Digital Democracy - Large Language Models for Transparency & Accountability

"The real problem of humanity is the following: We have Palaeolithic emotions, mediaeval institutions and godlike technology."

— Edward O. Wilson

Introduction

Abstract

Using an experimental methodology, expressive visualisations are able to capture key information about representatives in the UK House of Commons. Using divisions to pin politicians' political opinions, concerning trends of polarisation can be identified and clearly measured producing an indicator from which to understand UK politics.

Large Language Models (LLMs) are doing for Natural Language Processing (NLP), what Neural Networks (NNs) have done for Image processing. Using GPT4, years of work was compressed into a single project. This dissertation will demonstrate the capabilities, and limitations of LLMs for tasks such as classification, and clustering compared to traditional NLP techniques.

Key Terms: Large Language Models (LLMs), Natural Language Processing (NLP), Neural Networks (NNs), Division, Bill, Amendment, Motion, Prime Minister (PM), Member of Parliament (MP)

Intro

If we don't know how our local representatives vote, how can we hold them to account when they don't act in our interest?

It is the proposal of this dissertation that an information asymmetry exists between constituents and representatives, which prevents the proper dynamics between these two groups. The existence of an information gap may be a social problem, but its solution falls in the domain of data science & visualisation.

This study will demonstrate how new techniques, such as prompt engineering with Large Language Models (LLMs), can outperform traditional Natural Language Processing (NLP), while maintaining a 20-hour runtime; for a task that would take a human expert months or years.

The results will also show clear evidence of polarisation in the UK political system between 2006 - 2023, along with other interesting artefacts showing changing voting behaviour in specific MPs. In accordance with digitally democratic values, limited time will be given to interpretation of the data itself, with the exception of a few cases to demonstrate findings, instead the data will be made publicly available on github for those to conduct their own research.

This dissertation bridges two fields, examining Government and Politics, through the lens of Data Science and Modelling. It will review democratic theory from a modelling perspective, compare traditional NLP approaches with LLMs, and finally present digitally democratic solutions for closing the information gap.

Fundamentally Digital Democracy is based on an assumption, that democracy is more than just ‘the worst form of governance, apart from all the other ones’ - Churchill (1947). Instead it claims that democracy is an effective method of governance where the goal is to use information technology to identify new, or improve existing, methods of aggregating disparate information across vast groups. This is the epistemological theory of democracy, and is based on the wisdom of crowds (Surowiecki, 2004. Sunstein, 2008).

Questioning the previous assumption, grounding digital democracy in previous work, and identifying an appropriate path forward, will be the focus of the literature review. Methodology and findings will focus on data science techniques and visualisations.

Problem & Objectives

There are two key observations that triggered my interest in Digital Democracy:

1. Generally, the quality of democratic societies scales negatively with participation (Simon, 2017)

Counter intuitively the more democratic a nation is deemed to be based on the Economist Intelligence Unit’s Democracy Index (2016), the lower the average participation in general elections. ‘The highest levels of disengagement have occurred in 16 out of the 20 countries classified as ‘full democracies’ (Simon, 2017)

2. Youth participation in general elections is especially poor in Western democracies (UK Government, 2015. Dezelan, 2023)

The common logic that ‘the most politically motivated members of society are in the youth’ appears to be refuted by actual voter participation statistics. The simple explanation given for this is that the right to vote is ‘taken for granted’, or that people are ‘not interested in politics’ (UK Government, 2015).

While this is likely a partial explanation, it is also rather convenient. To suggest apathy, reveals no system level solution, and in effect no problem to be solved. A more difficult justification for the two observations above is that generally people feel disengaged, and under-represented by the current system.

Aims

The purpose of this dissertation is to provide UK citizens with the information they need to make informed decisions about their representatives. It will also demonstrate how large language models (LLMs) can be used to leverage existing data to develop this information.

Objectives

The objectives of this dissertation are:

1. To investigate to what extent low participation rates are caused by opaque, outdated democratic systems. (Answer through Literature Review)

2. To Identify to what extent Digital Democracy can present solutions to this problem. (Answer through Literature Review & Findings)

3. To use publicly available resources and data science techniques to produce ‘stage one’ digitally democratic visualisations and other ‘thick’ transparent information (Edwards, 2015). (Answer through Methodology & Findings)

Motivation

Accepting the premise that democratic systems are, all else being equal, greater than non-democratic systems of governance. And that democracy is not a default final state which societies tend toward, but rather a product of an difficult 2500 year history of evolving democratic theory (Hilbert, 2009). Growing threats to democratic UK institutions should be taken seriously, and opportunities to design more robust processes should be explored.

Given the global context of escalating military spending (World Bank, 2021). The rapid growth in competitiveness of authoritarian governments such as China (World Bank, 2021). Deliberate attempts to meddle in elections from foreign powers such as Russia (Mueller, 2019. Brattberg, 2018). And the more recent trends of claiming ‘election fraud’ by even the most democratic nations’ candidates such as Donald Trump in the 2020 US election (Berlinski, 2023). There is much reason for concern in the fragility of fledgling, or even, more mature democratic nations.

Ray Dalio (2021) describes the cyclical pattern of global unrest in his book ‘The changing world order’. If authoritarian governments such as China are really going to pose a meaningful threat to US dominance in our lifetimes, we need our democratic institutions to be stronger than ever. This means identifying problems, and seeking opportunities to integrate new capabilities with our existing democratic systems.

We should actively choose the best systems of government rather than let our technological capabilities dictate our systems and processes, however this dissertation will avoid discussion advocating between systems of government. The literature already conflates the debate on representative, vs more direct governance processes with digital democracy concepts. Rather, the focus of this dissertation is to identify aspects of digital democracy such as, awareness, deliberation, and participation that could improve government processes while being agnostic to the actual system itself.

Structure

Literature Review

Much of the digital democracy literature is based in the humanities, where possible I will attempt to interpret theory through the lens of data science, and modelling. Initially some context in political theory will help to ground the work in relevant literature.

Definitions

Democracy

The content of any process which we call democratic, is found in its ability to create ‘just systems’ (Hilbert, 2007). The fairness of the procedure has no bearing on the outcome of a decision, a democratic process can produce results that are good or bad for the system itself.

Democratic theory dictates that across the set of possible actions, given those actions impact a wide and diverse set of people, ‘just systems’ will in the long-run produce more equitable results than other systems of government. This same theory states that in our current representative system of government, transparency ultimately keeps politicians inline with moral and ethical considerations of their constituents (Heemsbergen, 2021. Hilbert, 2007).
As previously stated, it is the proposal of this dissertation that an ‘information gap’ exists between constituents and representatives. It is this dynamic, where constituents act to check and balance politicians conduct, that is impeded by information asymmetry. The nature in which we choose to close this gap will determine whether representation as a system of government is simply a product of technical limitations, or if more direct systems of government will begin to emerge.

At its core, democracy can be considered a process of information communication, with a focus on ideas and deliberation (Simon, 2017. Hilbert, 2007. Hilbert, 2009) . It is no wonder then, that the rapid advancement in Information Communication Technology (ICT) in the last 30 years, was assumed to transform our democratic processes enormously. Those same experts today echo each other's frustrations as technology continues to advance, while systems stagnate in comparison (Anttiroiko, 2003. Mahrer H, 2005. Bittle, 2009).

Digital Democracy

Ideas that involve democratic processes leveraging 21st century ICT fall under the banner ‘Digital Democracy’. Definitions have evolved several times, shifting typology as more nuanced understanding began to emerge. These definitions matured with the technology, such that even ‘recent’ literature in the early 2000s, appears outdated in a modern context (Kirschner, 2003. Bittle, 2009). This dissertation will focus on Tsagarousianou’s (1998) definition of Digital Democracy, aka ‘e-democracy’. This is a subset of the earlier ‘e-Governance’ concept, except it excludes ‘internal democratic processes’ which can be characterised by ‘digitalisation’, now a commonly understood process in the private sector (Lee, 2011. Pautz, 2009).
Digital Democracy in Tsagarousianou’s terms has 3 core traits which all Digital Democracy concepts can be categorised under:

1. Information Provision (Awareness)
2. Deliberation
3. Participation (Engagement)

In this case I have taken a further step, to number the categories into 3 distinct stages. This progression is hinted at in the literature, but not explicit. While projects in the past have by no means followed this chronological order, there is a logic in their relative importance. Deliberation without awareness leads to debate and division (Hilbert, 2007. Simon, 2017), while participation without awareness or deliberation is ineffective and inefficient (UK Government, 2015).
In some special cases, such as capturing 'The Wisdom of Crowds', it can be shown that engagement should precede deliberation, where conditions of independence are required for knowledge creation (Surowiecki, 2004), but this can be considered a special circumstance.

