UNIVERSITY OF NOTTINGHAM

DISSERTATION FOR DEGREE OF MASTER OF LAW (LLM)

LLM in Public Procurement Law and Policy

A CRITICAL ANALYSIS OF THE LEGAL IMPLICATIONS OF THE USE OF GENERATIVE AI FOR THE EVALUATION OF TENDER SUBMISSIONS UNDER THE PROCUREMENT ACT 2023

By

Andrew de Whalley, BA (Nottingham),
MSc (King's College London)

Candidate of the 2022 Executive Programme in Public Procurement Law and Policy
School of Law, University of Nottingham

……

I hereby declare that I have read and understood the regulations governing the submission of postgraduate dissertations, including those related to length and plagiarism, as contained in the LLM Manual and that this dissertation conforms to those regulations.

Introduction

On the 30th November 2022, OpenAI introduced the world to ChatGPT, its large language model (LLM) generative AI chatbot. Whilst precursors to the system, grounded in the field of artificial neural networks, have been in development since the late 1950s[1], the announcement was arguably transformational in two distinct ways.

First, due to the capability of the chatbot, with the general-purpose ability to both comprehend and generate language in ways that match or even surpass average human ability, a feat exemplified by its passing multiple graduate-level exams including the Uniform Bar Examination[2]. This clearly showed the potential of such models to have an immediate transformational effect on human interaction and commerce, and not in some distant future as depicted in science fiction.

Second, by making this capability freely available to the general public in a format easily accessible to those without computer science training, it opened the flood gates of human creativity and innovation. This has resulted in what some have described as a 'Cambrian explosion'[3] in applications of this technology to fields of human interest, with over 250 million weekly users at time of writing. Given the significant commercial value of procurement in general, and the highly regulated and written language dependent nature of public procurement specifically, it is only natural that this area of human activity is attracting attention from those wishing to use the capabilities of generative AI to gain a commercial advantage. Whilst multi-million pound investments have already been made into startups focused on assisting the supplier side of procurements[4], with the promise of lowering the cost of developing bids, it is the potential for this innovation to boost productivity within contract authorities that is the interest of this research.

---

[1] Frank Rosenblatt,' The perceptron: A probabilistic model for information storage and organization in the brain' (1958) 65(6) Psychological Review 386.
[2] Daniel Martin Katz and others, 'GPT-4 Passes the Bar Exam' (2023) SSRN <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4389233 accessed 12 Feb 2024
[3] Sean Leahy and Punya Mishra, 'TPACK and the Cambrian explosion of AI' (2023) Proceedings of Society for Information Technology & Teacher Education International Conference 2465
[4] AutogenAI 'Backed by Salesforce Ventures, AI Company AutogenAI Raises $39.5m series B' (2003) <https://autogenai.com/articles/salesforce-ventures-co-lead-a-39-5m-investment-round-in-autogenais-game-changing-proposal-writing-software/ accessed 12 Feb 2024

Research Question

The aim of this research is to perform a critical analysis of the legal implications of the use of generative AI for the evaluation of tender submissions under the UK's soon to be enacted Procurement Act 2023 (the Act), including associated secondary legislation, technical guidance and published procurement policies. More specifically, the research will analyse the use of generative AI to evaluate quality submissions within the award stage, as arguably many different forms of artificial intelligence are already used to support the evaluation of both price submissions (i.e. Microsoft excel) and conditions of participation (i.e. e-procurement systems). As this is an emerging area of innovation, with limited case law and legislation developed directly in response to it, the research aims to identify the areas of potential strengths and weaknesses of generative AI use for evaluation of quality responses, as well as identify approaches to its integration into the process that could potentially lower the legal risk of their use. The research follows an interdisciplinary method combining the field of Law with that of Computer Science, or more specifically the field of artificial intelligence. However, it is placed firmly in the doctrinal in context end of the interdisciplinary continuum, with both the history and detail of generative AI included to ground the subsequent legal analysis.

Importance of Research Question

This research can be viewed as significant in multiple ways. First, the widespread incorporation of generative AI for bid writing by suppliers is likely to lower the cost of bidding for public procurement contracts, by speeding up the process of text generation of responses. Following basic economic theory, this would assume that the number of bids for public contracts will increase, by reducing the barriers to entry.[5] Whilst this increased competition would appear positive from a value for money perspective, an increased number of bidders runs the risk of authorities exceeding their current capacity

[5] Jenny Alke and Joakim Hassel, 'Aspects of knowledge management applied to public bid writing: Lower the barriers of entry in public procurement by streamlining the bid writing process'(KTH Royal Institute of Technology 2023).

to evaluate, and in doing so, slow down public procurements further. Being able to utilise generative AI in evaluation of quality responses would not only mitigate this capacity risk, but also potentially unlock significant productivity gains by reducing the time and effort needed for evaluation. Providing a sound understanding of the legal basis for the use of generative AI would therefore assist in informing public agents of how best to achieve these gains.

Second, by successfully incorporating generative AI into the evaluation processes, contract authorities could be able to mitigate arguably the greatest risk of evaluation, human error. Given that evaluations are often carried out by stakeholders without deep knowledge of public procurement regulations, potentially alongside the demands of their day jobs, this high degree of manual effort can lead to potential biases and inconsistencies in application of evaluation criteria, resulting in suboptimal decisions[6]. The use of generative AI systems, which can be trained on not only legislations, but also the plethora of relevant case laws and government policies, could potentially provide an almost infallible evaluator, either as the sole evaluator of responses, or working alongside human evaluators.

Third, given that the technology is already available to the public, and the sheer number of public procurements that occur every year, it is not inconceivable that an evaluator somewhere has already surreptitiously used the software to evaluate responses. As such, the research is of importance to assist the orderly incorporation of this technology in a way that does not undermine the fundamental objective of sharing information about a contracting authority's decisions transparently that is at the core of the UK public procurement regulations.[7]

Finally, much in the same way as the introduction of eTendering tools spurred reforms in procurement law internationally[8], it is arguable that the widespread implementation of

[6] Mohammed Waseem and others, 'Artificial Intelligence Procurement Assistant: Enhancing Bid Evaluation' (2023) International Conference on Software Business 108
[7] Procurement Act 2023, s 12(1)(c)
[8] Don Wallace Jr. , Christopher R. Yukins and Jason P. Matechak, 'UNCITRAL Model Law: Reforming Electronic Procurement, Reverse Auctions, and Framework Contracts' (2005) 40 Procurement Law 12

generative AI across the procurement lifecycle will spur a new round of policy development that clarifies its usage from a legal perspective. This research could therefore be helpful for future international policy development, highlighting areas for further investigation.

Approach

To achieve its goals, the research is structured into three chapters. In order to place the analysis in context, chapter 1 focuses on a high-level overview of generative AI including: a short summary of its historical development and how it functions; the inherent limitations in the current generation of LLMs; and an overview of the UK policy response to generative AI in general. Next, chapter 2 outlines the rules, in both hard and soft law, that govern the evaluation process in the UK, and in doing so, attempts to summarise into key themes the attributes to which evaluations are required to align. Chapter 3 then provides a synthesis of the previous analysis, to identify where the current generation of generative AI systems either support or deviate from the key evaluation themes. The paper concludes with the author making a case for the incorporation of generative AI within the evaluation process, ending on a philosophical question of whether by attempting to introduce specific regulations in response to generative AI's use in procurement, one would effectively be applying higher standards to non-human intelligence than to the equally fallible human form.

Unfortunately, at the time of writing, the Procurement Act 2023 had not yet commenced following a delay to implementation date brought about by a change of UK government administration. This means that the analysis outlined in this paper is based on legislation that has not yet impacted contracting authorities' actions or been tested by the courts. In addition, there is a high likelihood that the guidance documentation may change, especially given both the stated intention of the delay and the fact that not all of the Act's technical guidance has yet been published. Whilst this is somewhat sub-optimal from a research perspective, it is this author's belief that the fundamental factors that

will impact the legality of using generative AI for evaluation of quality submissions are unlikely to be dramatically impacted by the legislative developments in the near term.

Chapter 1 An Overview of Generative AI

Generative AI, like nearly all modern technological developments, is not a product of immaculate conception. It stands on the shoulders of giants in the fields of mathematics, computer science, psychology, and linguistics. Whilst it is possible to plot the history of its development earlier, it is arguably Alan Turing, the Second World War polymath who cracked the Enigma code, who has the greatest claim to being AI's founding father. First, in a 1936 paper[9], Turing set out a hypothetical computing device, now known as a 'Turing Machine', that could perform computation by reading and writing binary code onto an infinite tape, and in doing so articulated the model for modern day computers. Next, in his groundbreaking 1950's paper 'Computing Machinery and Intelligence'[10], Turing set out to consider whether machines can think, describing a theoretical 'Imitation Game' as the threshold for this, which would later be known as the 'Turing Test'[11]. Turing theorised that if a machine can convince a human interrogator into believing that it is human when asked a series of questions, then the machine would have passed the benchmark for intelligence. Turing's work fired the starting gun for many in the field to develop programmes that attempted to clear this benchmark, such as Eliza, which used simple linguistic tricks to convince participants into believing they were talking with a human[12]. However, such approaches, whilst novel, could not in any meaningful way be viewed as intelligent.

In parallel, the field of artificial neural networks arguably began with the work of MuCulloch and Pitts, whose 1943 paper[13] claimed that networks of neurons, with their

[9] Alan Mathison Turing, 'On computable numbers, with an application to the Entscheidungsproblem' (1936) Proceedings of the London Mathematical Society Series 42
[10] Alan Mathison Turing, 'Computing Machinery and Intelligence' (1950) Mind 59
[11] Robert M French, 'The Turing Test: the first 50 years' (2000) Trends in cognitive sciences 4.3 115
[12] Joseph Weizenbaum, "ELIZA—a computer program for the study of natural language communication between man and machine." (1966) Communications of the ACM 9.1
[13] Warren S McCulloch and Walter Pitts, 'A logical calculus of the ideas immanent in nervous activity' (1943) Bulletin of Mathematical Biophysics 5

ability to fire in a binary nature, are capable of doing logical operations in a similar fashion to a Turing Machine. They outlined a simplified model of a neuron, with input weights, input values, threshold values for the neurons firing condition, and an activation function. When the result of a weighted sum of the inputs was higher than the threshold value it would activate, transmitting a binary output. Later, the psychologist Hebb proposed a theory for how these neural networks could 'learn'[14], claiming that the more neurons fired together, the stronger their connection became, increasing the weight that was associated with them, and in doing so the likelihood of exceeding the threshold.

