings.

Findings

Exersice 1: The most important ways that digitalization had already affected their work as judges

From the first exersice we could derive that digitalization had affected the day to day work of most jugdes as new work tools and work processes have altered how that they work: for instance they mentioned digial filing (e-files) and the access to legal sources on line. This, together with better hardware and technolgy having been implemented at the homes of the judges, enabled them to work more flexibel, both in regard to time and to geography. *"You can really work from where, and when, you want, so you do not have to get in to the court"*. However several judges mentioned that the physical meeting, being able to see and experience eachother (both in regard to citizens and to their colleagues), had become particulary important in these times of digital transformation. They claimed that it was particularly important to create a feeling for the client/citizen of "being seen". This was connected to the core value in the delivery of justice, and that justice is a broad term that is not only connected to substantial justice (getting a verdict that is correct) but also connected to procedural justice (getting a fair trial), and the importance in assuring the citizens of fair and correct procedural justice and establishing trust in this regard. This was also lifted in regard to larger societal trust and the importance of the society trusting the institution of the court. In summary, there was three words (themes) that emerged as vital in this first exercise: digital files, legal on line sources and the physical meeting.

Exersice 2: How can AI be used in judging?

It was generally agreed upon that AI can be a powerful tool in assisting human judges in making unbiased decisions in for instance cases where the judgement is based on how much (weight) of a drug posession, or how fast (speed) a car was going. Additional uses that were mentioned were calculations of what sums should be awarded for damages, determination of different sanctions, cost calculations and in some cases also help in the evaluation of evidence (for instance in regard to certain elements and probabilities). Also assessment of the risks of returning into criminal behaviour was mentioned as one potential application. The judges stated that getting automated (smart) suggestions for sentancing would help them do their work, in a collaborative effort, and would also reduce risks of human mistake, for instance in making

complicated estimations and calculations of likelihoods of different events. Also some judges exxpressed that AI could potentially help them with their own biases, as they did recognice that they often regarded things should not legally relevant, in their judgement: for instance being more lenient on suspects asking for forgivness or expressing regret, or to women in general.

In summary, it seems to be agreed among the judges that AI would be applicable in cases where ”amounts” measurements and calcutalions matter, and where a common handling process would benefit decision making (for instance when basing conclusions on large amounts of data). It was also stated that an increased use of AI and automation could simplify the flow in the court process and increase access to justice (by the citizens).

Exersice 3: What should we not use AI for?

Most judges stressed that human judges should not be replaced by algorithms or machines when it comes to more complex matters, involving for instance assessment of intent (crucial in determination of criminal justice). In fact most stressed that the judge should not be replaced by AI at all, but work together with AI. ”*It is in the combination of AI and judging that really interesting things could happen, maybe we could elevate quality as well as efficiency*”. Most judges were open to the application of AI to make their decision process more efficient and also more just, both across the geographical spread of the country (the entire jurisdiction) ensuring more homogenous verdicts. However also stressed was the fact that in order to use AI, the judges need to obtain a larger trust for the technology, building on an understanding of what is valued in the AI, and a transparancy in what parameters have been used in the decisionmaking and how it came to a certain conclusion. Thus, transparancy in the system matters, and the need of tecnhological competence to even assess and understand the basis of this technology. This being crucial for the willingness of judges to work together with AI.

Exersice 4: Where are we in 30 years?

