Journal & Proceedings of the Royal Society of New South Wales, vol. 157, part 1, 2024, pp. 123–129. ISSN 0035-9173/24/010123-07

Artificial intelligence: affordances and limits in the context of judging

Lyria Bennett Moses

Director, UNSW Allens Hub for Technology, Law and Innovation; Professor and Associate Dean (Research), Faculty of Law and Justice, UNSW Sydney

lyria@unsw.edu.au

Abstract

This paper is based on a presentation given on 2 November 2023 at the Royal Society of New South Wales and Learned Academies Forum "Our Twenty-First Century Brain," as part of a panel on Turbocharging Human Intelligence with Artificial Intelligence. A question posed in the panel was what changes we face as humans given the increased complexity of our interaction with artificial intelligence (AI). I explored that question through the lens a critical role played by humans in our society, namely that of judges. In this paper, I explore the extent to which AI might, alone or with humans, perform such a role and what this might mean for our understanding of the criticality of human involvement in high stakes decision-making.

What is AI?

AI is a confused term, but it would seem we are stuck with it. Part of the problem is the term "intelligence" itself, which often recalls one-dimensional metrics such as IQ. Engineered systems have a range of capabilities that produce outputs that in some circumstances are identical to, similar to, or more useful than that might be produced by an intelligent human, but we do not always call the result artificial intelligence. An example is the humble calculator. If I were to calculate 2,180,906 / 598 on paper, I would rely on my memory of the algorithm for long division and my ability to perform the calculations required. It is fair to say that my ability to execute the task involves *intelligence*, but the device used to do it in my stead, despite being "artificial," would not generally be described as "artificial intelligence." On the other hand, the ability of generative AI tools to write text, despite making the kinds of mistakes that would be rare for humans, is considered by many as the current pinnacle of "artificial intelligence."

Definitions of AI typically focus on either a field of research comprising subfields such as machine learning, computer vision, natural language processing and so forth or an adjective to describe a kind of system. For both, some definitions focus on anthropomorphic comparisons: the classic example being the Dartmouth Summer Research Project in 1956 which referred to the field of research as "making a machine behave in ways that would be called intelligent if a human were so behaving."[1] Other definitions of the field of research focus on rationality rather than similarity to humans.[2]

[1] John McCarthy et al. (1955) A proposal for the Dartmouth Summer Research Project on artificial intelligence, report, 31 August, https://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

[2] Russell S and Norvig P (2016) *Artificial Intelligence: A Modern Approach*, (3rd ed, Pearson Education Ltd), pp. 2–5.

The OECD defines AI systems rather than the field of research, stating that:[3]

> An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.

Note that "content" was not in the original definition but was added following developments in large language models and other generative AI techniques. The definition of AI system in ISO/IEC 22989:2023 is similar:

> Engineered system that generates outputs such as content, forecasts, recommendations or decisions for a given set of human-defined objectives.

The multiplicity and evolution of definitions suggests we are still coming to terms with the kinds of things we are creating. It might be, as Roger Clarke suggests, that we are defining the wrong concept and that we might be better off with a term that better captures the fact that humans and systems co-produce outputs that influence physical and virtual environments.[4] However, for the purposes of this paper, I will adopt the definition in ISO/IEC 22989. This treats the concept of artificial intelligence as distinct from the property of "autonomy." It thus allows for socio-technical systems that include both human beings and AI subsystems, which are likely to be the basis for most useful applications in the context of judging.

Judging and Artificial Intelligence

There is no doubt that judges use a range of AI tools in the context of their work.[5] For example:

1. Legal research tools increasingly rely on AI in addition to more straightforward techniques of phrase matching and cross-linking; and
2. Word processing devices (internal or external to standard word processing software) encourages stylistic and grammatical enhancements and may increasingly also identify repetition and opportunities to improve signposting and structure.