Historical context

Since Democracy was first established in Athens, the system of governance defined by 'power' and 'people' has changed dramatically.
Originally characterised by a great deal of power, in the hands of a narrowly defined 'people' (women, the poor, and slaves didn't have a right to vote). Modern democracy in contrast, distinguishes itself by extremely limited power, with almost universal right to vote among the people (Chapman, 2023).

Even today the majority of government pro-democratic policy is focussed on expanding the dimension of people through equality, rather than the dynamic of power as democracy is defined (Uberoi, 2022). This was assumed a necessity until the end of the 19th century due to the size of modern nations.

In theory, this is known as the trade off between the masses, and expression of will (Hilbert, 2007, Hilbert, 2009). If a community of 10 people have 3 minutes to explain their point of view, expression of will might take half an hour. A population of 10 million, would take 57 years to express their will, in a similar process. An unacceptable timeframe considering 8.98 million citizens in London alone.

This is where the concept of representation derives its necessity / legitimacy in democratic theory. A cynical postmodern perspective would point out that representation was less a conscious compromise of circumstance, and more like a justification given to legitimise the bourgeoisie perspective of an elite ruling class (Vedel, 2006).

A realist perspective would point to theory to justify representation, that the UKs free mandate, protects minority groups from the 'tyranny of the majority' (Barber, 2003). And the pragmatic idea that any movement to improve government has a greater chance of success by building on the existing system, rather than risk being opposed by those already in power. The 'tyranny of the majority' occurs when majority rule is used to legitimise inequitable actions, while the latter point is best described by H, Mahrer (2005) in the middle man paradox, where the real decision makers in a system are resistant to change that diminishes their own power.

Digital Democracy

Digital Democracy is presented in the literature as if it's the salvation to any given critique or shortcoming identified in democratic theory. Because of the blanket use cases, the scope of its application ranges from electronic voting, to direct democracy, as well as digitalisation of internal government processes (Lee, 2011. Weiss, 2022. Pautz, 2009. Spakovsky 2015).

The literature on digital democracy is split into a series of epochs, roughly starting in the 1960s until the start of the 1990s, typically using the prefix ‘cyber’, the era is characterised by extreme optimism and grand visions for future societies (Vedel, 2006. Tsagarousianou, 1998).
From the end of the dot-com bubble 2000 - 2010 the literature shifts typology describing projects as being ‘e-democratic’ short for electronic democracy (Lee, 2011. Hilbert, 2009. Anttiroiko, 2003. Şendağ, 2010). The sentiment has shifted at this time, recognising that the technology itself is just a tool, and not inherently democratic (Hilbert, 2007). This had been recognised as a potential threat in some of the prior literature (Anttiroiko, 2003). It’s marked by an underlying disappointment as the technology, while advancing rapidly, is still limited by early web 2.0’s lack of user friendliness, and general inability to facilitate desired projects (Vedel, 2006. Mahrer, 2005. Simon, 2017).
Finally from 2010 onwards, really gaining momentum in the UK around 2015 with parliaments ‘Open-up’ initiative, we begin to see the infrastructure and groundwork laid for meaningful decentralised, bottom-up initiatives to take off (UK Government, 2015).

The key message from this current epoch is that of caution. A lot of the groundwork has been painstakingly put in place in the past decade. This removes the limitations of the early 2000s but policy from government and NGOs alike advocate for slow, long-term thinking to be the backbone of fledgling initiatives (The Democratic society, 2023. UK Government, 2015).
Many barriers in the UK, such as the digital divide persist, but are being diminished (UK Government, 2015. Edwards, 2015). Successful social media platforms such as reddit, and collaborative websites such as wikipedia, have inadvertently progressed the trial and error work to design co-operative online strategies. These strategies will be necessary to implement deliberative processes first articulated in the early literature.
By no means is the path clear for ‘innovators & entrepreneurs’ (UK Government, 2015), but the environment is ripe, the technology matured, and low hanging fruit has presented opportunities for straightforward progress.
The UK may have been criticised early on for the lack of initial progress in the space compared to peer nations. However a stream of steady progress has not only begun to turn those tides, but numerous case studies show, there may have been justification for initial hesitancy (Pautz, 2009. Huetlin, 2016). The UK’s policy on digital democratic initiatives is to remain agnostic, providing thin transparent material from which the private sector and tech innovators can build. This strategy is slower than creating projects internally, but by remaining agnostic to any particular party it also maintains the democratic integrity of future projects.

Data Science & Democracy

The wisdom of crowds can be directly understood through the lens of data science and modelling. This section will address the wisdom of crowds to demonstrate the connection between the two fields, and better understand the purpose of digital democracy.

To summarise papers primarily based in the humanities, and focus the literature review on the problem of ‘low participation’ / ‘disengagement’, I am going to attempt to articulate the modern understanding of an individual’s ‘Political Will’ via a simplified model. This should be a useful crutch for understanding the problems we face with representation, rather than useful in practice, ‘All models are wrong, but some are useful’ - George Box (Barroso, 2019).

Wisdom of Crowds

The epistemological theory of democracy, behind all types of democracy, deliberative, direct, and representative, is based on the ‘wisdom of crowds’ (Surowiecki, 2004). Its application to democracy strikes at the question - what is democracy for? Is it simply the least worst form of government, or is it an effective system for making uncertain decisions?

The wisdom of crowds occurs when the average estimate from a group is a better prediction than any single expert. This works for qualitative and quantitative decision making, such as brainstorming potential solutions to a problem, or estimating the number of jelly beans in a jar, but has limits when laymen groups are asked about specialist subjects. It is also shown to be subject to bias when estimating probabilities which results from inherent human biases (Sunstein, 2008. Wolfers, 2004).

In subjects of ethics, and complex problems where significant trade-offs may exist, the definition of an ‘expert’ is suspect (Sunstein, 2008. Rittel, 1973). Therefore vast realms of politics may benefit from stronger democratic participation. It is important to identify when and where greater participation may or may not be effective. Conditions of independence, diversity, and decentralisation must hold for the technique to be effective, and many poorly executed case studies can be found (Sunstein, 2008. Surowiecki, 2004).

The model for this technique is based on a simple equation, assuming previous conditions are met and that there is actual information in the system (crowds cannot predict purely random events). An individual's prediction is based on their given Information (I) and some Error term (e). Error can be assumed to be normally distributed in a diverse group, while aggregate Information will shift the mean of this distribution toward a centre of collected knowledge. Using a preferred system of aggregation, the expectation of the crowd takes the form:

$$\mathbb{E}[\bar{I}] = \frac{1}{\infty} \sum_{i=1}^{\infty} [(I + e)_i]$$

Where: $\bar{I}$ = Mean Information, I = Individual Information, e = Error-term (normally distributed). $(I + e)_i$ = An Individuals Best Prediction, $E[\bar{I}]$ = The True Value

Page (2017) elaborated on this by introducing the diversity prediction theorem. Page demonstrates via an identity that even in the absence of experts, diversity will improve accuracy.

It makes intuitive sense that adding ever more brilliant experts to a group, will cause the average estimate to increase in accuracy. But so will diversity, measured as the difference between individual predictions, and the crowd's prediction based on its collective knowledge. The crowd's collective error is the difference between the group's collective knowledge and the truth.

$$(\bar{I} - \theta) = \frac{1}{n}\sum_{i=1}^{n}[(I+e)_i - \theta] - \frac{1}{n}\sum_{i=1}^{n}[(I+e)_i - \bar{I}]$$

Where: $\bar{I}$ = Mean Information, θ = The True Value, I = Individual Information, e = Error-term. $(\bar{I} - \theta)$ = Crowds Collective Error, $(I+e)_i$ = An Individuals Best Prediction, $(I+e)_i - \theta$ = Average Error , $(I+e)_i - \bar{I}$ = Diversity

There is a problem with applying this theory to democratic process. We have no definitive metric that can be used to verify decisions are in the public interest. So while we can check how many jellybeans are in the jar, we cannot check the error in alternative policy decisions. Finding methods to overcome this problem and effectively aggregate information across disparate groups is at the core of digital democracy.
This is the role of deliberation in parliament, where MPs are accurate because of their expertise. It is not clear however, that our MPs are paragons of independent, or diverse thought which is assumed for the process to work effectively. This is concerning in the context of ethical debates where the idea of an 'expert' is suspect.