Yet it is arguably the work of Frank Rosenblatt that brought the field from the theoretical to the practical. In 1957, Rosenblatt built the 'Perceptron', which was an analogue system that mirrored the McCulloch-Pitts Neuron[15]. This had an input layer, an association layer, and an output layer. A key development was that the weights in the Perceptron were initially randomly assigned. Rosenblatt then outlined an error correction procedure, which would change the weights within the network dependent on the result of the output layer, in essence 'teaching' the network to either increase or decrease the likelihood of firing when given a similar stimulus. Rosenblatt theorised that in the Percepton, he had demonstrated a fundamental law for all information handling systems.

However, in 1969, Minksy and Papert published 'Perceptrons'[16] which criticised Rosenblatt's work, specifically claiming that single layer Perceptrons, as a linear classifier, could not accurately classify non linearly separable problems, and that they have a fundamental scaling problem which limited their ability to deal with more complex tasks. In addition, Rosenblatt's original error correction algorithm was not suitable for use within a multi-level Perceptrons. This criticism, coupled with the technical limitations with the computational power available at the time, significantly limited the practical

[14] Donald O Hebb, '*The Organization of Behaviour*' (Wiley New York 1949)
[15] Rosenblatt (n 1)
[16] Marvin Minsky and Seymour Papert, *Perceptrons* (MIT Press 1969)

application and interest in these early neural networks. This resulted in what colloquially was termed the 'AI winter[17]' for research into artificial neural networks.

Yet, as ever with consensus opinions, there were contrarian researchers who continued to develop the field. Chief amongst these was the now Nobel Prize winning computer scientist Geoff Hinton, often called the 'Godfather of AI'[18]. Of his many contributions to the field, his 1986 paper with Rumelhart and Williams is arguably one of the most important[19]. This set out a back propagation algorithm for training multi-layered neural networks, which could adjust the weights that apply to the hidden layers within a multi-layered neural network, by computing the gradient of the error between the predicted output and the actual output, until the error was minimised.

Whilst Hinton's work proved that multi-layered neural networks, given enough training data, would be able to accurately compute more complex tasks, the computational constraints and the paucity of training data of the time continued to undermine the practical effectiveness of the neural networks that could be built. It was not until the development and widespread adoption of graphical processing units (GPUs) in the late 1990s that the first of these constrains began to fall away[20]. GPUs, originally developed for rendering graphics primarily for computer games, have multiple cores which support the parallel computation that was well suited to the development of artificial neural networks. These were then able to grow in complexity in the number of layers and nodes, which kicked off a rapid explosion in development of the field.

These developments culminated with the arguably paradigm shifting AlexNet, which brought this new computing power to bear on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) dataset of over 14 million annotated images. In their

[17] Crevier D, *AI: the tumultuous history of the search for artificial intelligence* (Basic Books 1993) 203
[18] Rannard G and Fraser G, ''Godfather of AI' shares Nobel Physics Prize' (BBC, 8 October 2024) <https://www.bbc.co.uk/news/articles/c62r02z75jyo#:~:text=The%20Nobel%20Prize%20in%20Physics,and%20said%20he%20was%20flabbergasted> accessed 29 November 2024
[19]David E Rumelhart, Geoffrey E Hinton and Ronald J Williams, 'Learning representations by back-propagating errors' (1986) Nature 323.6088
[20] Chris McClanahan , 'History and evolution of gpu architecture' (2010) A Survey Paper 3

2012 paper[21], Krizhevsky, Sutskever and Hinton, outlined how their large convolutional neural network of 60 million parameters, 650,000 neurons across 8 layers, trained on 1.2 million images across 2 GPUs, was considerably better at classifying the dataset than other leading non-neural network techniques of the time. Even with this apparent scale, when compared to the size of Rosenblatt's Perceptron, the authors highlighted that their results would likely be improved with greater computing capacity and larger datasets. Apart from this increase in accuracy, what was innovative about AlexNet, was how its convolution layers functioned, with the first layers learning to recognise simplistic features, such as shapes or edges. As one progressed through higher layers, these were able to recognise more abstract representations[22]. The success of AlexNet upended the conventional wisdom in the field of artificial intelligence and commenced a race to develop ever more sophisticated neural networks utilising the key ingredients of more computing power, variables and data.

A final further development to note before reaching current generative AI systems is the 2017 paper[23] 'Attention is all you need' by eight Google researchers, which introduced the world to a new network architecture called the Transformer. Transformers are made of two parts, an encoder and decoder, which work on the input and output sequences. Where this paper was groundbreaking was a move away from a sequential processing of previous neural networks via the use of an 'attention' mechanism. This provided context to the input sequence of natural human language by converting it into units, known as tokens, and applying weights to each dependent on the importance of its meaning to the sequence. This allowed the Transformer to focus on the most important part of the sequence, significantly increasing both the performance and efficiency. Whilst the paper's main focus was on language translation, the architecture it outlined could be

[21] Alex Krizhevsky, Ilya Sutskever and Geoffrey E Hinton, 'Imagenet classification with deep convolutional neural networks' (2012) Advances in neural information processing systems 25

[22] Robert Kozma, Roman Ilin and Hava T Siegelmann, 'Evolution of abstraction across layers in deep learning neural networks' (2018) Procedia computer science 144

[23] Ashish Vaswani and others, 'Attention is all you need' (2017) Advances in Neural Information Processing Systems

applied in many different contexts, with the ability to both understand and then generate novel text, images, music and videos that is practically indistinguishable from those created by humans. The Transformer arguably was one of the final keys that unlocked the development of the LLMs we see today. It provides the 'T' in ChatGPT's acronym of Generative Pre-trained Transformers, an LLM which reportedly used 20,000 GPUs[24] and practically the entirety of the internet as its training data. Once developed, these raw models can then go through a process of fine tuning to become specialised at a given task, such as the evaluation of tenders. Which brings us to the present day, where the latest version of ChatGPT-4 passed the Turing test in February 2024, exhibiting behaviours and personality traits that were statistically indistinguishable from humans[25]. Yet, whether this represents a crossing of the threshold for true intelligence is debatable.

What is Prompt Engineering?

Whilst so far, we have delved into the history of generative AI in general, it is worthwhile to briefly touch on the topic of prompt engineering. As has been outlined previously, LLMs are able to respond to inputs in natural human language, rather than requiring specific programming code. The clear benefit of this is that they have the potential to open up the field of computer science to a wider audience who lacked the knowledge of specific programming languages. The downside, however, is that given both the innate ambiguity of human language and the idiosyncrasies of these specific models, they can generate hugely different outcomes dependent on minor differences in inputs. The emerging field of 'Prompt Engineering' is the process of constructing input prompts in a specific way that results in the models generating the desired outputs[26]. Effective prompt engineering can help to minimise the limitations of the models discussed later in this

[24] Keumars Afifi-Sabet, 'Most formidable supercomputer ever is warming up for ChatGPT 5 — thousands of 'old' AMD GPU accelerators crunched 1-trillion parameter models' (Tech Radar 12 January 2024) < https://www.techradar.com/pro/most-formidable-supercomputer-ever-is-warming-up-for-chatgpt-5-thousands-of-old-amd-gpu-accelerators-crunched-1-trillion-parameter-models> accessed 29 November 2024

[25] Qiaozhu Mei and others, 'A Turing test of whether AI chatbots are behaviorally similar to humans' (2024) Proceedings of the National Academy of Sciences, 121(9) 1

[26] Giray Louie, 'Prompt engineering with ChatGPT: a guide for academic writers' (2023) Annals of biomedical engineering, 51(12) 2629

paper. This can be done by either focusing the attention of models onto domain specific knowledge[27] when generating its output (e.g. a specific piece of legislation) or setting out a chain of thought[28] that the prompter wants the model to follow, in essence breaking down a larger task into smaller steps, and guiding the model through these steps before it provides a final answer. Given the material impact that the construction of the underlying prompts would have on the outputs of systems built on top of LLMs, these prompts are arguably a key part of its decision-making process.

The limitations of current generative AI

Despite being hugely powerful, generative AI systems are not the infallible tools that many users often mistake them to be, potentially undermining the case for their application in human domains where the need for accuracy is greatest. In their 2024 paper, 'Easy Problems That LLMs Get Wrong'[29], Williams and Huckle set out a linguistic benchmark to evaluate the limitations of current LLM models across multiple domains including comprehension and reasoning. They highlighted nine known limitations with these models, of which five are arguably most relevant to this research: *Linguistic Understanding* – the misinterpretation or overlooking of nuanced meanings in human language; *Common Sense* – the absence of an embodied experience crucial to its development; *Contextual Understanding* – a failure to understand the implicit context in which something relates; *Popular Misconceptions* – the vulnerability to propagating and reinforcing inaccuracies found in their training data, such as misconceptions and outdated information commonly perpetuated online; and *Logical Reasoning* – with results generated based on statistical likelihood rather than by applying logical steps. On applying their benchmark to leading models of the time, they found the average accuracy performance score of the best model to be 38% when compared to a human score of 86%. They did however find that by refining the wording of inputs via prompt

[27]Marton Ribary and others, 'Prompt Engineering and Provision of Context in Domain Specific Use of GPT' in Legal Knowledge and Information Systems (IOS Press 2023) 306
[28] Andrew Gao, 'Prompt engineering for large language models' (2023) SSRN 4504303 4
[29] Sean Williams and James Huckle, 'Easy Problems That LLMs Get Wrong' (2024) arXiv preprint 2405.19616.

engineering, they could achieve significantly more accurate outcomes. However, this only increased the highest accuracy score to 52%[30]. They concluded by cautioning on the commercial use of LLMs for tasks '*that require high-stakes decision-making, nuanced reasoning, or understanding subtle linguistic cues unless supervised or complemented by human judgment*'[31]. Whilst William and Huckle's paper is an excellent overview of the more internal issues of generative AI systems, there are also significant strategic issues that need to be understood and overcome before their effective implementation.

First is the potential issue of 'Lock-in', where contracting authorities end up being technologically and operationally dependent on either a single or limited group of suppliers[32]. Given that the most capable generative AI systems are all developed by profit driven corporations, mostly from the USA, there are risks involved in integrating these systems into public sector delivery and decision making. Although arguably, this is a risk that is live today for many other areas of technology deployment across government. Then, given the current reliance on a handful of companies, there is the non-trivial risk of conflict of interest. This is where the underlying model is biased in its decision making, either intentionally or unintentionally, to favour its own organisation or that of connected or favoured firms. This risk is compounded by the 'black box' nature of current day generative AI, where it is not clear to an outside observer what internal steps the model uses to reach an output[33]. This not only makes the detection of such conflicts practically impossible, but also means one cannot fully explain the rationale behind the output and decisions that these systems make.