What is most controversial, however, is the use of AI tools in constructing reasoning or reaching decisions. Here, there is an important distinction, not along the lines of "artificial intelligence" but along the line of "autonomy." Sourdin uses the terminology "Judge AI" versus "supportive Judge AI" to capture the distinction between autonomous AI systems that substitute for a human judge and systems that assist a human judge in their work.[6] However, it is less a binary than a scale of decreasing levels of human involvement into the final decision and reasons. A human judge who

[3] OECD (2023) Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449 (adopted 22 May 2019, amended 8 November, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

[4] Clarke R (2023) The re-conception of AI: Beyond Artificial, and beyond Intelligence. *IEEE Trans. on Technology and Society* 4(1): 24.

[5] See generally Bell F et al. (2022) *AI Decision-Making and the Courts: A Guide for Judges, Tribunal Members and Court Administrators* (AIJA).

[6] Sourdin T (2021) *Judges, Technology and Artificial Intelligence*, Elgar, p. 16.

has little understanding of the affordances of the technology being used and takes a trusting attitude to its outputs is very close to Judge AI.

Three dimensions of measurement

In earlier work,[7] I set out a three-dimensional framework for assessing AI in a context such as judging. My goal in doing so was to counter the narrative around an AI "singularity" which imagined a one-dimensional comparison between the "intelligence" of humans and machines. The three dimensions, shown in Figure 1, are: (1) the extent to which available tools perform well in the context of a clearly defined purpose (do we?); (2) the extent to which AI as a discipline has the capability to perform particular functions (can we?); and (3) the extent to which the use of available tools in the particular context would be appropriate (should we?). The first dimension, while seemingly mundane, is important because we often get excited about capability and concerned about appropriateness, leading to simplistic utopian/dystopian visions that ignore the fact that most of the problems experienced in practice involve an inadequately thought-through purpose and poor implementation. One reason why legal projects often measure poorly on this dimension is the relative lack of expertise and understanding among legal experts involved in commissioning AI projects.[8]

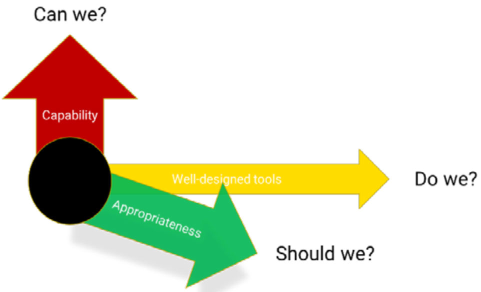

Figure 1: Applications of artificial intelligence can be measured in three dimensions

Improving decision-making around AI adoption in courts and tribunals

It was out of a desire to improve decision-making in critical contexts such as courts around the uses of AI that some colleagues and I partnered with the Australasian Institute for Judicial Administration to create a guide on AI for judges, tribunal members and court administrators.[9]

The guide does several things. It begins with a basic explanation of terminology as well as the various capabilities of different kinds of AI (mirroring the "can we" axis). It then describes the most common applications in the domain of courts and tribunals, as well as some of the limitations of these in practice (mirroring the "do we" axis). After that, it sets out the most critical judicial values, and explains the implications on these of the various use cases, given the affordances of the different kinds of tools being deployed (mirroring the "should we?" axis). The conclusions are not in the form of answers or prescriptions, because all of these

[7] Moses LB (2020) Not a single singularity. In: Deakin S and Markou C (eds) *Is Law Computable? Critical Perspectives on Law and Artificial Intelligence*. Hart Pub., pp. 205–222.

[8] For a solution, see Hildebrandt M (2023) Grounding comptuational "law" in legal education and professional legal training. In: Brozek B, Kanevsaia O and Palka P (eds) *Elgar Handbook on Law and Technology*.

[9] Bell et al. (2022) *op cit* n 5.

questions are highly context dependent. Whether or not a particular tool should be used will depend on the purpose of doing so, the approach or methodology (for example, whether code-driven or data-driven), the performance of the tool (including whether it has been evaluated and as against which metrics), and so forth. Instead, the guide is constructed around questions that we argue are critical in deciding whether to use a particular tool in a particular context.