Working model of 'Individual Political Will'
Many studies cite a 'dissatisfaction with the current system' as a cause of low participation in the UK, although this is vague and poorly articulated. In this section, I will attempt to exemplify a problem with the current system through a crude model based on political theory. The suggestion being that representatives should play a more meaningful role when an individual is considering to vote.

Each member of society is at once:

> A citizen who owes their allegiance to the prosperity of the Nation (Democratic Theory)
>
> Identifies as a member of a group with collective interests (Identity Politics)
>
> An individual with their own Self-interests (Liberal Theory)
>
> And traditionally, associated with a political Party. This fourth dimension results from the previous three and the system of government which they are expressed (Party Politics)

A persons 'political will' can be understood to result from a combination of National (N), Identity (I), Self-interest (S), and Party (P). In a general election, we ask every individual with a complex position resulting from the above dynamics to select between Representatives (R).
We are in effect taking a selection of inputs, N, I, S, P, and being asked to select an output in the form of R.

In the absence of adequate information about R, the moderate correlation with P is used as a heuristic, voting for P irrespective of the actual representative or their prior performance.

At first the problem might be assumed to stem from a ‘lack of expression’,where mirriad complex, overlapping, and often shifting Identities (I), combined with the other dimensions (S,N,P), are forced to be expressed by selecting one of 3 or 4 leading parties.

This would appear to be exaggerated in a referendum, where this same complexity is narrowed into a binary classification; For or Against.
However the Brexit referendum sparked mass participation, exceeding that of any recent general election (Uberoi, 2022). This seems to indicate that the voting ‘format issue’ of over simplified classification is not the primary issue, despite so much of the literature using ‘limited expression of will’ as a base to argue for a deliberative democratic solution (Benhabib, 1998. Hilbert, 2009. Barber, 2003. Wolfers, 2004).
Instead it suggests, part of the problem ‘dissatisfaction with the current system’, is created when instead of a direct connection between the question being asked and the resulting action being taken; as in the EU referendum.
General elections ask us to vote on R as a proxy for P. Where our participation seems twice removed from the issues we actually care about; from the issue to the party, from the party to the representative. When we have no information on the representative, they have no relation to a person's political will, it is no wonder that people feel a disconnect or unrepresented by the current system.

It is like going to a restaurant, but instead of being given a menu, you are being asked to select the chef you wish to prepare your meal based on his nationality, and assuming that this represents his preferred cuisine. At least in a referendum you are being asked if you would like a specific dish, even if they neglect to inform you of other options.

Further problems with the representative system in the UK are highlighted by parliament, for example the cost of candidacy presents significant barriers for the working class to run (Uberoi, 2022). As such representatives are no more selected by the people, than delegated by the national party. Fostering a political class of financial elite; this compounds the lack of connection with candidates, and reduces the diversity of thought in parliament.

It is the suggestion of this dissertation that if we can inform the public about the representatives from which they are asked to cast their vote, constituents would be able to hold their representatives to account, affecting political change more effectively. By introducing the information required to judge representatives, R then becomes an input factor to consider and express directly rather than a proxy for P. ‘From the issues we care about, directly to the representatives we have chosen’, and that this may in some cases, and to a limited degree decrease the feeling of 'dissatisfaction with the current system’, resulting in improved participation.

The ‘political will’ for an individual is then the product of:

$$CR = N * I * S * P * R$$

Where: CR = Chosen Representative, N = National Interest, I = Identity Groups, S = Self-Interest, P = Party Affiliation, R = Representative

This does not reduce the financial barriers to selecting candidates, nor does it solve the ‘compression of will’ into restricting classifications. But if it makes representatives more vulnerable to bad decisions, then it also creates the opportunity for new candidates, while deterring entrenchment. This competition, Hilbert (2007) suggests, is key to a healthy democracy.

UK Democracy

Participation among the general public has declined then plateaued since the 1960s, where post war Britain recorded 80% turnout; unprecedented in mature european democracy in recent decades (House of Commons , 2004. UK Government, 2015). The ‘galvanising spirit of war’ clearly stimulates a strong national sensibility. However, Europe after World War 2, is twice shy having been burnt by the darker side of national interest, whose perverse instantiation leads to nationalism or Fascism (Barber, 2003). As such the UK tends to rest heavily on the liberal ethics of self-interest, and individual human rights.

The two problems at the crux of this dissertation highlight the disillusionment especially of youth in western democracies, despite growing engagement in alternative political activity such as protests, online petitions, and community action (House of Commons , 2003. Uberoi, 2022). If apathy cannot explain low participation in youth, then disengagement must stem from a different source.

'They are apathetic because they are powerless, not powerless because they are apathetic' - Barber (2003)

UK Digital Democracy

Case Studies
A few successful UK initiatives are worth highlighting.

Assessing their success will be done with a simple methodology. Rating each initiative between 0-5 based on how:

> *Democratic* - Is the initiative built on principles of a just system? Specifically looking at fairness, accessibility, and bias.
>
> *Digitally Democratic* - Which, and how well does the initiative achieve its given stage as Digital Democracy is defined by Tsagarousianou? Information provision, deliberation, or engagement. Does this initiative empower people? Specifically with reference to both dimensions of power & people.

Hansard

Resulting from the UK's transparency initiatives such as 'Open up' (UK Government, 2015), which was a commision led by then speaker of the house John Bercow and the Speaker's Commission on Digital Democracy, primarily focused on the house of commons.

The Online Hansard, is part of a collection of sources published at this time, including MP votes, and other raw data, that later included the House of Lords.

Deemed 'thin transparency' by Edwards (2015), 'thin' and 'thick' differentiates raw data, from visualisation or analysis which has been left to entrepreneurs and innovators to prepare for consumption by the general public. Thin transparency is a deliberate attempt by the UK parliament to remove itself from the position of judge.

This is an example of a top down initiative required as the backbone for future projects, on which many of the latter case studies came to rely. This dissertation can be included in this set of projects.

While generally bottom up is to be desired, this first step can be considered a minimum requirement for transparent governance and digital democracy projects. This is an example of type 1 digital democracy - 'Information Provision'.

Democratic - Score: 4/5

Although it has some limitations in available data it has taken remarkable steps to remain agnostic and forward planning. The data is accessible to anyone although it can be difficult to find specific information even for the technically gifted, as such some points are lost for the 'accessibility' dimension of this criteria.

Digitally Democratic - Type: 'Information Provision & Awareness' - Score: 2/5

While being the backbone of future Digital Democratic initiatives, the data has been notoriously difficult to navigate, and deliberately takes the form of raw data. This does little to empower the public with meaningful information, despite being a gold standard in transparent governance. The hansard also uses pseudo random page URLs which prevents meaningful data extraction, a major barrier to data collection, which the API does not fix at the time of writing.

They work for you

They work for you is a premier UK project, designed to address directly the information gap articulated in this dissertation. It converts data from the Hansard, into simple consumable bites of information. The data tracks the voting behaviour of elected MPs, it does this through a text based methodology.

By categorising legislation and tracking the voting behaviour, simple outputs such as 'always votes for POLICY X', 'inconsistently votes for Policy X', or 'Consistently votes against POLICY X' are provided. The website has steadily grown in popularity for good reason.

Democratic - Score: 4/5

A phenomenal job at addressing a shortcoming in UK democracy.

Slight deductions are made because of data analysis methods, although this may be an unavoidable consequence when simplifying big data. Issues are determined from the top down, imposing the will of the organisation on issues deemed 'relevant'.

Some website accessibility issues have been documented, 40% of users stating they could not find what one is looking for (Edwards, 2015).

Digitally Democratic - Type: 'Information Provision & Awareness' - Score: 4/5

The site, while growing in popularity, is by no means universally adopted. Many don't know about the project that would otherwise use it. As a third party website, no simple way of sharing graphics limits the empowerment any one person can feel.