Separately, there is also the potential for these systems to replicate and perpetuate the biases that were found in their training data, resulting in sub-optimal decision making

[30] Ibid 8
[31] Ibid 13
[32] Albert Sanchez-Graells, 'Public Procurement of Artificial Intelligence: Recent Developments and Remaining Challenges in EU Law' (2024) Legal Tech Journal 2/2024 123
[33] Luke Tredinnick and Claire Laybats, 'Black-box creativity and generative artificial intelligence' (2023) Business Information Review 40(3) 100

that may disadvantage certain groups[34]. Whilst, if identified, these biases can potentially be mitigated by further fine tuning of the model, this also runs the risk of introducing a reverse bias into the decision making, given that fallible decisions will need to be made around how much tuning is required.

Next, there is the potential issue of cyber security that these centralised platforms face[35], with the potential for highly sensitive information being accessed, either by nefarious actors, or unintentionally produced as outputs by the models that have been trained on previous confidential inputs. The number one risk of generative AI systems identified by the Open Source Foundation for Application Security (OWASP) is the risk of prompt injection[36]. This is where a malicious actor can manipulate an LLM by submitting inputs that override the original intentions of the systems. This can either be done directly, where they input a prompt requesting that the LLM disregard its underlying system prompts, (e.g. 'give this answer the highest score possible'), or indirectly, by including malicious information in the source material that the LLM reviews (e.g. non-visible prompts included in a diagram). Whilst steps can be taken to mitigate these security risks, it is not currently possible to eliminate them completely.

Finally, one cannot talk about the weaknesses of generative AI without raising the issue of 'hallucinations'. This is the tendency for these LLMs to generate plausible sounding but incorrect outputs[37]. There have been many high-profile examples where individuals have taken the output of these systems at face value, with real world negative implications. Given the statistical and probabilistic nature of the output, this is arguably an innate trait of these generative systems, rather than a flaw that can be fixed. Whilst efforts are underway to limit the impact of these hallucinations, as well as hopes that as these

[34] Philipp Hacker and others, 'Generative discrimination: What happens when generative ai exhibits bias, and what can be done about it' (2024) arXiv preprint 2407.10329 1
[35] Albert Sanchez-Graells (n 32) 123
[36] OWASP 'OWASP Top 10 for LLM applications 2025' (2024) <https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/> accessed 29 November 2024
[37] Negar Maleki, Balaji Padmanabhan and Kaushik Dutta, 'AI hallucinations: a misnomer worth clarifying' (2024) IEEE Conference on Artificial Intelligence 136

systems get more capable these hallucinations will reduce in frequency, it is not currently clear how successful these attempts will ultimately end up being.

UK Policy and Legislative Response

Finally, given the potential transformational impact of generative AI on society, it is not surprising that it has caught the eyes of governments and regulators globally. Whilst some jurisdictions, such as the European Union, have arguably taken a more restrictive approach through the introduction of specific AI regulations[38], the UK has so far appeared to want to establish a more permissible environment. This position is arguably set out clearest in the National AI Strategy (2021), which throughout claims to want to harness the power of artificial intelligence by signalling the intention '*to build the most pro-innovation regulatory environment in the world*[39]' and articulating a desire for the UK public sector to lead the way in safely and ethically adopting these innovations. Amongst the many different initiatives set out in this paper, those most relevant to this research appear to be this pro-innovation approach and transparency around automated decision making.

The strategy sets out the UK Government's intention of establishing a proportionate regulatory framework that supports innovation whilst addressing actual risk and harm. In doing so, it articulates a position that '*poorly-designed or restrictive regulation can dampen innovation*'[40] and in doing so, impede growth. It argues against the blanket approach to AI regulation akin to the EU AI Act, stating that many existing regulations (e.g. data protection) already apply to the systems, and that '*existing sector-specific regulators are best placed to consider the impact on their sectors*'[41]. The strategy claims this to be an optimal approach to regulation due to the highly complex use cases and impacts that the incorporation of AI will have and the significant grey area in determining the risks and harms caused by it. However, given that procurement

[38] Artificial Intelligence Act [2024] OJ L 1689
[39] The Office for Artificial Intelligence, National AI Strategy (Command Paper 525, 2021) 5
[40] Ibid 51
[41] Ibid 52

processes don't have a specific regulator, it seems like there is a potential gap in this approach when applied to this field, suggesting that some additional targeted reforms and regulations may be of benefit in this area.

The strategy then highlights the Central Digital and Data Office (CDDO) work to develop a cross-government standard for algorithmic transparency[42]. This follows on from recommendations from the Commission on Race and Ethnic Disparities that there be mandatory transparency obligations on all public sector organisations applying algorithms that have an impact on significant decisions affecting individuals[43]. This standard sets out a two-tier level of transparency split between non-technical and detailed technical description. As the evaluation process is an innately decision making one, this standard would likely impact generative AI's use in procurement.

Finally, whilst procurement is mentioned within the strategy, with statistics included from an AI Council survey highlighting a 72% agreement rate that government should take steps to increase buyer confidence and AI capabilities[44], the main focus of the strategy appears to be on the procurement of AI, rather than the use of AI in procurement. This is a focus that is repeated in the Procurement Policy Note (PPN) covered in the next chapter.

To conclude this chapter, whilst generative AI has a long history, the recent significant gains in capabilities means that it is breaking into more fields of human endeavour. The approach of the UK to date of taking being more permissible, whilst also requiring a high degree of transparency, appears supportive, at least at a strategic level, of the use of generative AI within the procurement process. However, it must be noted that whether this remains the case following the change in administration is yet to be determined. Yet, in order for contracting authorities to harness the potential benefits from these systems

[42] Ibid 59
[43] Commission on Race and Ethnic Disparities, Commission on Race and Ethnic Disparities: The Report (CCS0221976844-001, 2021) 12
[44] National AI Strategy (n 39) 46

in the evaluation process, one first needs to understand what rules governing this process these systems would need to comply with.

Chapter 2 – UK rules on evaluation

Before diving into the analysis of the rules, it is worthwhile setting out what is meant by evaluation in public procurements. A compelling explanation was provided by the Government Commercial Function (GCF) in 2021, which stated that '*bid evaluation is the process of assessing bids to identify the most economically advantageous tender (MEAT) submitted for a project*'.[45] This guidance note then goes on to state that *"A good evaluation is not only about the final award decision - it is about the design and execution of the whole process… including ensuring the process is properly documented and can stand up to internal and external scrutiny".*[46] Whilst the terminology used in this explanation, specifically the use of MEAT, is outdated when compared to the Act's use of 'Most Advantageous Tender' (MAT)[47], the continued relevance of this would appear to be supported by the fact that this guidance note is directly referenced in the 'Assessing Competitive Tenders' technical guidance[48].

Given the pivotal role that the evaluation process plays in the decision-making process of awarding contracts, it is unsurprising that it is subject to significant regulation in public procurement regimes globally. The rules governing the evaluation of tenders in the UK are set out in both hard law, in the Procurement Act 2023 and Procurement Regulations 2024 (the Regulations), and soft law, in a range of technical guidance documents and procurement policies. This chapter provides a summary of the key requirements across each of these and attempts to encapsulate them into a series of key evaluation themes to be used as the basis for subsequent analysis.

Procurement Act 2023

[45] Government Commercial Function, Bid Evaluation Guidance Note, (OGL 2021) 4
[46] Ibid 4
[47] Procurement Act 2023, s 19 (1)
[48] Government Commercial Function, Guidance: Assessing Competitive Tenders (OGL 2024) 9

The sections of the Act that are most relevant to the evaluation process are arguably: Section 1 (Procurement and covered procurement); Section 3 (Public contracts); Section 11 (Covered procurement only in accordance with this Act); Section 12 (Covered procurement: objectives); Section 19 (Award of public contracts following a competitive tendering procedure); Section 23 (Award criteria); Section 50 (Contract award notices and assessment summaries); Section 56 (Technical specifications); Section 81 (Conflicts of interest: duty to identify); Section 82 (Conflicts of interest: duty to mitigate); Section 95 (Notices, documents and information); Section 98 (Record-keeping); and Part 9 (Remedies for breach of statutory duty).

Section 1 outlines the differences between a 'procurement' and a 'covered procurement'. This makes clear that a 'covered procurement' is a subset of procurements for '*the award, entry into and management of a public contract'*. [49] Section 3 defines a 'public contract' as a contract for the supply of goods, services or works which *'(a) has an estimated value of not less than the threshold amount for the type of contract, and (b) is not an exempted contract*.'[50] The same definition is also stated for both frameworks and concessions contracts. Section 11 (1) then states that *'A contracting authority may not carry out a covered procurement except in accordance with this Act'.*[51] These three sections combined would appear to support the view that the general rules related to 'covered procurements', and crucially for this analysis, those related to the evaluation of tenders, do not apply to procurements that are not 'covered procurements', the first key evaluation theme of importance to the later analysis within this paper. This view is reinforced by the fact that the rules governing 'regulated below-threshold contract procedures' in Section 85[52], themselves a subset of non-covered procurements, make no reference to either award criteria or assessment methodologies, two fundamental concepts that will be discussed later in this chapter.

[49] Procurement Act 2023, s 1 (1)
[50] Ibid, s 3 (2)
[51] Ibid, s 11 (1)
[52] Ibid, s 85

Next, Section 12 sets out the objectives for all covered procurements, which contracting authorities must either have regard to, or, in the case of the same treatment objective, must follow. Of these, the objectives arguably most relevant to the evaluation process are; *'sharing information for the purpose of allowing suppliers and others to understand the authority's procurement policies and decisions'*[53]- given that the evaluation process is at its core a decision-making process, '*acting, and being seen to act, with integrity'*[54]- given that the selection of preferred bidders during an evaluation has the potential to be corrupted by procurement agents, or be perceived as being corrupted*,* and *'treat suppliers the same unless a difference between the suppliers justifies different treatment*[55] - given that an evaluation process that did not treat suppliers the same would undermine fair competition*.* It is therefore argued that any evaluation process for a covered procurement, either human or AI, would need to demonstrate its alignment with these objectives to have confidence in being compliant with the Act if challenged. Being able to therefore evidence alignment with the objectives of the Act is identified as the second key evaluation theme within this paper.