The first set of questions are at the highest level and provide a starting point for analysis:

1. Why is AI being used? What problem does it solve?
2. Is the use of AI authorised in the context in which it is deployed?
3. In what contexts is AI being used, and is its use in those contexts appropriate? Does the context involve high stakes, vulnerable people, novel situations, or high levels of emotion?
4. How is AI being used? How can system requirements (through a procurement process) better fulfil its purposes and meet the needs of courts and tribunals, including in relation to core judicial values? How will the system be checked, tested and evaluated to ensure it meets those requirements?
5. Who is consulted about the deployment of AI systems? Are all stakeholders including users and litigants included in decision-making about whether and how AI will be used?
6. Will the use of AI impact on public confidence in the judiciary?
7. Will the use of AI in the courtrooms be accepted by the public?

Other questions relate to the various aspects of "should we," revolving around the identified judicial values, being open justice, accountability, impartiality and equality before the law, procedural fairness, access to justice, and efficiency.[10]

The technological imaginary

While the guide provides a tool that courts and tribunals can use in decision-making, it operates within the domain of current capabilities. Indeed, we will soon publish a second edition that brings the first up to date, both in relation to the increasing number of examples of AI deployment internationally, but also recognising the growing capability of and interest in generative AI tools such as OpenAI's ChatGPT.

Going beyond current systems, increasingly optimistic hypotheticals are being posed about how our society generally and courts in particular might respond to significant inflation in the second dimension, namely in the capabilities of AI. What if, for example, generative AI was linked with a reasoning engine that solved the problem of hallucinations? What if there were a "legal singularity" with AI systems able to produce judgments that were, to a critical observer, indistinguishable from those authored by humans?[11] For the purposes of the exercise, it is not necessary to decide whether any of these things are technically possible or likely. However, it does draw us back to the issue

[10] The second edition of the Guide, currently in draft, will link accountability with another important value, namely independence.

[11] Alarie B (2016) The path of the law: toward legal singularity. *University of Toronto Law Journal* 66: 443.

of what role judges (or indeed other humans undertaking critical tasks) perform.

One answer that is sometimes given relates to human traits such as empathy.[12] For example, many insist that presenting a case to a human with an ability to empathise with the parties before them, and receiving judgment from someone able to relate to the impact of their decision, is important. However, there is a need to be careful here. In some contexts, such as first instance judges deciding matters between individual litigants or against an individual defendant, the ability to connect and relate can help people feel heard and better able to cope with a negative outcome. However, many judges, particularly in higher courts, would minimise the importance of this. At least some would argue that their performance should be evaluated primarily on the basis of their judgments and, in particular, the doctrinal rigour of their reasoning. This leaves them vulnerable to the argument of Alarie that it would be reasonable to replace human judges with a system that can simulate that output where it is judged (by a third-party observer) as of equivalent quality.[13]

To try to get at the question of whether artificial intelligence *ought to* (given sufficient capability) replace judges, it is necessary to dwell on *purpose*. What judges do, even in higher courts, goes beyond producing text containing valid doctrinal arguments. What is most important is that they are exercising judgment. This is different from both prediction (working out the expected outcome of litigation using probability) and simulation (which is what ChatGPT does when asked to produce the text of a judgment). The manner of the decision is as critical as its content. An analogy might be elections — even if the accuracy of polling could be improved to the point that the chances of same-day polling yielding a different answer from the formal election were minimal, we would not want to replace elections with polling. What matters is not simply the ability to predict an outcome, but the judgment made by the electorate at a particular solemn moment of decision. Writing a judgment is slower than completing a ballot paper, but the point is similar — the exercise of judgment in reaching a decision is more critical to the function of judging than the production of text as such. Predicting that judgment (who will win the case, amount of damages, length of sentence, etc.) or generating reasons artificially cannot substitute for the exercise of judgment, even if it is impossible for a third-party observer to tell the difference. This also goes beyond issues of empathy, both as a capacity and as interpersonal sharing of affect. Empathy may be important, particularly in how litigants themselves perceive the process, but is not the only the only thing lost in a shift towards automation.