Contact them

A simple initiative with clearly stated aims. The website helps integrate antiquated governance systems with modern technology. Specifically it addresses the problem of low response rate from local MP's, identifying missdirected emails as a potential cause, which leaves action in incorrect, or unmonitored inboxes.

The website helps to correctly target relevant MPs using a simple tree algorithm to identify the correct MP, & targets accountability by monitoring final response rates.

It also takes measures to prevent cyber attacks with verification of user emails, which is later used to monitor response rate.

Democratic - Score: 4.5/5

The website remains agnostic, integrating efficiently into the existing system.

Bias in response rate may be created by its collection method, because response rate data collection is not mandatory, and negative feedback is generally a greater motivator for humans. This means assumptions must be made in cases where the data is not returned, which may cause inaccuracies despite being highly useful if the data was fully credible for accountability purposes.

Digitally Democratic - Type: 'Engagement' - Score: 5/5

While modest in its ambitions it is also robust, and effective in its design. A gold standard for decentralised digital democracy projects.

38 degrees

38 degrees is a petition and signature collection tool. The site is designed to stimulate community action.

Anyone can suggest a petition, the signature target is set, and is well integrated with social media and designed for funding.

The site's function is two fold, bringing pressing issues to the foreground in a simple process of deliberation via vote, while stimulating community action by bringing individuals together with similar views. The later function is an impressive example demonstrating how digital democracy is especially effective at facilitating alternative political engagement.

A great example of the limitations of Tsagarousianou's definition of digital democracy, 38 degrees blurs the line between deliberation and engagement, highlighting the precariousness of strict delineation in this way.

Democratic - Score: 5/5

While you could point to bias in the user base, which self selects for those already most politically engaged, the lack of a solution to this problem and attempts to mitigate this by spreading messages over social media is commendable.

Digitally Democratic - Type: 'Deliberation & Engagement' - Score: 4/5

Credit is due for its user friendliness and aesthetic design which was a non-trivial problem in previous decades.

An ambitious project, the simple signature based deliberation method leaves no room for discussion, despite effectively bringing pressing issues to the forefront.

While the projects regularly appear to stimulate alternative political engagement, 38 degrees struggles to integrate with existing government systems, and explicitly states its exasperation as government action on moderately signed petitions is limited.

The Information Gap

From this point forward the dissertation will narrow its focus to stage 1 digital democracy, 'information provision', specifically to address the information gap. In the context of this dissertation the information gap refers to the inability for more than half of UK constituents to name their local representative, let alone what they voted on, or how they voted in a division (House of Commons, 2003).

In the spirit of the current era of digital democracy, I will briefly address some potential concerns that a better informed citizenry may produce.

Namely, if the general population becomes motivated to vote primarily based on their representative's performance, does this lead to populism and a limited mandate? Significant theoretical debate questions the effectiveness of a state where the most successful candidates contort themselves to the will of the people, rather than lead on decisions (Hilbert, 2007). The fear being that the tendency toward short-term thinking is exaggerated, risking the long-term prosperity of the nation (The Democratic Society, 2023).

I would suggest that given the current UK context this threat is unrealistic, but credible. I would also point to the earlier model of political will, purely to illustrate how representative (R) data is introduced simply as another consideration factor rather than the sole consideration in decision making.

If for example a labour representative votes against their party on a particularly sensitive issue to a constituent, this may not be enough to shift an individual's vote alone if the party is generally in favour, and there lacks an especially strong alternative representative.

This context shows how the current party based system limits the degree to which populism can take hold. Borris Johnson & Donald Trump also demonstrated that populist attributes such as humour can already be used to seize and entrench power. Charismatic leaders in political parties may encourage populism in a system that lacks transparency, because no metric can be used to measure their cynicism or performance, thus we cannot hold them to account by voting (Hilbert, 2009).

A healthy scepticism toward the positive impact digitally democratic projects can have, introduces a third dynamic to which visualisations should be constrained. We should consider what is possible, and what is realistic, but also what will actually achieve our objectives.
It may be a safe assumption that closing the information gap is a good idea, however Hilbert (2007) expresses concern over an Orwellian future where algorithms collect data and suggest who to vote for without any critical evaluation from the consumer. Such capabilities seem achievable with even limited user data, but should we build an algorithm that tells you who to vote for? No.

As such when identifying which visualisations to create, one must consider:
What **could** we create?
What is **actually** available?
What **should** we create?

Large Language Models

Why use LLMs?
Admittedly, part of the attraction is the novelty of such models. Their limitations are yet to be understood thoroughly, so testing them against traditional data science methods is exciting.

The launch of chatGPT from OpenAI in November 2022, is the first time the public has access to models as powerful as GPT3 (OpenAI, 2022). The rapid rate of progress in the field during the dissertation process meant that models such as GPT4 became available by March 2023 which I was able to utilise for latter tasks (Dilmegani, 2023).

When discussing the ‘power’ of a LLM, I am really discussing three dimensions.

1. The number of parameters
2. The model architecture, and training
3. The feature list

Parameters
The number of parameters in GPT3.5, reached 175 billion (Prakash, 2023). OpenAI has declined to disclose the number of parameters in GPT4, but reports suggest it should be measured in trillions of parameters (Shevchuk, 2023. Lubbad, 2023). Significant debate existed around the effectiveness of parameter scaling, but the success of this approach was largely settled from the ‘Sentiment Neuron’ paper by Radford (2023. Brockman, 2022) onwards.
We can assume two limiting factors on the impact parameters can have on model ‘power’. The first is diminishing returns. An approximate 1,000x bump in parameter size between GPT3 & 4 doesn’t correspond to a one-to-one improvement in performance. For comparison between the two models, benchmarks such as ‘ability to pass the Bar’ are used. GPT3.5 scored in the bottom 10%, while GPT4 scored in the top 10% (OpenAI, 2023).
The second is the quality of parameters, ‘Garbage in, Garbage out’ is the phrase often used to describe this limitation (Simmons, 2023).

Architecture

'Large Language Model' is a somewhat misleading name, because GPT is not based on any single technique. GPT4 is described as a large multimodal model (OpenAI, 2023), fundamentally based on a decoder only transformer architecture (Prakash, 2023). It can simply be understood as a text-based generative model to avoid mischaracterisation due to non-disclosure of certain specifications.

A series of improvements occurred between GPT3 and 4 according to Sam Altman (2023), the major innovation with chatGPT itself was to train the model with Reinforcement Learning with Human Feedback (RLHF), during the tuning stage of model training (OpenAI, 2022. Chamanth, 2023). This A/B testing of results helps to improve the readability of model outputs for human users, among other tuning effects. This technique induced some unintended consequences such as biassing probability predictions along with other human biases (Itzhak, 2023).

Features

The feature list is a different kind of dimension, because it is related to the model's usability in practice. As will be discussed in model selection, model capabilities such as character limits, cost, and API extensions became the dominant consideration factor.

Hallucinations

Hallucinations are imagined references to ideas the model has about a topic (OpenAI, 2023). This limitation is reduced in GPT4 but persistent (OpenAI, 2023). Anecdotal evidence exists that clever prompt engineering can significantly reduce hallucinations. Asking the model to begin a response 'my best guess is..', potentially works because it has some idea of what 'truth' or 'real' is within the model (OpenAI, 2023), however this prompt engineering technique is not formally tested (Benson, 2023).

We can assume that a percentage of the responses produced in the final output files contain these hallucinations, however it shouldn't impact the classification task it was attempting to complete.

Prompt design vs engineering

Prompt design refers to the method of interaction between user and LLM. Prompt engineering is a specialist version of prompt design which focuses on optimising the accuracy and consistency of outputs to achieve a desired goal (Google, 2023). With the models only available to the general public 8 months before the project, prompt engineering is a relatively novel technique requiring much testing, trial and error. Best practices are sparse, and misinformation on how to best optimise prompts are frequent.

In the end, LLMs played an integral role in successfully producing the results. I will share the techniques used throughout the methodology, and findings.

Methodology

The project was fundamentally a NLP task. To avoid bias, it became desirable to organically produce results from the data up, rather than dictate categories from the top down.