Continuing, Section 19 sets out the rules that govern the award of public contracts following a competitive tendering procedure. This introduces the concept of the 'Most Advantageous Tender', and defines this as the tender that *"(a) satisfies the contracting authority's requirements, and (b) best satisfies the award criteria when assessed by reference to—(i) the assessment methodology under section 23(3)(a), and (ii) if there is more than one criterion, the relative importance of the criteria under section 23(3)(b)."*[56] It is these dual concepts of 'award criteria' and 'assessment methodology', a concept that was not explicitly defined in the previous UK regulations, that are the key to understanding the rules governing evaluation, and also the feasibility of using generative AI. This is because the ability of AI to both comprehend the stated award criteria and to

[53] Ibid, s 12 (1) (c)
[54] Ibid, s 12 (1) (d)
[55] Ibid, s 12 (2)
[56] Ibid, s 19 (2)

rationally apply the assessment methodology would need to be evidenced in a court of law in order for any system to be compliant. Both of these concepts are defined in more depth in Section 23, but in the author's view in an insufficient level of detail to provide legal clarity. Award criteria are defined as '*Criteria set in accordance with this section against which tenders may be assessed for the purpose of awarding a public contract under section 19'*[57]*.* This would appear to be a somewhat unsatisfying and autological description, given that it includes within it the properties it is attempting to define. The Act then goes on to outline a number of considerations that a contracting authority would need to satisfy when setting the award criteria, including that it relates to the subject matter of the contract, are sufficiently clear measurable and specific, do not break the rules on technical specifications and are an appropriate means of assessing the tender.[58] However, it is arguably section 23(3)(a) that is of paramount importance in relation to this research, as it states that contracting authorities must *'describe how tenders are to be assessed by reference to them and, in particular, specify whether failure to meet one or more criteria would disqualify a tender (the "assessment methodology")*[59]. It is fascinating that, for its centrality in understanding of evaluations, a process that is so integral to the award of contracts, the definition on this methodology appears so skeletal in nature. Whilst this can be argued as leaving a high degree of flexibility for contracting authorities, an objective that has been repeated throughout policy development of the Act[60], it does leave significant ambiguity as to how detailed the description of the methodology would have to be. Now, whether this ambiguity is problematic in practice will be dependent on the availability and clarity of formal guidance provided on the topic, something we will investigate later in this paper. But regardless, this general point of transparency of an assessment methodology and the decision-making process is identified as the third key theme.

[57] Ibid, s 23 (1)
[58] Ibid, s 23 (2)
[59] Ibid, s 23 (3) (a)
[60] Cabinet Office, Green Paper: Transforming Public Procurement, CP556 (The Stationery Office 2020) 5

Next, after completing the evaluation process and prior to entering into a public contract, Section 50 sets out the requirements for contracting authorities to publish both assessment summaries to suppliers of assessed tenders, and contract award notices to the general public. Of key relevance to the topic of this research is paragraph 4, which states *"An "assessment summary" means, in relation to an assessed tender, information about the contracting authority's assessment of (a) the tender, and (b) if different, the most advantageous tender submitted in respect of the contract."*[61] Yet again, the Act itself can be argued as providing limited details into what information should be included, although at least in this case this is further outlined in Regulation 31, which will be discussed later. This appears to reinforce the author's view of transparency as a key evaluation theme, also aligned to the 'sharing information' objectives. However, all of these appear to leave open the question of the level of transparency required in order to be compliant, which is identified as an additional sub-theme of transparency relating to the 'depth of information'.

Given that it was directly called out in Section 23, it is worthwhile briefly focusing on the rules governing technical specifications set out in Section 56. This states that procurement documents should, where possible, be constructed in a way that refers to their performance or functionality, rather than to a '*design, a particular licensing model or a description of characteristics*[62]*.* There are also prohibitions on referring to UK standards without allowing for an internationally recognised equivalent if ones exists. In addition, requirements must not refer to *'(a) trademark, trade name, patent, design or type, (b)place of origin, or (c) producer or supplier'*[63] unless it is necessary to make the requirements understood, and if so must ensure that equivalent quality or performance will not be disadvantaged.[64] It then defines procurement documents, as including both the process and criteria for the award of contracts[65]. This requirement for the acceptance

[61] Procurement Act 2023, s 50 (4)
[62] Ibid, s 56 (2)
[63] Ibid, s 56 (7)
[64] Ibid, s 56 (8)
[65] Ibid, s 56 (9) (b)

of equivalence is therefore identified as a key evaluation theme, as any evaluation processes that disadvantaged a supplier offering such would appear to be in breach.

Next, Section 81 outlines the requirement for contracting authorities to take all reasonable steps to identify both actual and potential conflicts of interest.[66] This defines a conflict of interest as occurring where either *'(a) a person acting for or on behalf of the contracting authority in relation to the procurement has a conflict of interest, or (b) a Minister acting in relation to the procurement has a conflict of interest'*[67]. It is arguably the first leg that is most relevant to this research, given that it is unlikely a Minister will be replaced by generative AI in the near future, regardless of potential productivity gains. The Act then goes on to state that any *'person who influences a decision… is to be treated as acting in relation to the procurement'.*[68] This appears to provide a very broad coverage of conflict of interest requirements on individuals involved even tangentially with a procurement decision. In the event that a conflict is identified, contracting authorities have a duty to take all reasonable steps to mitigate it under Section 82. Should it be considered that a supplier has an unfair advantage due to this identified conflict, and that it cannot be mitigated, then the supplier must be treated as an excluded supplier from the procurement.[69] It is the potential ramifications for both competition in general, and supplier exclusion specifically, that leads this author to view conflicts of interest as being another key evaluation theme within this research. Although, it is noteworthy to highlight that the requirement only relates to a 'person'.

Section 98 sets out the requirements for contracting authorities to *'keep such records as the authority considers sufficient to explain a material decision'*[70]*.* The keeping of such records is generally a fundamental pillar of public procurement regulations, as their disclosure supports the seeking of justice by suppliers wishing to challenge an outcome.

[66] Ibid, s 81 (2)
[67] Ibid, s 81 (2)
[68] Ibid, s 81 (3)
[69] Ibid, s 82 (4) (a)
[70] Ibid, s 98 (1)

Yet, in relation to this research, it is this author's view that the key theme is not simply the keeping of records of decision, which is arguably as easily achievable, if not more so, when utilising generative AI as those done by humans. Instead, the key theme is whether such records are 'sufficient to explain', and whether both the generative AI's probabilistic output and the current black-box nature of these systems would ever make records of decisions sufficient to explain.

Finally, Part 9 of the Act covers the remedies that suppliers can seek where a contracting authority has breached its obligations, in Sections 100 to 107. This states that the duty to comply with the Act only applies to either UK suppliers or treaty state suppliers,[71] and that proceedings can only be brought by that supplier where they have suffered, or been at risk of suffering, loss or damage as a consequence of that breach.[72] This arguably is supportive of, and a subset of, the first evaluation theme of 'covered procurement', as the general obligations in the Act, and those to do with evaluation specifically, do not appear to apply to non-UK / non-treaty state suppliers.

Next, Section 102 (1)(c) gives the courts powers to make *'an order suspending the effect of any decision made or action taken by the contracting authority in carrying out the procurement',*[73] as a form of interim relief. Given, as previously stated, the evaluation process is fundamentally a decision-making process, it seems clear that the courts would be able to challenge and overturn a decision-making process that it viewed as not compliant with the Act. The Act then sets out similar abilities for the court to intervene with both pre-contractual and post-contractual remedies. Combined, this ability to overturn decisions that do not conform with the obligations of the Act is identified as the next, and arguably one of the most important evaluation themes.

Lastly, Section 106 outlines the time limit that claims can be made by suppliers to the courts. This is generally set at *'30 days beginning with the day on which the supplier first*

[71] Ibid, s 100 (2)
[72] Ibid, s 100 (3)
[73] Ibid, s 102 (1) (c)

knew, or ought to have known, about the circumstances giving rise to the claim"[74], albeit set-aside proceedings have an additional leg of six months beginning the day the contract was entered into[75]. It is this author's view that these requirements are an additional subset of the previously identified 'transparency' evaluation theme, acting as an incentive for authorities to publish relevant information early to reduce legal risks.

To conclude, whilst the review of the primary legislation has brought to light a number of key themes for further analysis, it is argued that it has also identified a general deficit in detail required to reach thorough and substantial conclusions. It is therefore logical for the author's attention to move onto the secondary legislation, the Regulations, in an attempt to unearth this detail. Unfortunately, this also appears to be lacking.

Procurement Regulations 2024

A review of the Regulations identified the following as arguably the most pertinent to the analysis within this paper: Regulation 16 (Planned procurement notice); Regulations 18 to 21 (Tender notices); Regulation 23 (Associated tender documents); Regulation 24 (Below-threshold tender notices); Regulation 31 (Assessment summaries); Regulation 33 (Contract Details Notice: Frameworks).

Regulation 16 sets out the information that must be included in a planned procurement notice. This notice, which contracting authorities are not mandated to publish, informs the market of its intention to launch a procurement. Whilst it does not expressly require contracting authorities to provide information in relation to the award criteria or assessment methodology, section 16(2) requires contracting authorities to provide *"as much of the information relating to tender notices which is referred to in regulation 18(2), 19(2), 20(2), 21(2) or 22(2) (as the case may be) that is available to the contracting authority at the time of publishing."*[76] As discussed later, these sections

[74] Ibid, s 106 (2)
[75] Ibid, s 106 (1) (b)
[76] Procurement Regulations 2024, s 16 (2)

include reference to publishing the 'the award criteria'[77]. Therefore, if a contracting authority at the time of publishing this notice has made decisions relation to its award criteria or assessment methodologies, such as the use of generative AI, then the details of this would seemingly need to be provided within the planned procurement notice.

Next, Regulations 18 sets out what must be included in tender notices for the open procedure. These are notices that invite suppliers to participate in a procurement. Specifically of interest to this research, paragraph 2(k) requires the notice to include *'the award criteria, or a summary of the award criteria, for the public contract.'*[78] It is somewhat ambiguous as to whether this is just the award criteria themselves, or if it also includes the assessment methodology, which is not referenced explicitly. This position is confused as, section 23(3)(a) of the Act requires contracting authorities, in setting award criteria, to describe the "assessment methodology"[79], appearing to intrinsically link the two concepts together, a position further supported by the drafting in regulation 31 which repeatedly refers to *'the award criteria, including the assessment methodology'*.[80] It is proposed that, given the overarching objective of sharing information outlined in section 12 of the Act, and the linking of the two concepts throughout both the Act and the Regulations, the assessment methodology would need to be included either in the tender notice itself, or be included in the associated tender documents published alongside the notices.[81] Assuming this is the case, then the same requirements to publish information on the award criteria and assessment methodology would apply to tender notices for competitive flexible procedures, as regulation 19(2) requires the same information be provided as regulations 18(2). It would also apply to tender notices for frameworks and dynamic markets, as the requirements for each of these points to either regulation 18(2) or 19(2). Interestingly, this information is not explicitly referenced in the regulations related to tender notices for qualifying utilities

[77] Ibid, s 18 (2) (k)
[78] Ibid, s 18(2)(k)
[79] Procurement Act 2023, s 23 (3) (a)
[80] Procurement Regulations 2024, s 31 (i)
[81] Ibid, s 18 (u)

dynamic market notices,[82] in keeping with the general increased flexibilities that utilities suppliers have under the Act. Finally, Regulation 23 outlines that the information required for tender notices may instead be included in the associated tender documents as a supplement to the Tender Notice[83].