There are also other potential manifestations of artificial intelligence that have different affordances, including the capacity

[12] Empathy is an ambiguous term, see Hall JA & Schwartz R (2019) Empathy present and future. *The Journal of Social Psychology* 159(3): 225. For the purposes of this paper, I adopt the approach of Decety and Jackson, who describe three functional components of empathy, being interpersonal sharing of affect, self-other awareness with clear regulatory mechanisms to distinguish between the two, and a cognitive component that involves adopting the perspective of another: Decety J and Jackson PL (2004) The functional architecture of human empathy. *Behavioral and Cognitive Neuroscience Reviews* 3: 1.

[13] Alarie (2016) *op cit* n 11.

to exercise judgment. Imagine the possibility of duplicating human brains *in silico*, so that a judge living today could be intellectually duplicated in an engineered system. There are philosophical questions about whether the system would indeed be a "duplicate," but for current purposes assume that it could exercise the same kind of judgment as the human judge from which it was copied.

There are still concerns, albeit different ones. The first is social licence for this kind of practice, in the very practical sense of whether people would subjectively recognise the engineered judge-clone as a legitimate decision-maker. The second is the problem of moral evolution — an eighteenth-century judge could be taught new legal doctrine, but could he recognise women and non-white people as worthy of the same dignity as himself? But the third problem comes back to a twist on the idea of empathy.

My third issue relates to the purpose of the rule of law in tempering power.[14] Where a judge perceives themselves subject to the same law that they are interpreting and applying, that belief acts as a constraint on arbitrariness. At least theoretically, if a judge committed the same offence as the person before them, they would be susceptible to sentencing by someone in a similar position to themselves. Similarly, if they themselves or an entity in which they held an interest was involved in a civil dispute similar to that before them, the same rules and interpretations would apply. As a result of that awareness, the judge might be less likely to act arbitrarily than a despotic ruler not subject to the same rules as everyone else. An engineered system, even if it is effectively a human clone, would not have the same awareness because, even if it were conscious, it would not experience the law in the same way. The "experience" of an engineered system of a jail cell would not be the same qualitatively as the experience of a biological human. Similarly, the system's relation to money (such as may be payable in damages) is qualitatively different to that of a human for whom it may help provide for themselves and others. If the entity making, interpreting or enforcing rules experiences those rules fundamentally differently, then the rule of law as a means of tempering power breaks down. This is not the *same* as experiencing empathy, but it may in practice be related.

Conclusion

Many words have been spilt over the question of whether and when artificial intelligence might replace humans. Much of this links to the idea of a singularity when machines become more "intelligent" than we are, although the multi-faceted nature of that concept is usually ignored in the comparison. But humanness is more than intelligence and is certainly more than an exercise in the prediction of the outputs of an intelligent mind or the simulation of its work products. If we are to analyse whether machines might replace humans not just at a task (like playing chess) but in an important social role (like a judge), we need to go beyond comparing intelligence. Instead, we need to understand purpose and what it is, often unspoken, that links that purpose to humanness. The outputs of a human brain may be indistinguishable from the outputs of an engineered system — I certainly got the same answer for 2,180,906 / 598 as my

[14] Krygier M (2016) Tempering power. In: Adams M, Ballin EH and Meuwese A (eds) *Bridging Idealism and Realism in Constitutionalism and Rule of Law*. Cambridge: Cambridge University Press.

calculator, and we can assume that future systems may be able to write like me and speak like me in ways that would fool even those who know me well. This Turing test, however, is not enough when considering what *roles* engineered systems might be able to perform.[15]

To perform the role of a judge, I believe an entity will need at least three things: (1) the ability to exercise judgment; (2) being attuned to the morality of the community in which decisions are made (more or less, acknowledging there are a range of acceptable moral views in any community); and (3) being subject to law (the same law being applied to humans) in a meaningful sense. There are inevitably more — these are just the ones revealed by the hypotheticals considered above. But they reveal something important — we cannot look solely at intelligence in comparing humans and AI — we need to understand more about ourselves and our society to decide where we can and should stand aside in favour of our tools.

[15] Turing AM (1950) Computing machinery and intelligence. *Mind* 59: 433.