As such the challenges in this dissertation can be broken down into three parts:

> **Clustering** – Can the vast amounts of divisions be simplified into digestible categories? How many categories?
>
> **Classification** – Does a division positively or negatively impact the given cluster?
>
> **Visualisation** – How to best present the data?

Initially I planned to cluster the divisions using a traditional NLP method, such as K-means, DBSCAN, or hierarchical clustering. Then use LLM's to classify the debate-as-text files as either positive, or negative for the previously defined clusters.
In practice LLMs were better at both tasks, classifying, and clustering the data.

'They work for you' had claimed to previously solve the classification task but not explained how. Their description indicated that a combination of expert humans, a tree logic based algorithm, or sentiment analysis was used. Passing the division title, containing words like 'Amendment' and 'motion to disagree' to achieve the categorisation. This felt like a very difficult and fragile approach. Assuming there was enough information in the title, I didn't have time to work through the dataset by hand in order to create a substantial test set of data. I also lack the expertise to consistently categorise the data myself.

The lack of accurately labelled training data, and relatively small dataset for machine learning (2,000 vs 10,000 for an ideal minimum machine learning dataset), ruled out neural nets as a classification technique. This could potentially be revisited with labelled data.

The large language model 'GPT4' was used not because of the model design itself, but because of its accessibility, and feature list. Specifically 'document scan' which was an addition to 'Bing Chat' via Microsoft's acquisition of OpenAI. While large language models at the scale of GPT4 are successfully transferable, smaller models are inaccurate when prompted with language outside of the models training data (Chamanth, 2023). Without access to the parameters inside the GPT4 model, document scan was a key feature to ensure the model was exposed to some relevant data during classification.

The process will be described in greater detail in the classification section, however in the interest of reproducibility, I will describe some functional steps that were required to produce the results.
Start by opening the 'Hansard debate' .txt file in Microsoft Edge as a fresh tab, disabling safety, and enabling viewer access to the 'BingChat' sidebar. Each division had a unique algorithmically

produced prompt crafted in R, which was pasted into the sidebar manually because no API was available at the time of writing. The model took approximately 30s to produce a response, theoretically this could have had a total run-time of 16 hours excluding prompt engineering and testing. The lack of an API meant that the manual file extraction turned a 16 hour task into 20 hours even running two browsers in parallel.
Further limitations included a daily limit which could be exceeded by running multiple accounts, and a hard 2,000 character limit which occasionally broke prompts due to long division titles.

Testing

Testing the model output meant running the same prompt multiple times to check consistency of output format, and classification. Often using the model itself to identify opposing perspectives on an issue was useful, because it knew what ideas and words it considered to be different. It also contains subject specific expert knowledge which I did not, and this helped build A/B perspectives on an issue.

Data

Voting data is useful to hold MP's to account, because it pins down their opinion at a given time. The Hansard public record of divisions (votes) is missing data between 2004 & 2006, as such I have selected a time frame between 01/01/2006 - 23/07/2023 to which all data will be accurate.

The Parliment.data API contained 3,972 divisions in that time period, the data contains [Division Title, Date, Number of Votes For, Number of Votes Against, MP names that Participated in the Division].

The API date and title was used to match divisions on the Parliament online website which counted 3,967 divisions in the same time frame. The website was used to manually collect 2,617 'Debate as Text' .txt files which could be matched to divisions in the API. These files contained between 100 & 10,000 words approximately, with an average around 2,500 words linked to discussion preceding a given division.

The missing divisions are due to a combination of missing entries and discrepancies in title and dates being between the API & Parliament websites.

Data Engineering

More than 2,000 divisions is overwhelming, making the information difficult to consume. The first task was to simplify the data by clustering divisions into similar clusters. The debate as text files were too large for many NLP techniques, however 'Titles', and later 'Summaries' of the debates as text files created using LLMs could be used for clustering.

Titles took the form:
‘Illegal Migration Bill: motion to disagree with Lords Amendment 1’
Where: Yellow - Key title information preceding ‘:’ , Green - The type of division called , Blue - The specific subject of the division.

This pattern was inconsistently applied, and would often hide critical information in brackets after the colon which was irretrievable because brackets also contained non-target information.

Preprocessing took 12 steps. Which included, replacing several abbreviations so that EU & European Union could match across titles, removing all text after the colon, removing ‘Stopping words such as ‘a’ or ‘and’, before finally reducing the length of the processed string to the first 3 words. Three words were found to optimise information for clustering based on silhouette scores. 2 words experienced significant loss of information shown by low silhouette scores, while 4 contained extra information that grouped non-similar titles.

Clustering

Most traditional clustering techniques required pivoting the titles into a normalised term document matrix, where rows represent titles & columns represent words across all titles. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) preferred a Term Frequency–Inverse Document Frequency (TF-IDF) matrix, but were otherwise similar in process.

The methods are discussed below:

K-means clustering

Using the elbow method, I failed to identify ideal clusters for K, because the curve was smooth, gathering the majority of its effectiveness by 550 clusters.
Obscure clusters were identified, putting ‘Adjournment Summer’ & ‘Holocaust Memorial’ in similar clusters. Little effort was given to optimising this technique, due to the poor initial clusters, although it effectively grouped like terms. The uniqueness of certain titles hindered k-means significantly.

DBSCAN

An exciting method, successfully clustered groups when given a low minimum cluster value of 2, and epsilon 2.14. The problem with this method is that no matter how the parameters varied one cluster would explode in size before a significant enough number of total titles would become clustered. Of the 779 unique titles after preprocessing, 603 were left unclustered before cluster 6 expanded massively absorbing all other clusters. The maximum clusters that could be made was 32, with 603 titles unclustered.

Hierarchical clustering

Forming clusters using a Jaccard distance matrix, and Ward.D2 method was the most successful strategy for clustering using traditional techniques. Using silhouette scores to identify peaks, the optimal amount of clusters was around 700 with a small secondary peak around 30. Formatted as a dendogram the clusters can be visualised but with 700 clusters, all information is lost.

Initially an attempt was made to try and use the 30 large clusters as inputs to ChatGPT3.5 which would attempt to produce a cluster title for that group. This capability would be similar to early language models such as 'Word2Vec'. The clusters suffered as they were at the same time not unique enough between clusters and two diverse within clusters. I believe the titles did not contain enough information alone to successfully complete this task using traditional NLP methods.

LLM

At this stage ChatGPT-3.5 was used to identify 12 clusters, including 1 miscellaneous used to identify poorly defined categories. The model is trained with human interaction, which meant the clusters were relevant and interesting.
The model was highly successful at identifying clusters & combining the titles in a single heading such as 'Immigration & borders'. GPT3.5 was inaccurate at a secondary task asking it to iteratively assign titles to each cluster via chat GPT's API, because the separate prompts would launch a new 'conversation' not remembering the previous cluster process. This would regenerate clusters with each prompt. The model's inherent randomness meant much of the cluster assignment had to be checked manually. 2,189 titles out of the 2,617 divisions fit into the 11 clusters excluding miscellaneous.

In the first task, the LLM took only the list of pre-processed titles and a prompt. The parameter for creativeness, 'Temperature' in chatGPT-3.5, was set to 0.1 to try and reproduce similar clusters if necessary. The prompt engineering used a few techniques to achieve the desired output: Parameters, Demonstration of expected output, Defined objective function, and Explained inputs.

Prompt:
"temperature 0.1, I want you to cluster a list of titles into a few new categories. I expect your answer to state: 'Cluster 1 title' - 'title in that cluster 1', 'title in that cluster 2'.... Each cluster should be a single word, or maximum 3 words long. You will be rewarded for fewer words, but punished for loss of accuracy, or information. To help, each group will be associated with political discussions, and you are trying to group them. Each title will be separated with '/'. I will give you the titles in the next message, for now just reply if you understand?"

While generating clusters, a secondary prompt utilised GPT3.5s chat memorisation to split the prompt across two conversations, avoiding the character limit. The second prompt was simply a

list of each unique titles post-processing. This was 779 titles long, separated in the manner described in the prompt: ‘immigration borders/summer adjournment/etc’.