Continuing, Regulation 24 sets out the information to be provided in a below-threshold tender notice. These are notices that must be published when a contracting authority wishes to advertise a non-covered procurement that is below the Act's thresholds. Of relevance to this research is the requirement that the notice must contain '*an explanation of the criteria against which the award of the contract will be assessed*'.[84] It is interesting to consider why this apparently more ambiguous wording has been used, rather than the previously discussed terminology of award criteria and assessment methodologies. This appears to the author to place less onerous requirements in this area than for a covered procurement. This arguably modifies the first key evaluation theme of 'non-covered procurement', in that at least some consideration around what should be included in this explanation will be required.

Regulation 31 is perhaps the most relevant to this research, as unlike the regulations reviewed so far, it sets out further granular details of what must be included in relation to the award criteria within the assessment summaries. Paragraph 2(i) explicitly requires information on '*(i)the award criteria, including the assessment methodology, set out in full, or a summary of the award criteria including—(aa) the title of each criterion, (bb) the relative importance of each criterion, and (cc) how each criterion was to have been assessed by reference to scores and what scores were to have been available for each criterion*'[85], as well as where the full version of the information can be accessed. This requirement for the information to be set out in full, is of paramount importance to this analysis, given that it modifies the evaluation theme of transparency, significantly

[82] Ibid, s 22
[83] Ibid, s 23 (2)
[84] Ibid, s 24 (2) (i)
[85] Ibid, s 31 (2) (i)

reducing the ambiguity as to the level of detail required. However, as we will see in the next chapter, this could result in practical difficulties and unintended consequences when determining what 'full' details are required when considering generative AI evaluation.

Finally, regulation 33 sets out the information that must be included in a contract details notice for the establishment of a framework. Of particular interest is paragraph (2)(j) which requires inclusion of '*details of the selection process to be applied on the award of a contract in accordance with the framework*[86]'. This information is not mirrored in the requirements for contract details notices for contracts awarded under open or competitive flexible procedures in regulation 32. This appears to make logical sense, given that a framework agreement is a contract for the award of future contracts, and how those subsequent contracts would be awarded is of significant consequence. Again, the use of different terminology of selection process, which is not defined within the Regulations, rather than the previously used award criteria and assessment methodology is notable, potentially providing more discretion in this area. However, it is this author's view that this will likely encompass similar requirements to both these terms, especially when read alongside the overarching objectives of the Act in section 12.

On completing the review of the Regulations, they appear to provide greater clarity as to what should be covered in the key evaluation theme of transparency as well as when in the process this information should be provided. However, when considered in their entirety, the hard law governing the evaluation process appears skeletal in nature, with very little detail on how these processes should be executed by contracting authorities. This lack of details provides a significant degree of ambiguity, which arguably leaves open the potential for using generative AI within evaluation, so long as they were done in a transparent way, and taking account of the other themes identified. However, whilst this detail is lacking in the hard law, much of it is instead filled in by the soft law, in a series of technical guidance documents and procurement policies. Since the 2021 ruling

[86] Ibid, s 33 (2) (j)

of *Good Law Project v SoS H&SC*, contracting authorities would likely have a common law duty to comply with published government guidance '*absent good reason for depart from it*'[87]. A thorough review of these documents is therefore required.

Technical Guidance

In order to assist contracting authorities in both understanding and implementing the Act, Cabinet Office has published a comprehensive set of technical guidance documents. Whilst many of these documents contain information that relate to the topics already discussed, a thorough review highlighted the following as arguably the most relevant: Assessing Competitive Tenders; Assessment Summaries; and Covered Procurement Objectives. The subsequent analysis will focus on where these appear additive.

Assessing Competitive Tenders

In Assessing Competitive Tenders, it is stated that the basis for the award of contracts is largely unchanged from that of the previous regulations, claiming that the move from MEAT to MAT previously discussed does not represent a change in policy, rather one of emphasis[88]. It does however highlight that contracting authorities now have the ability to refine the award criteria during a competitive flexible procedure within certain circumstances[89]. Whilst an interesting reform, this does not appear consequential to this research. The document then goes on to lay out the key points and policy intent of the rules in this area. In line with previous analysis, it highlights the need for contracting authorities to have regard to the objectives in the Act when designing and applying award criteria.[90] However, it is interesting to note that the objectives that it claims as being of particular relevance are different from the ones argued in this paper, instead initially highlighting maximising public benefit, delivering value for money and the duty to have regard to the barriers that SMEs face. Later it does highlight the objective of

[87] Good Law Project Limited v Secretary of State for Health and Social Care [2021] EWHC 346 (Admin) [2021] ACD 49 153
[88] Government Commercial Function, Guidance: Assessing Competitive Tenders (OGL 2024) 2
[89] Ibid 2
[90] Ibid 3

sharing information – specifically highlighting that this means '*sharing all relevant information at the earliest opportunity*'[91]. Whether this difference in emphasis would have any practical impact is unclear, but it does appear to align with the general government position of reforming procurement regulations to provide a greater focus on value for money[92]. On this point of value for money, the guidance then states that '*value for money can be directly affected by the choice of assessment methodology and contracting authorities should undertake appropriate scenario-testing to understand the impact of different methodologies*'[93]*.* This appears to support the inclusion of an additional sub-theme under the previously identified 'objectives' theme, that of a value for money review of the chosen assessment methodology. The guidance then states that contracting authorities have wide discretion when selecting award criteria, that no list of criteria has been included within the Act, and that any criteria that meet the requirements in section 23 can be considered[94]. This appears to support the view that the lack of detail previously highlighted is an intentional choice of the drafters to provide greater flexibility to contracting authorities.

It is on the subject of assessment methodology that the guidance is most relevant to this analysis. The guidance provides a single example of what information must be set out. It states that '*this will include any scoring matrices to be used by evaluators when assessing tenders against the award criteria*'[95]. Whilst this drafting clearly is not meant to be exhaustive, it is of interest that there is no mention of the process that must be followed (e.g. moderations), focusing instead only on the written inputs into the assessment process. Furthermore, the guidance highlights Section 21(5) of the act, which '*requires the contracting authority to provide information sufficient to allow suppliers to prepare tenders in the tender notice or associated tender documents*'[96]*.* This

[91] Ibid 3
[92] Cabinet Office (n 60) 7
[93] Government Commercial Function, Guidance: Assessing Competitive Tenders (OGL 2024) 4
[94] Ibid 4
[95] Ibid 4
[96] Ibid 5

suggests that, whilst the award criteria and scoring matrices need to be provided, the nature of the evaluator, be they human or generative AI, may not need to be included.

Assessment Summaries

This guidance states that, where a contracting authority has carried out a competitive tendering procedure, they must issue an assessment summary '*which provides information to enable a relevant supplier to understand why its tender was either successful or unsuccessful*'[97]. It explains that assessment summaries fulfil broadly the same function as standstill letters under the previous regulations, although now contracting authorities are not required to make a direct comparison between suppliers' assessments.[98] In seeking to explain the key points and policy intent, the guidance highlights that the assessment summary aims to ensure unsuccessful suppliers can *'see how the contracting authority has determined the most advantageous tender (MAT) in accordance with the award criteria and assessment methodology'*[99]*.* It then states that the intention of the regulations is to provide a level of consistency of the information provided across all procurement, regardless of the contracting authority. It makes clear that there is no obligation to provide assessment summaries to suppliers that have not submitted assessed tenders, but that these suppliers should be informed in writing as soon as reasonably possible as to why they were not taken forward.[100] Of interest, given the latter was not highlighted in the guidance note of assessing competitive tenders, is that it claims that the structure of the assessment summary has been designed having regard to the objectives of sharing information and acting and being seen to act with integrity.[101] It then states that full details of the award criteria and assessment methodology do not need to be provided in the assessment summary, only a summary, as this information should have been provided in or alongside the tender notice.[102]

[97] Government Commercial Function, Guidance: Assessment Summaries (OGL 2024) 2
[98] Ibid 2
[99] Ibid 3
[100] Ibid 3
[101] Ibid 3
[102] Ibid 4

What appears significant to this research is that the guidance expands on the Regulation 31 requirement to provide an explanation of the score by '*reference to relevant information in the tender*'[103]. It states that this '*requires the contracting authority to make a judgement as to the appropriate level of detail to provide*'[104], and then sets out the guiding principles that the assessed tender should be recognisable from the information provided. This appears to provide contracting authorities with significant scope in determining the appropriate level of detail to provide.

Finally, the guidance sets out a number of best practice principles that are not mandated in either the Act or the Regulations, such as the provision of information on why a higher score was not awarded for each criterion[105], that the assessment summary should address the requirements of each criterion 'as fully as possible'[106], and using a similar approach to assessment summaries to call-offs under a framework.[107] Each of these are supportive of the key theme of transparency already identified, yet their status as best practice rather than requirements mean that they do not need to be followed.

Covered Procurement Objectives

To conclude the section on the technical guidance, it was the author's original intention to review the guidance on the covered procurement objectives, given the apparent importance of these for interpreting all the other obligations with the Act. Unfortunately, at the time of writing, this guidance has not yet been published. This is an unsatisfactory position, and leaves unanswered many questions related to these, such as whether, when there are clashes between the objectives, there is a hierarchy that contracting authorities need to consider, or ultimately 'how much regard' is appropriate for a contracting authority to take. Some of the finer conclusions of this research may need to be amended following publication of this guidance.

[103] Procurement Regulations 2024, s 31 (2) (e) (i)(aa)
[104] Government Commercial Function, Guidance: Assessment Summaries (OGL 2024) 5
[105] Ibid 6
[106] Ibid 6
[107] Ibid 9

Despite this, and as anticipated given the stated objectives for their publication, the review of the technical guidance has provided some additional details as to what needs to be considered as part of the evaluation process. They appear to reinforce the position previously raised of significant discretion and judgement on the part of the contracting authority on both the manner in which the evaluation process is practically implemented and the level of detail that is provided in meeting the requirements of the Act.