In regard to where we will be in 30 years the judges discussed different digital alternatives to the formal court process, and mentioned that such alternatives have already emerged on the private arena in terms of different providers of on line dispute resolution. A fear was expressed that if the public institutions did not transform in line with the general expectaions, the citizens would turn to these private dispute resolution providers instead and that would ultimately damage society and decrease the rule of law. At the supreme court the judges discussed the

purpose of the court, being to provide justice to the people, and that it is highly important to be open to change and ensure that justice is provided in a matter that the citizens desire. The lower courts also discussed that not all cases should reach the courts, as it is today, but that ”it would be good to create a fast-track for certain smaller cases” or to ”put up some rules of what cases the court should attend to”. ”*We should be more like a hospital, and say that you are not supposed to come in with a common cold – that you have to cure at home – it is simply not reasonable that public spending is allocated to that*.” The judges also discussed the possibility of a completely digital court in the future. ”this would suit the new generations better”. Also discussed was the current division into different geographical areas with common courts in the city that serves (has jurisdiction over) the surrounding geographical area. It was discussed that a better division in the future might be based on different legal areas. ”*As society, and the issues that come to us, are all the more complex, it is not viable for us to be experts of everything.” ”Why don’t we just have one Swedish court with different departments and judges that are specialized in different legal areas*.” However a risk was raised in this transformation that there are certain symbols that carry trust for the institution of the court ”*you know, the large wooden club to bring order in the court room, and the way that the court rooms are designed with the judges behind a high bench in the back, if we are to change all this we need to find new symbols that can create and carry our professional identities*.”

Discussion

Based on our findings we argue that there are certain tasks, and cases, that would be more suitable for the implementation of AI, whilst there are other cases where judging should remain in the capacity of judges. Our findings stress that judges are open to the application of AI in order to make their decision process more efficient and also more just (by eliminating human bias and reducing risks stemming from human behavior and error). The judges were even open to a future were smaller cases would be served in a ”fast-track” manner, where automation and the application of AI could ultimately increase the speed of the court process, lower the public cost for the procedures and even increase the quality in the verdicts being the same across the country (and over the time of the day – without any effects of lunch breaks). The judges however argued that they should not be replaced by algorithms or AI when it came to more complex matters. Examples that were brought up involved the assessment of intent, which matters on order to get a suspect commited for a crime. However, even in these more complex cases the judges stressed that they could be assisted by AI, for instane in complex calculations, in determining sizes of damages or in sentencing. Most of the judges saw that there was some expert workt that they would not leave to AI, but were simultaneously open to the assistance of AI in most tasks. ”*It is in the combination of AI and judging that really interesting things could happen, maybe we could elevate quality as well as efficiency*”.

This paper stress that there is a boundary to what AI/machines can do, and that there is an element of being human – that stemmes from our processes of thinking that do not take place in the frontal lobe (Rock, 2008; Lieberman, 2013; Kahneman, 2019) that cannot be recreated artificially (at least not yet). By combining knowledge from fields of theory, that are not commonly discussed together, this paper provides a novel understanding of the limits of AI – and gives us some baseline to discuss what work is suitible for machines and what work tasks should remain in the sole capacity of humans. While we agree with Huang and Rust (2017) that different work demands different intellegences, where mechanical and analytical intellegences are the first to be replaced but intuitive and empathetic intellegences are replaced at a later stage, we add that there is a large difference between these intellegences, in regard to their origin. Where in the brain the thinking takes place. Where mechanical and analytical thinking occurs in the frontal lobe, more intuitive and empathetic thinking resides in other parts of the brain. We consequently argue that the current status of AI is that work demanding the attention of the frontal lobe is being replaced, while work demanding connections to the emotional part

or the of the brain can not yet be replaced. Moreover, we argue that there might be another part of the brain, the ventromedial prefrontal cortex, that hold room for a sense of justice, or ethics and that is crucial for expert work, which we term the ”veto brain”, which with current technology is far from being replaced, and that potentially should not be replaced. This part of the brain is less discussed in regards to decision making but seem to be connected to deeper image one has of oneself – what is being perceived as “you”, This could connected to work by for instance Senge (1990) around mental models, which is described as deeply held images affecting how you think and act. As long as this part of the brain cannot be replicated artificially – we can and should not replace human expert work with artificiall capacity (especially not in regard to judging) but merely use the machine as a tool that can thelp us do the human specific work in the very best way. This combination of human and machines does not only hold superiority (to the alternatives consisting of one of them), but has the potential of gaining public acceptance, being acceptable as authentic and ethical in the minds of the public, which is essential for the sence of justice and trust that modern democracies builds upon.