Response:
“Clusters: ● Cluster 1: Immigration & Borders ● Cluster 2: Economy & Financial Services ● Cluster 3: Education & Learning ● Cluster 4: Health & Healthcare ● Cluster 5: Environment & Energy ● Cluster 6: Crime & Justice ● Cluster 7: Welfare & Social Housing ● Cluster 8: Standards & Technology ● Cluster 9: Parliamentary Procedures ● Cluster 10: European Union & Foreign Affairs ● Cluster 11: Defence & Armed Forces ● Cluster 12: Miscellaneous”

Classification

Before visualising an MPs voting record, I need to identify how each vote impacts its given cluster.
A problem identified by ‘They Work for You’ is the complexity of divisions. A division can be a vote for, or against a bill, an amendment to a bill, or a motion to disagree with an amendment to a bill. It takes experts to understand the complexity of any debate, before breaking down the logic behind each division call. They Work for You expect frequent errors, and so rely on crowdsourcing to review their data.

I propose that with very tight prompt engineering, a large language model could outperform individual experts, and take a fraction of the time. For each division we are looking for either a 1 or a -1 corresponding to the relationship between the division and its cluster. I.e.
Will, voting ‘For’ the division:

‘Illegal Migration Bill: motion to disagree with Lords Amendment 1’

Tighten or loosen immigration & border control?
If it ‘tightens’ we can call this Sentiment: 1, and if it ‘loosens’ we can call this Sentiment: -1. That way when paired with a vote: Ayes: 1, Abstain: 0, Noes: -1, we can identify a politician's stance by taking the product of sentiment and vote.

An average prompt of length 1500 characters can be used to produce:
‘Summary of debate (50 words), Summary of arguments For division (30-50 words), Summary of arguments Against division (30-50 words), Sentiment, and Justification for decision’.

To produce an output like this by hand would be a day's work for some of the larger text files. GPT4 could produce the output in 30 seconds.

Prompt Engineering

Some very important 'prompt engineering' strategies are used or the technique will fail. The LLM produces one word at a time, it produces the most likely word given what it knows.
Asking it to justify its decision, before making a final classification drastically improved its accuracy, and reduced hallucinations. This appeared to work because justifying its classification gave it time to 'think', and the predictive nature of the next letter A or B would be consistent with its prior logic.

Other techniques such as tagging the text, which involves repeating words across the prompt and document it is scanning, help to locate key information.

For an unknown reason GPT4 will break if using 1 or -1 as a categorization in the text. Through experiment, it failed far less frequently using A / B classification, which I can extract through a text file later on. I have no intuition as to why this error took place although it was persistent in early prompt designs, and appeared to generate the number randomly.

In the final prompt, parameters were not set, because in testing it seemed to have little effect. There was no way to check if setting the parameter in text was being applied consistently.

Explicitly demonstrating what the output structure should look like was key to producing consistent outputs. Consistent outputs were key for post-processing and visualisation.

Improvements to the final prompt design would be to simplify the prompt. Suggestions of simple prompts producing better outputs are common online. I found that this created a trade-off with output consistency which was very important for the reasons above.

Example prompt:
`Review the text file open in my current webpage. First, summarise the debate generally in 50 words. Next briefly summarise the arguments For (Ayes), and Against (Noes) in the division related to: 'Illegal Migration Bill: motion to disagree with Lords Amendment 1'. Finally this file is classified under 'Immigration & Borders', I want you to consider if votes 'For (Ayes)' will A: tighten Immigration & borders , or B: soften Immigration & borders. Please format your response strictly in the following manner: 'Summary - (50 words, about the txt file)', 'For - (30-50 words, voting for (Ayes) in the division: motion to disagree with Lords Amendment 1)', 'Against - (30-50 words, voting against (Noes) in the division: motion to disagree with Lords Amendment 1)', 'Decision - (Justify if the motion to disagree with Lords Amendment 1 either 'tighten Immigration & borders', or 'soften Immigration & borders')' , 'Sentiment - (Based on your 'Decision', return either 'A', or 'B')'
Where: Yellow - Document scanning , Light Green - Tagging , Light Blue - Division Title, Blue - Cluster, Orange - Sentiment, Purple - Demonstrate Output, Green - Explanation of decision making

Example Response:

Bing

Here is my response based on the text file open in your current webpage:
Summary - The text file is a transcript of a debate in the UK Parliament on the Nationality and Borders Bill, which aims to reform the asylum system and deter illegal migration. The debate focuses on the Lords amendments to the Bill and the Government's objections to them.

For - The Government argues that the Bill is necessary to stop the abuse of the asylum system by people who enter the UK illegally, often using criminal smugglers. The Government says that the Bill will create a fairer and firmer system that prioritises those who need protection most and discourages dangerous journeys.

Against - The Opposition and some Conservative MPs argue that the Bill is cruel, counterproductive and unlawful. They say that the Bill will create a two-tier system that discriminates against refugees based on how they arrive, and that it will violate the UK's obligations under international law and human rights conventions. They also say that the Bill will not stop the boats, but will increase the backlog, the cost and the suffering of asylum seekers.

Decision - Voting for the motion to disagree with Lords Amendment 1 would tighten immigration and borders, as it would reject the amendment that requires the Bill to comply with international law and treaties. This would give the Government more power to remove or detain asylum seekers who enter illegally, regardless of their protection needs or rights.
Sentiment - A

Visualisation

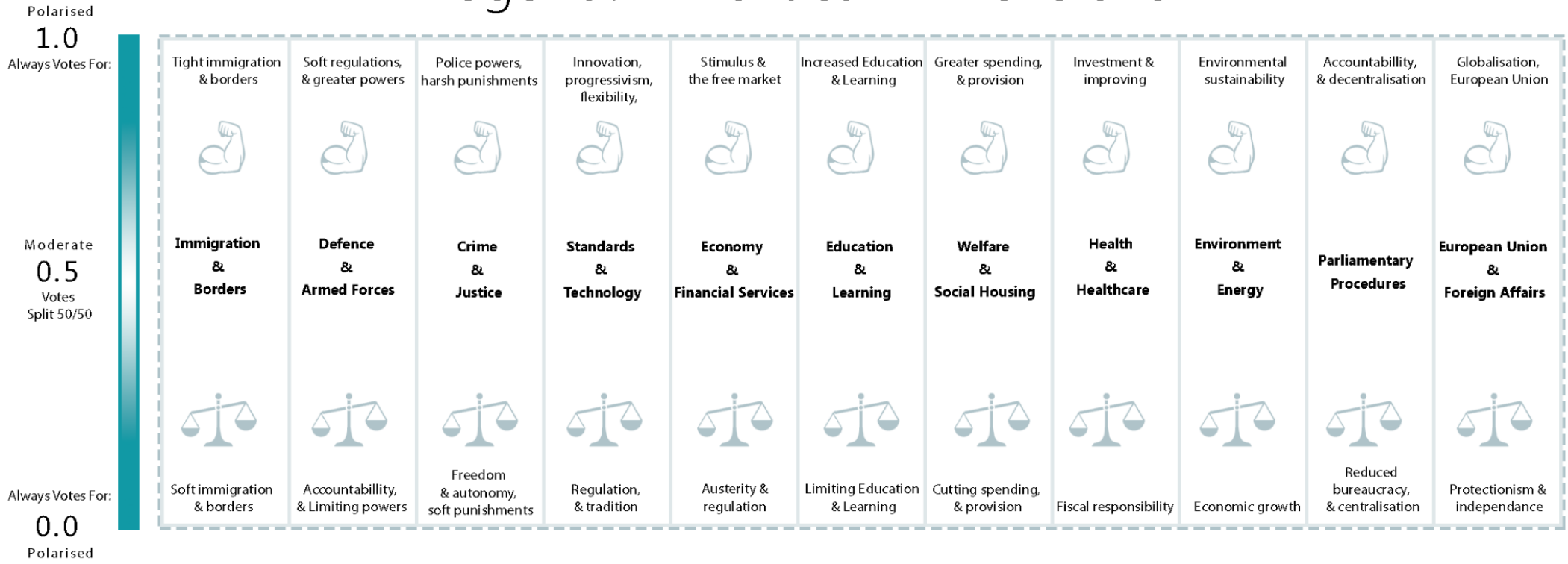
To visualise the diagram, I normalised the votes within each dimension. This meant that the greatest number of votes for a given dimension was set to 1, and the greatest number of votes against was set to 0. In some cases this meant that the neutral stance (abstaining from all votes on an issue) was slightly above or below 0.5, however this is almost unnoticeable in the diagrams. This also helped to limit noise created by miss-classified divisions, which would have been visible if a politicians score was instead out of the maximum possible votes for against a division.