To conclude this chapter, the author will broaden out the soft law review by analysing two wider procurement policy documents that appear relevant to the research.

Relevant Procurement Policies

Outside of the Regulations, the UK has historically published targeted procurement policy documents, either as Procurement Policy Notices (PPNs) or in the form of more general guidance notes as a way of influencing the activities of contracting authorities. Whilst many of these are due to be updated to align with the requirements of the Act, the two that appear most relevant to this research are the Bid Evaluation Guidance Note and PPN 02/24 Improving transparency of AI use in Procurement.

Bid Evaluation Guidance Note

Published in May 2021, this note was intended to be a collection of good practice guidance to support the compliant delivery of evaluations during the award phase of a procurement. Whilst, as previously mentioned, the focus was compliance with the previous UK Regulations, its continued relevance is supported by its inclusion within the assessing competitive tenders technical guidance. Of interest is the comprehensive detail that this document contains on the process of evaluation, especially when compared to the technical guidance, providing practical recommendations across the entire process. Arguably most relevant to this research are the chapters on sufficiency of resources, evaluating quality and applying the evaluation model.

First, on sufficiency of resources, the guidance states that the evaluation process can take as little as a month to complete for a simple process, but often much longer for

complex ones[108]. It identifies that at least two evaluators should independently score the tenders, and that it is best practice that the same evaluators score all the responses to the same question. It then outlines that an additional moderator should be appointed to facilitate the reaching of a compliant consensus score. All of these individuals require training in both the regulations, the specifics of the tender, the evaluation process and appropriate record keeping. All together this guidance suggests that, even for a simple procurement, the resource requirement to run a best practice evaluation is significant, and that's without taking account of the number of tender responses received. Given the potential for generative AI to materially reduce this resource requirement, as well as the previously discussed potential need to undertake a value for money review on the choice of assessment methodology, this appears supportive of the case for its use. Therefore, the sufficiency of resources is identified as complementing the value for money assessment sub-theme within the objective's key evaluation theme.

Next, given that the title of section 6 is Evaluating Quality, one would expect this to be highly relevant to this research. Yet, on review, this section is mostly focused on the creation of compliant award criteria, rather than on the running of the evaluation itself. However, it does state that contracting authorities should *'Consider whether evaluators should be allowed to take account of information contained in one question response when evaluating a separate question response[109]'.* This consideration could have a direct impact on the previously discussed process of prompt engineering built into a generative AI evaluator, and the transparency requirements that this might entail, as the models could potentially need to separately evaluate each question without using previous question submissions as part of its domain of knowledge.

Finally, under applying the evaluation model, the guidance reiterates the position previously discussed that *'Bids must be evaluated in accordance with the published*

[108] Government Commercial Function, Bid Evaluation Guidance Note, (OGL 2021) 7
[109] Ibid 16

criteria and evaluation methodology.[110]' It then goes on to emphasise the importance of a written record which would demonstrate this, in the event that the award was challenged. Of some interest is the suggestion that bidders' identities should not be anonymised during this process, due to the practical challenges in ensuring this was genuinely achieved. The guidance then sets out the process of moderation, a key practical step in the evaluation which has been conspicuous in its absence in both the hard and soft law reviewed so far. The guidance lists that the role of the moderator is to identify and challenge a range of potential issues within individual evaluator's evaluation reports, from the inappropriate use of language through to poor quality and inconsistency of scoring and justifications[111]. It then states that moderators should chair a meeting of evaluators to agree a single consensus score per question. Once these consensus scores are reached, this is identified as the end of the scoring process, with scores not amended unless there are exceptional circumstances[112]. Given how important the role of the moderator is in ensuring both the compliance of the evaluation and in reaching the final consensus scores, it is identified as the final key evaluation theme. It is notable that this vital part of the evaluation process is not covered in either the Act, the Regulations or the technical guidance. It would seem to be somewhat inadequate for the dissemination of best practice in this area to be reliant on a legacy guidance note that is likely to be missed by practioners focused purely on the new legislation.

Improving transparency of AI use in Procurement

Published in March 2024, PPN 02/24[113] was the first procurement policy note to explicitly focus on the phenomenon of AI. On reviewing this relatively short note, its primary purpose appears to be to act firstly as a signpost towards a range of different AI guidance and best practice set out in Annex A. Its second, and probably of equal

[110] Ibid 23
[111] Ibid 24
[112] Ibid 25
[113] Cabinet Office, Procurement Policy Note: Improving Transparency of AI use in Procurement (Information Note 02/24 2024)

importance, was to clearly set out that there is no prohibition on the use of AI by suppliers in the drafting of tenders. Yet, whilst not prohibited, the PPN does provide in Annex B a standard disclosure questionnaire that would require tenderers to outline if and where they had used AI in both the submitted tender and in the products/services being offered. The note also outlines that, in instances where a disclosure of the use of AI was made, additional due diligence checks may need to be applied to these tenders to ensure their accuracy, and that this time should be incorporated into procurement plans.

Of interest to this research is that, given the potential for generative AI to improve the productivity of the procurement process, and the ambition set out in the National Strategy for AI for the public sector to lead the way in its adoption, the PPN does not directly focus on the topic of the use of AI across the procurement lifecycle, let alone in evaluation. It does briefly hint at this, stating the potential for AI to accelerate and support the decision-making process, but then says that '*It is essential to ensure that decisions are made with the support of AI systems, not a reliance upon them'*[114]. It goes on to say that this should be done in accordance with principles set out by the Government's Data Ethics Framework, a separate guide for the appropriate and responsible use of data.

Whilst this PPN did provide some useful clarity to support greater adoption of AI in supplier organisations, it is unfortunately lacking in the detail needed to provide confidence to contracting authorities for its use on the evaluation side, potentially even relegating their use to a support role only. However, given the non-statutory nature of this and the other guidance documents, and the ruling in *Good Law Project v SoS H&SC* there is the potential for contracting authorities to disregard the best practice outlined within them where there is a good reason for doing so.

To conclude this chapter, the comprehensive review of both the hard and soft law has identified a total 8 key evaluation themes that any generative AI evaluation system

[114] Ibid 3

would likely need to align with to ensure compliance. The next chapter attempts to bring together the previous analysis in this paper against each of these themes, to determine if and where the use of generative AI evaluations would likely be compliant, and identify potential mitigation strategies for limiting the legal risk.

Chapter 3: Generative AI's alignment to the Key Evaluation Themes

Evaluation Theme 1: Covered Procurement

As the obligations in the Act generally apply to covered procurements, it does not appear a controversial position to claim that the use of a generative AI system, with all their current flaws and limitations, would not be prohibited by the Act for non-covered procurements. Given that contracting authorities are not required to run any type of competition for such contracts, the use of an AI tool, even an imperfect one, would likely be acceptable. Even for the sub-category of regulated below-threshold contract procurements, where contracting authorities are required to set out the criteria for award, this does not require the publishing or following of the assessment methodology, the likely hook for the incorporation of information on the use of generative AI. This potentially has significant practical implications for procurement practioners, with the feasibility of using generative AI systems for the often long-tail of procurements that sits below the thresholds. Whilst it may still be preferable to keep a human in the loop for the final decision, in order to stop the negative impact that a hallucination could have, this would likely be to ensure optimal purchasing occurs in line with an organisation's internal governance procedures rather than compliance to the Act.

Pushing this hypothesis further, it would also appear likely that, where a covered procurement does not require a competitive tendering procedure (i.e. a direct award), a contracting authority may be able to use a generative AI tool for evaluation of submissions from suppliers. This position seems supported by the technical guidance on direct award which states that, where a direct award ground applies, '*contracting authority may want to consider undertaking some form of informal competition before*

awarding the contract[115]'. It is this author's belief that there is a strong case for use of such systems for the justification of protecting life, as exemplified by the recent experience of the Covid-19 pandemic. During this time, many contracting authorities were overwhelmed with offers to procure desperately needed products. This required significant manpower to sift through, and ultimately opened up authorities to perceptions of conflicts of interest and not acting with integrity. A generative AI evaluation system could be built in advance of the next pandemic, to methodically sift offers and award contracts in a way that was transparent and agreed outside of an emergency.

Finally, and potentially more legally ambiguous, would be the potential to use such for the evaluation of tenders from non-treaty state suppliers. For instance, where UK embassy staff within a non-treaty state use generative AI for the evaluation of tenders that were only of interest to suppliers in that local geography. However, given neither the Act nor Regulations specifically call this out as an option, and the difficulty in determining when a contract would not be of interest to a UK or treaty state supplier, it would likely require either a change to the legislation or inclusion in the National Procurement Policy Statement (NPPS) to provide additional clarity.

Key Evaluation Theme 2: Alignment to the Objectives

Next, if a contracting authority intended to use a generative AI evaluation tool for a covered procurement, then they would likely need to build a robust case that doing so was aligned to each of the objectives of the Act. As we will see below, some of these appear more supportive than others.

It is to the first objective of delivering value for money that a generative AI tool would likely be most aligned. This is due to the potentially significant productivity gains that could be achieved through its use, by dramatically reducing the time, effort and ultimately costs involved in running the evaluation process. In addition, the use of generative AI would likely score highly in any potential value for money review of the

[115] Government Commercial Function, Guidance: Direct Award (OGL 2024) 7

proposed assessment methodology as recommended in the guidance. There is also a potential second order benefit of its use through the likely increase in the number of competitive procurements that a given contract authority could run with the same level of resources, and in doing so reducing instances of extensions to existing contracts that often occur where authorities have run out of time to deliver a compliant procurement.

Next, for the objective of maximising public benefit, the use of a generative AI evaluator would appear to be neutral. This is presumed as maximising public benefit generally is achieved via the selection of specific questions in the award criteria (e.g. the inclusion of targeted social value questions), rather than in the process of the assessment methodology. However, given that Cabinet Office is due to provide an updated NPPS with the intention of reforming how social value is delivered, this position could change.

It is the objective of sharing information where the use of generative AI appears the most problematic. Whilst it is feasible for a contracting authority to articulate in its tender documents that it will be using generative AI for the evaluation, whether this would be sufficient to allow a supplier to understand the decisions is questionable. From the transparency on the prompt engineering within the systems, to whether one is truly able to explain the output of the generative AI, there are many aspects that would appear not to be aligned to this objective. This is a point that will be covered in more detail as part of the wider key evaluation theme of transparency.