Challenges and Questions for this Conference:

We are happy that you have devoted your time into reading our paper this far and would now like to get your assistance in several different areas. We believe that the paper would benefit from a good discussion and hope that this forum would provide such opportunity!

Primarily, we need some help in connecting our main idea (that discusses the capacity boundry of AI as compared to the human brain) to the empirical field of expert work. In essence: we need your help understanding how to best tie the conceptual idea and the suggested framework to the empirical domain of expert work, in particular the work of judges. We also need to discuss how we should frame the paper to capture the interest of the reader, and what audience we should seek. Are there any particular journals that you could see fit this paper?

Also, as this paper is still at a very early stage, do you like the empirical stance of it, even if it is just a collection of data based on workshop participation and observations? Do we need to gather more empirical data? Should the paper even be empirical or is it better to have a conceptual appraoch? Since there is already some knowledge available in respective field, the novelty of this paper might be in combining these – and analyzing them together to reach a new understanding. If the empirical approached however is deemed more suitable, we would like to explore if you think that what we already have holds, or else what methods you think should be appropriate. I.e. how should we study this intersection of human expert work and AI, and who should we target in sampling? Adjusting the research questions in accordance, would also be a key part of this discussion. As we have already established contact with several courts it would not be a problem to conduct further studies on site.

References:

Brynjolfsson, E., & McAfee, A. (2014) The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies. WW Norton & Company

Danziger, S., Levav, J., & Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. Proceedings of the National Academy of Sciences, 108(17), 6889-6892.

Denzin, N. K., & Lincoln, Y. S. (2005). The Sage Handbook of Qualitative research, 3rd ed. Thousand Oaks, CA: Sage Publication.

Eisenhardt, K. M., & Graebner, M. E. (2007). Theory Building From Cases: Opportunities And Challenges. Academy of Management Journal, 50(1), 25-32. doi:10.5465/amj.2007.24160888

Gioia, D. A., Corley, K. G. and Hamilton, A.L. (2012). Seeking qualitative rigor in inductive research: notes on the Gioia methodology, Organizational Research Methods, Volume: 16 issue: 1, page(s): 15-31

Huang, M. H., & Rust, R. T. (2018) Artificial intelligence in services. Journal of Service Research, 21(2), 155-172

Jago, A. S. (2019). Algorithms and authenticity. Academy of Management Discoveries, 5(1), 38-56.

Kahneman, D. (2011). Thinking, fast and slow. Macmillan.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). Human decisions and machine predictions. The quarterly journal of economics, 133(1), 237-293

Kronblad & Pregmark forthcoming (2021) How COVID-19 has changed the digital trajectory for professional service firms, Chapter in "Covid19 and the future of the service industry post-pandemic: insights and resources" Springer

von Krogh, G., Rossi-Lamastra, C., & Haefliger, S. (2012). Phenomenon-based Research in Management and Organisation Science: When is it Rigorous and Does it Matter? Long Range Planning 45(3) 277-298

Lieberman, M. D. (2013). Social: Why our brains are wired to connect. OUP Oxford.

Hare, T. A., Camerer, C. F., & Rangel, A. (2009). Self-control in decision-making involves modulation of the vmPFC valuation system. Science, 324(5927), 646-648.

Rock, D. (2008). SCARF: A brain-based model for collaborating with and influencing others. NeuroLeadership Journal, 1(1), 44-52.

Senge, P. M. (1990). The art and practice of the learning organization.

Schwarz, G., & Stensaker, I. (2016). Showcasing phenomenon-driven research on organizational change. Journal of Change Management, 16:4, 245-264,

Susskind R., Susskind D., (2015) The future of the professions, New York: Oxford university press

Susskind R., (2019). Online Courts and the Future of Justice. Oxford University Press.