When discussing visualisations with my supervisor Stuart Towley, we discussed producing a radar plot for each of the 11 dimensions would be a strong method for visualising the beliefs of a politician. He suggested that the real value in this plot would be visualising the politicians shifting stance at different points in time, and that this would be reflected in the shape of the diagram. For this suggestion I am extremely grateful, as it led to some fascinating insights and helped to identify patterns that would otherwise not have been so clear.

It is important to note that 1 does not mean good, and 0 does not mean bad. Instead if a politician scores closer to 1, or 0 it implies they have stronger opinions on an issue, while approximately 0.5 indicates a more moderate compromising stance.

The key is very important to understand the plot. Due to the complexity of political issues such as ‘Economy & Finance’, where nobody chooses to vote ‘Against’ the economy. The difference in A/B classification reflects opposing policy, e.g. 'Stimulus & the Free Market’ vs ‘Austerity and Regulation’. This appeared to still work effectively, even compared with simple dimensions such as ‘Education & Learning’ which could be easily classified into ‘For’ vs ‘Against’.

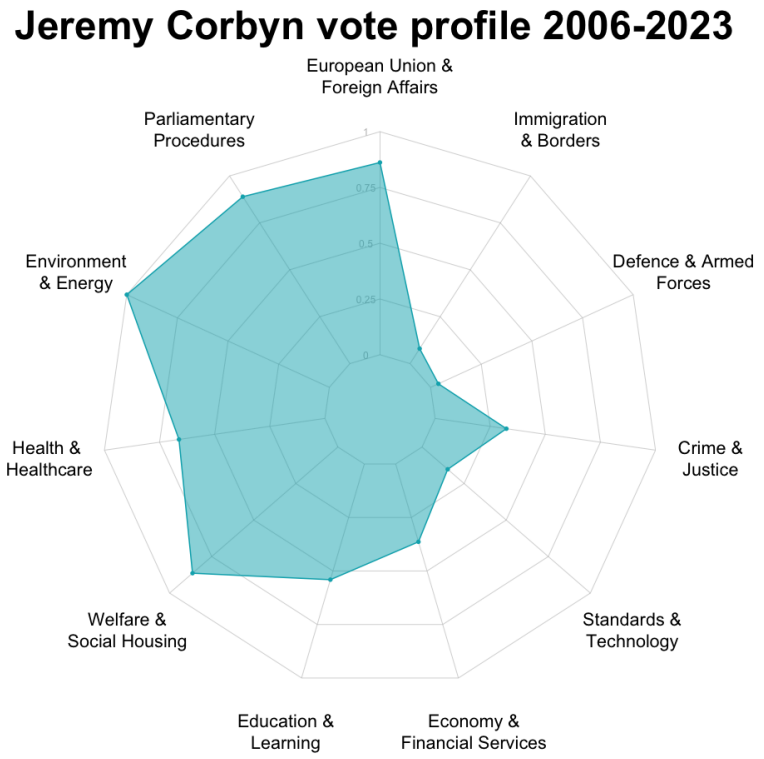
Legend: 11 Political Dimensions

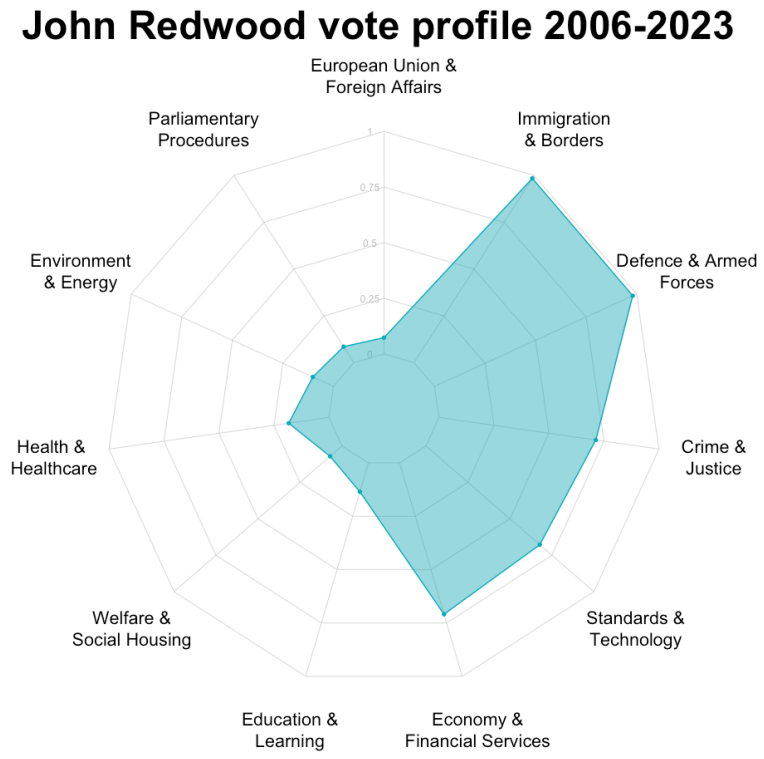


Deciding which policy should be associated with 1 and which with 0, poses a further problem. This is supposed to be arbitrary, but as will be discussed in the limitations people tend to associate 1 with 'pro' an issue, and 0 with against. This meant that the decision could be made logically in most cases, although judgement calls were made for 'Economy and Financial Services' where stimulus was associated with 1 and austerity 0. The chart was organised so that traditionally left wing issues associated with 1 are on the left of the diagram and traditionally right wing issues associated with 1 are positioned on the right of the diagram.

This helps differentiate between political stances by shape in the diagrams [1] [2] below:

(Jeremy Corbyn - Labour Vs. John Redwood - Conservative)





Findings

Transparency & Accountability

Visualising the voting profile of politicians helps to achieve the aims of this dissertation. By demonstrating how representatives are likely to vote, constituents can hold them to account by voting for, or against them in the next election cycle. This was best achieved by animating the profile 10 divisions at a time, effectively representing one month per frame. These gifs can be found in the supporting material to this dissertation.

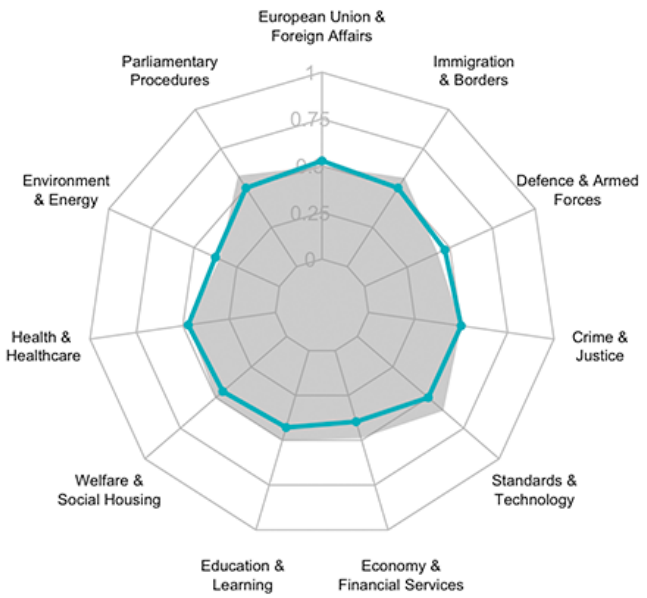
I have split 4 MPs voter profiles into 5 periods to be compared within the dissertation itself. Boris Johnson, Jeremy Corbin, Keir Starmer, & John Redwood, were selected arbitrarily, for the purpose of demonstrating how the visualisation can capture changes in position over time. The below plots are broken into a series of 417 divisions which represents 3.4 years on average, approximately the duration of a sitting parliament.