Next, the objective of acting or being seen to act with integrity also appears to be problematic for a generative AI system, for reasons discussed in greater depth in key evaluation theme 4. In short, the use of an underlying model that was provided by a private sector supplier would potentially be susceptible to perceptions of bias and, in doing so, undermine integrity. However, it is this author’s view that it is not outside the realms of possibility to construct a system in a way that supports the objective of acting with integrity. Or at least, passing a threshold for integrity similar to procurements delivered by humans currently.

There is arguably a mixed position for the objective of treating suppliers the same. On the surface level, it is possible to envisage an approach that would meet this objective, where each submission was evaluated by the same generative AI evaluation tool, against the same award criteria and assessment methodology. As a process, this would appear to be compliant. However, as has been discussed previously, due to potential biases in the training data, it is not possible to give complete assurance that the model would not treat a supplier differently, for example where there has been a significant degree of historic bad publicity that is present in the training data. Whilst there are potential ways to mitigate this risk, such as anonymising the organisation names, as has been raised previously there may be practical difficulties in ensuring that this is effective. In addition, whether such mitigations would meet the threshold of '*all reasonable steps*'[116] to ensure there was no unfair advantage or disadvantage will be something for either the courts to determine, future guidance or legislation.

Next, the use of a generative AI evaluation system appears to align with the objective of having regard to the barriers that SMEs face and seeking to remove them. One major barrier to SME participation in public procurements is the elongated timelines and the associated costs, sometimes known as the 'valley of death[117]'. By speeding up the process, generative AI could significantly reduce this barrier, directly supporting the second leg of the objective.

Finally, whilst not directly under the objective section of the act, the requirement to have regard to the NPPS would also need to be aligned with. However, with the withdrawal of the previous statement, and the replacement not yet published, the author is unable to provide a case either for or against this at the current time.

Yet one cannot conclude this section on the objectives without considering the topic of 'have regard'. Given that only the objective of 'same treatment' must be followed by a

[116] Procurement Act 2023, s 12 (3)
[117] Clyde Frank and others, 'Surviving the "valley of death": A comparative analysis' (1996) The Journal of Technology Transfer 21

contracting authority, it appears that complete alignment of the use of generative AI to the other objectives is not mandated. So long as an authority has rationally considered its use against each of the objectives, put in place reasonable mitigations, and has an audit trail of this decision-making process, it appears at least feasible that the use of a generative AI evaluator would not be in breach the Act's obligations. However, to have greater confidence in this position, one would need to review the currently unpublished covered procurement objectives technical guidance.

Key Evaluation Theme 3: Transparency

It is on the theme of transparency where the use of a generative AI evaluation system appears most problematic, primarily due to the potential depth of information that would need to be provided in both the assessment methodology and the assessment summary. It is this author's opinion that simply providing the evaluation criteria and stating that a generative AI system would be used to evaluate this would not meet the threshold of being set out in full in Regulation 33(2)(i). This is due to the fundamental role in the decision-making process that both the choice of model and the underlying prompt architecture would have, potentially being akin to undisclosed sub-criteria. However, if contract authorities are required to disclose details of both the model and prompts, as well as the domain knowledge that the model is using, this would likely have negative practical implications. This level of detail would allow suppliers to replicate the contracting authority's evaluation tool, and in doing so help craft what could be the perfect answer. Not only would this create an unfair advantage to those companies that could afford to develop such tools, at least in the short-term whilst their costs are still high, but ultimately make it difficult to meaningfully differentiate between tender submissions, undermining the entire process of quality evaluation. Yet a failure to be transparent on the use of prompts runs the risk of materially disadvantaging a supplier, especially if individual contracting authority prompts were constructed in a way that potentially favoured a specific tenderer, or UK suppliers over treaty state suppliers.

However, there is a potential solution to this conundrum. First, the government could either reform the legislation or provide additional technical guidance providing clarity that, where generative AI evaluation tools were to be used, then the underlying prompt architecture would not need to be provided to tenderers. Then, to limit the potential risk that a contracting authority constructs prompts in a way that does not align with the obligation in the act, a central body, such as Cabinet Office, would develop and make available a single instance of the tool that all other contracting authorities would be required to use. This single instance could then have regular independent audits to ensure that it continued to be compliant, and suppliers did not gain unfair advantages or disadvantages. This could also support the perception of integrity, as tenderers would have confidence that results were not being skewed by individual contracting authorities.

Key Evaluation Theme 4: Equivalence

At first glance, the need to ensure the acceptance of equivalence outlined in the Act's technical specification requirements would appear neutral to the use of a generative AI evaluator given that the compliant drafting of the requirements and associated tender documents would need to have occurred prior to commencing the evaluation process. However, on closer inspection, the need to ensure that equivalent standards or brands are not disadvantaged would appear to be a continuous requirement throughout the evaluation decision-making process. There is therefore a potential risk, stemming from either the training data or domain knowledge that a generative AI evaluator consumed, that where certain brands or standards appear more frequently and favourably than others in the training data, the tools could give preferable scoring when identifying these in a submission. As previously mentioned, the black box nature of AI systems would mean that it would not be possible to ascertain whether this was occurring. This risk could potentially be mitigated by building specific underlying prompts that instruct it to ensure equivalent responses are not disadvantaged within its scoring, yet whether this would be sufficient is unclear.

However, a generative AI solution could potentially assist in determining whether an unknown foreign standard was equivalent to a target UK standard. This is due to the ability of these tools to provide a detailed comparison of the tenderer's submitted standard to the target one, as well as their ability to accurately translate text. This could either provide a more detailed justification for an AI generated score or assist human evaluators in determining whether the submitted standard was a '*true equivalent*'[118].

Key Evaluation Theme 5: Conflicts of Interest

It is this author's belief that whether the use of a generative AI evaluation tool was aligned to the conflict of interest requirements has a greater dependency on practical impediments than theoretical ones. Given that the leading generative AI models are currently produced by the private sector, there is the potential that any evaluation tool built on top of one of these could be 'conflicted' either intentionally or unintentionally by their makers, to favour responses from either their own company, or connected companies. This represents a significant risk to the integrity of the evaluation process, and one that could not be easily identified from the outside. This would likely result in a position where the perceived conflict of interest could not be mitigated, forcing the authority to either exclude the supplier or not use the generative AI evaluation tool to evaluate tenders from them. This potential conflict would also extend to whichever organisation helped to design the evaluation tool that sat on top of the underlying model if this was outsourced to a third party with the necessary skills to develop it.

Yet there are potential practical solution to mitigate this risk. One would be for the UK government to develop their own sovereign model. Whilst it may be prohibitively expensive to reach the capabilities of today's frontier models, the costs of doing so are likely to reduce over time. This sovereign model could either be created within the public sector, or commissioned from a third party that would then be excluded from future procurements. This would likely limit the interest of commercial suppliers in building

[118] Government Commercial Function, Guidance: Technical Specifications (OGL 2024) 6

such a tool, with an academic institution being a more natural candidate for its development. An alternative approach would be to use multiple different providers' models for evaluations, much in the same way that multiple human evaluators are required to evaluate a tender. This way, potential conflicts from one system could be offset by the others, with a single score taken as an aggregate of all the AI systems used. Whilst not perfect, this would at least appear be a reasonable step.

Finally, the biggest ambiguity in relation to conflict of interest is that the drafting of the Act's obligations is entirely focused on a person, rather than a system conflict. Were generative AI systems to become widely used, then the legislation would likely need to be reformed to cover this separate category of system conflict risks.

Key Evaluation Theme 6: Keeping of Records that are 'sufficient to explain'

In many ways, the use of generative AI could make it significantly easier for a contracting authority to meet its record-keeping obligations. This is due to the plethora of tools available to both record and transcribe conversations that occur during the life of a procurement project. This is combined with its ability to understand the context of a question and identify relevant documentation from unstructured data sets, providing unparallelled productivity gains to the due diligence process. However, there is a potential downside to such a 'big brother' approach, in that where procurement agents know everything they say is being recorded, then they are likely to significantly reduce their candour and risk-taking behaviour, undermining the discretion and achievement of value for money which is at the core of the UK's reforms.

Yet, on closer inspection, it is this author's belief that again the inherent black box nature of generative AI systems would make any record of decisions generated by such systems unlikely to cross the threshold of 'sufficient to explain'. Whilst it is technically possible to ask such an evaluation tool to provide a detailed output of 'why' it gave a certain score, highlighting parts of the tenderer's submission so that it was recognisable from the information provided, this would in practice be an eloquent probabilistic output

rather than a reasoned decision. This lack of reasoning is a potential Achilles' heel for any contracting authority seeking to defend its decision-making process as a rational one from an unsuccessful tenderer. Given the novelty of this issue, it is unclear how the case law would develop on this. To resolve this ambiguity, it would be helpful to either make reforms to the legislation or include in a future guidance on AI evaluation the extent of information that would likely be sufficient. It is proposed that this would likely be at different thresholds whether this was based on a single instance of a sovereign model previously discussed, or whether the system used a commercially developed model, which would likely warrant a higher degree of disclosure.

Key Evaluation Theme 7: Remedies

It is the ability for an aggrieved supplier to seek remedies against a contracting authority, where they believe they have suffered damages due to the authority not meeting their obligations in the Act, that turns the analysis in this chapter from a purely theoretical exercise into one of real practical import. It is argued that the ability for the courts to order suspension of any decision made or force an authority to take any where a decision breaches its duties, would directly impact the ability to confidentially use a generative AI evaluation tool. On the surface level, their propensity to hallucinate, exhibit bias and the potential threat of prompt injections, would likely result in a structural increase in aggrieved suppliers. But even assuming that such issues could be resolved, or that humans in the loop catches them before award, whether the courts would accept the decision making of such a tool, with its unobservable reasoning due to the black box nature previously discussed, is in doubt. It would seem that a defence akin to 'computer says no' would not be a convincing argument. Until such models are able to show clear capabilities for reasoning, a feat that many companies are working towards but may never be reached, this would likely remain one of the largest barriers to incorporating generative AI evaluation solutions.

Key Evaluation Theme 8: Moderation

It is in the final evaluation theme of moderation that we potentially find one of the most fertile grounds for the use of a generative AI system. As previously mentioned, this vital step in the evaluation process is completely absent from the legislation. Yet both in challenging non-compliant evaluator reports and facilitating the reaching a single consensus score, the moderator's role is paramount in maintaining the integrity of the evaluation process and protecting the contracting authority from challenge. For this role to be effective, the individual moderator needs both a detailed knowledge of the legislation and the soft skills of people management and meeting facilitation. Dependent on the complexity of the procurement process, this could be organising and cajoling tens or even hundreds of individual evaluators into a consensus position. Given this role ultimately does not require the making of any direct decisions about a tenderer's submission by the moderator, it is a stage in the evaluation process that could integrate generative AI in a way that potentially represented a low legal risk. One could envisage a generative AI moderator chat bot, trained on detailed domain knowledge of the legislation, guidance notes, and the specific procurement documentation, used as a tool to support evaluators in understanding whether their initial scorings and justifications were compliant with the obligations in the Act (i.e. highlighting where an evaluator has marked down a tenderer for not using a specific brand), or where their reasoning for a score did not directly relate to the published award criteria. The evaluators would then be able to review their initial response and assess whether they would like to change it. Generative AI tools could also be used to summarise the key differences in scores between evaluators, and guide consensus scoring activities between them until a final score was reached. A potential strength of such a use case is that it would not need the generative AI systems to always be accurate, as it still would have the human evaluators in the loop, should the systems hallucinate.

To conclude, on reviewing the use of generative AI against each of the key evaluation themes, it is this author's opinion that the strongest use cases for a generative AI evaluation system would be for non-covered procurements, direct award, and as a

moderator. These appear to provide the lowest legal risk and are worthy of both further analysis and practical experimentation and implementation.

Conclusion

This paper has attempted to bring together two very different fields of knowledge, in computer science and public procurement law, to critically assess whether generative AI systems could be incorporated in a compliant manner within the evaluation of quality submissions under the UK's public procurement legislation. It is the author's view that this is a question of profound importance, as generative AI becomes more powerful and embedded in the everyday lives of citizens. To fail to integrate such systems would not only be a disservice to the public procurement profession, but also risk further elongating the length of public procurements as increasing numbers of tenders are received following wider adoption of the technology by suppliers. It is proposed therefore that this should not be a question of whether generative AI evaluation should be used, but one of how it can be implemented in the most legally compliant manner. This paper argues that even with the known limitations, there are a number of immediate use cases, such as for non-covered procurement, direct award for urgency and the process of moderation, where the legal risk of its use would appear low and the potential productivity benefits high. Whilst it would be a brave contracting authority that sought to use such a tool for the evaluation of a covered procurement, due to the potential avenues for challenging the outcome. But that doesn't mean it shouldn't and couldn't be tried. In any piloting of such an approach, mitigation steps could reduce the legal risk: from detailing the use clearly in the tender notice; using a sovereign model or; having centrally audited prompts that could draw on the domain knowledge of the legislation, technical guidance, and that specifically sought to ensure equivalence.

Yet, given the lack of legal precedence covering such an approach to public procurement evaluation, and the significant ambiguity as to the level of detail required to both explain the assessment methodology and the contracting authority's decision, it would be highly likely that an unsuccessful tenderer would be successful in seeking a judicial review

against its use. It is therefore proposed that, alongside a series of targeted reforms and guidance outlining both the non-prohibition of generative AI evaluation and the level of detail required to be published where it was used, it would be of wider public benefit for a single contracting authority, potentially cabinet office, to pilot its usage and be resourced appropriately to defend a challenge. Doing so would ultimately provide other contracting authorities with greater clarity as to what would, and would not, be deemed compliant.

Finally, this author believes it is worthwhile to conclude with a more philosophical consideration, in that many of the limitations of generative AI systems outlined in this paper mirror those found in human agents that are permitted to evaluate tenders every day. Whether that be the presence of bias, conflicts of interest, general inaccuracies and miss-remembering (akin to hallucination), these are all well-known and understood human traits. Furthermore, our lack of understanding of the black box that is generative AI is arguably equal to our lack of understanding of human intelligence. In this author's opinion, many in society appear to want to hold artificial intelligence to a higher standard than we hold our fellow humans. This would appear an illogical position given the significant benefits that integrating these systems would bring to all parties. Having sat through many public procurement evaluation procedures in their career, there is no doubt in this author's mind that the current generation of AI tools would be able to perform as well as an evaluator, if not better, than the average human evaluator, author included. And, assuming the rapid developments in this field continue, such generative AI evaluation tools are likely to become more powerful and accurate. Given anyone with a passion for public procurement can see this new world is approaching, and recognising the length of time it typically takes to develop new comprehensive policies and legislative reform, it is not acceptable to wait until these systems are 'perfect' before we start considering how to integrate them into the evaluation process. That work should begin now.

Primary Sources

UK Cases

Good Law Project Limited v Secretary of State for Health and Social Care [2021] EWHC 346 (Admin) [2021] ACD 49

UK Legislation

Procurement Act 2023

Procurement Regulation 2024

EU Legislation

Artificial Intelligence Act [2024] OJ L 1689

Bibliography

Afifi-Sabet K, 'Most formidable supercomputer ever is warming up for ChatGPT 5 — thousands of 'old' AMD GPU accelerators crunched 1-trillion parameter models' (Tech Radar 12 January 2024) < https://www.techradar.com/pro/most-formidable-supercomputer-ever-is-warming-up-for-chatgpt-5-thousands-of-old-amd-gpu-accelerators-crunched-1-trillion-parameter-models> accessed 29 November 2024

Alke J and Hassel J, 'Aspects of knowledge management applied to public bid writing: Lower the barriers of entry in public procurement by streamlining the bid writing process'(KTH Royal Institute of Technology 2023).

AutogenAI 'Backed by Salesforce Ventures, AI Company AutogenAI Raises $39.5m series B' (2003) <https://autogenai.com/articles/salesforce-ventures-co-lead-a-39-5m-investment-round-in-autogenais-game-changing-proposal-writing-software/ accessed 12 Feb 2024

Cabinet Office, Green Paper: Transforming Public Procurement, CP556 (The Stationery Office 2020)

Cabinet Office, Procurement Policy Note: Improving Transparency of AI use in Procurement (Information Note 02/24 2024)

Commission on Race and Ethnic Disparities, Commission on Race and Ethnic Disparities: The Report (CCS0221976844-001, 2021)

Crevier D, AI: the tumultuous history of the search for artificial intelligence (Basic Books 1993)

Department for Science, Innovation & Technology, A pro-innovation approach to AI regulations: government response (Command Paper 1019, 2024)

Frank C, Sink C, Mynatt L, Rogers R and Rappazzo A, 'Surviving the "valley of death": A comparative analysis' (1996) The Journal of Technology Transfer 21

French R M, 'The Turing Test: the first 50 years' (2000) Trends in cognitive sciences 4.3

Gao A, 'Prompt engineering for large language models' (2023) SSRN 4504303

Giray L, 'Prompt engineering with ChatGPT: a guide for academic writers' (2023) Annals of biomedical engineering, 51(12)

Government Commercial Function, Bid Evaluation Guidance Note, (OGL 2021)

Government Commercial Function, Guidance: Assessing Competitive Tenders (OGL 2024)

Government Commercial Function, Guidance: Assessment Summaries (OGL 2024)

Government Commercial Function, Guidance: Direct Award (OGL 2024)

Government Commercial Function, Guidance: Technical Specifications (OGL 2024)

Hacker P, Mittelstadt B, Borgesius F Z, and Wachter S, 'Generative discrimination: What happens when generative ai exhibits bias, and what can be done about it' (2024) arXiv preprint 2407.10329

Hebb D O, 'The Organization of Behaviour' (Wiley New York 1949)

Katz D M and others, P 'GPT-4 Passes the Bar Exam' (2023) SSRN https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4389233 accessed 12 Jan 2024

Kozma R, Ilin R, and Siegelmann H T, 'Evolution of abstraction across layers in deep learning neural networks' (2018) Procedia computer science 144

Krizhevsky A, Sutskever I and Hinton G E, 'Imagenet classification with deep convolutional neural networks' (2012) Advances in neural information processing systems 25

Leahy S and Mishra P, 'TPACK and the Cambrian explosion of AI' (2023) Proceedings of Society for Information Technology & Teacher Education International Conference

Maleki N, Padmanabhan B, and Dutta K, 'AI hallucinations: a misnomer worth clarifying' (2024) IEEE Conference on Artificial Intelligence

McClanahan C, 'History and evolution of gpu architecture' (2010) A Survey Paper

McCulloch W S and Pitts W, 'A logical calculus of the ideas immanent in nervous activity' (1943) Bulletin of Mathematical Biophysics 5

Mei Q, Xie Y, Yuan W and Jackson M O, 'A Turing test of whether AI chatbots are behaviorally similar to humans' (2024) Proceedings of the National Academy of Sciences, 121(9)

Minsky M and Papert S. *Perceptrons* (MIT Press 1969)

OWASP 'OWASP Top 10 for LLM applications 2025' (2024) <https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/> accessed 29 November 2024

Rannard G and Fraser G, ''Godfather of AI' shares Nobel Physics Prize' (BBC, 8 October 2024) <https://www.bbc.co.uk/news/articles/c62r02z75jyo#:~:text=The%20Nobel%20Prize%20in%20Physics,and%20said%20he%20was%20flabbergasted> accessed 29 November 2024

Ribary, M, Krause P, Orban M, Vaccari E and Wood T, 'Prompt Engineering and Provision of Context in Domain Specific Use of GPT' in Legal Knowledge and Information Systems (IOS Press 2023)

Rosenblatt F, 'The perceptron: A probabilistic model for information storage and organization in the brain' (1958) 65(6) Psychological Review

Rumelhart, D E, Hinton G E and Williams RJ, 'Learning representations by back-propagating errors' (1986) Nature 323.6088

Sanchez-Graells A, 'Public Procurement of Artificial Intelligence: Recent Developments and Remaining Challenges in EU Law' (2024) Legal Tech Journal 2/2024

The Office for Artificial Intelligence, National AI Strategy (Command Paper 525, 2021)

Tredinnick L and Laybats C, 'Black-box creativity and generative artificial intelligence' (2023) Business Information Review 40(3)

Turing A M, 'Computing Machinery and Intelligence' (1950) Mind 59

Turing A M, 'On computable numbers, with an application to the Entscheidungsproblem' (1936) Proceedings of the London Mathematical Society Series 42

Vaswani A, Hazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L and Polosukhin I, 'Attention is all you need' (2017) Advances in Neural Information Processing Systems

Wallace D, Yukins CR and Matechak JP, 'UNCITRAL Model Law: Reforming Electronic Procurement, Reverse Auctions, and Framework Contracts' (2005) 40 Procurement Law

Waseem M and others, ‘Artificial Intelligence Procurement Assistant: Enhancing Bid Evaluation’ (2023) International Conference on Software Business

Weizenbaum, J, "ELIZA—a computer program for the study of natural language communication between man and machine." (1966) Communications of the ACM 9.1

Williams S and Huckle J, ‘Easy Problems That LLMs Get Wrong’ (2024) arXiv preprint 2405.19616.