[1] Lilley-Townley Diagram Left - Jeremy Corbyn
[2] Lilley-Townley Diagram Right - John Redwood

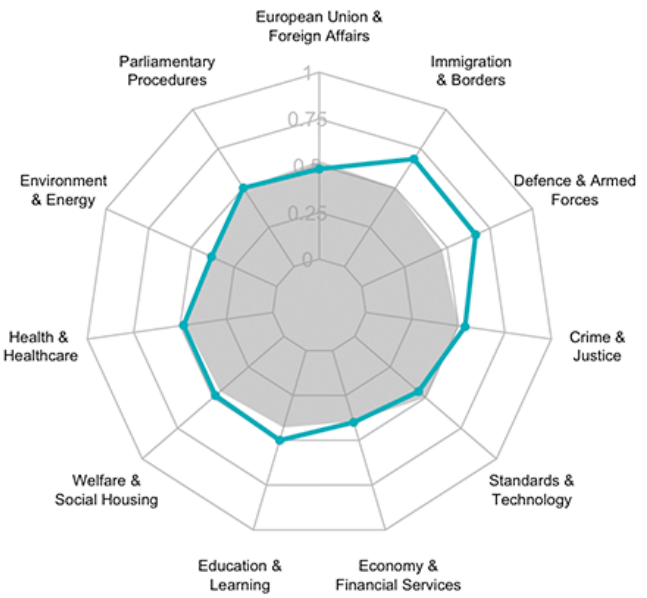
Vote Profile - Boris Johnson

Boris Johnson 2006-2009

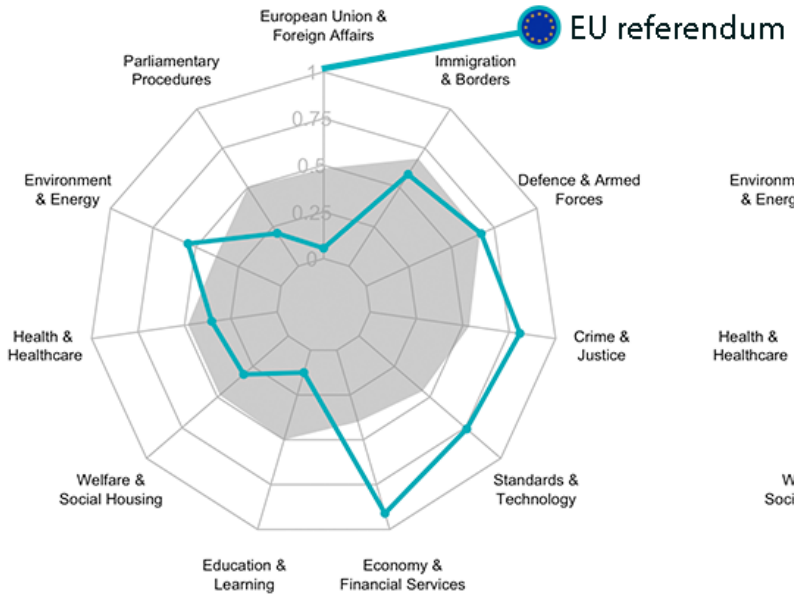
Boris Johnson 2009-2012



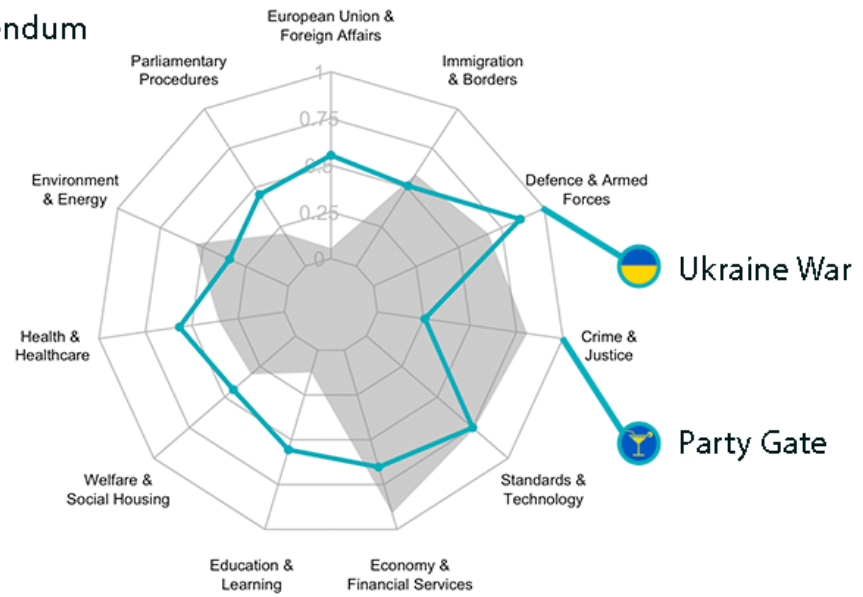
Boris Johnson 2012-2016

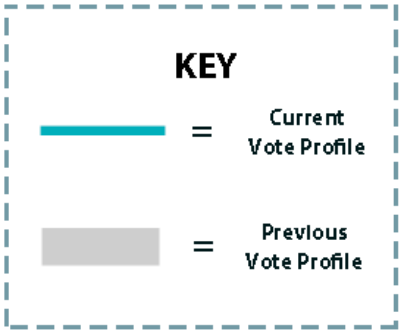


Boris Johnson 2016-2021



Boris Johnson 2021-2023



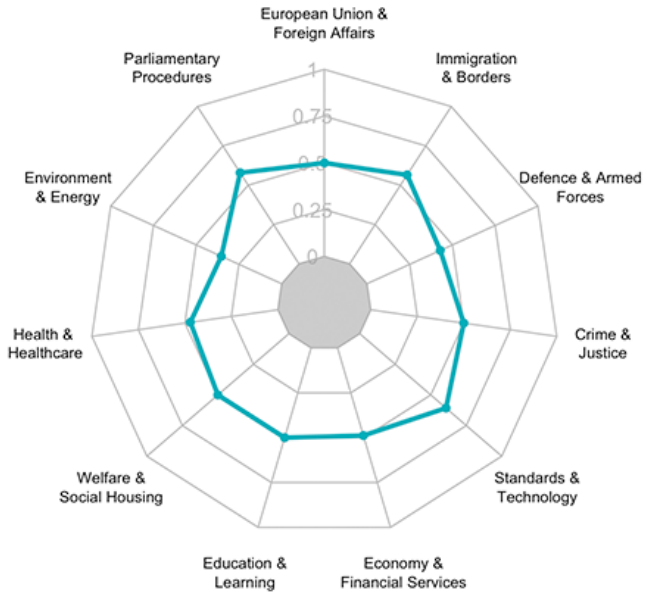


Notice in the graphs above Boris Johnson in his early career is fairly moderate. As mayor of London, and MP for Ruislip. Only after the EU referendum (2016) does Boris begin to hold any strong views about ‘EU and Foreign Affairs’. Later as MP for Uxbridge, and as Prime Minister his profile captures both his pro military stance toward Ukraine in the 2021 - 2023 profile. Also a shift in Crime & Justice as he is accused of breaking lock-down rules, and parliamentary procedure (2021 - 2023).

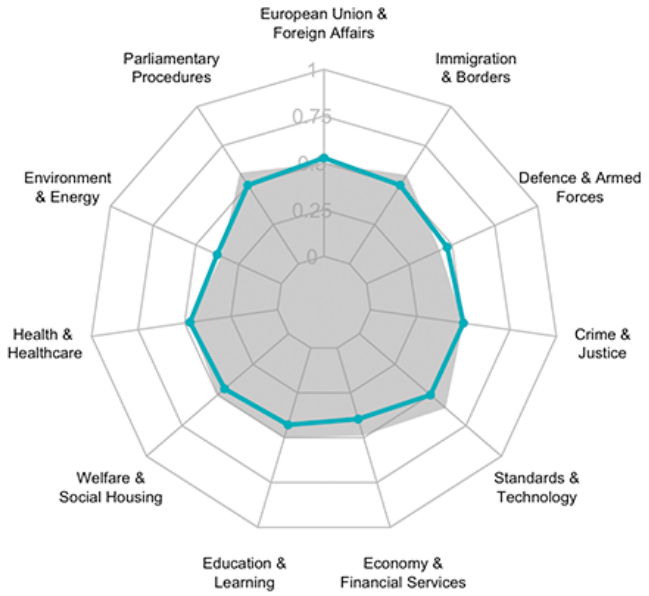
The graph is able to capture much of these events as it contorts based on current affairs, but shows very little lasting trends. The shape’s convulsions between 2016-2023 reflect a fickle stance toward politics generally.

Vote Profile - Keir Starmer

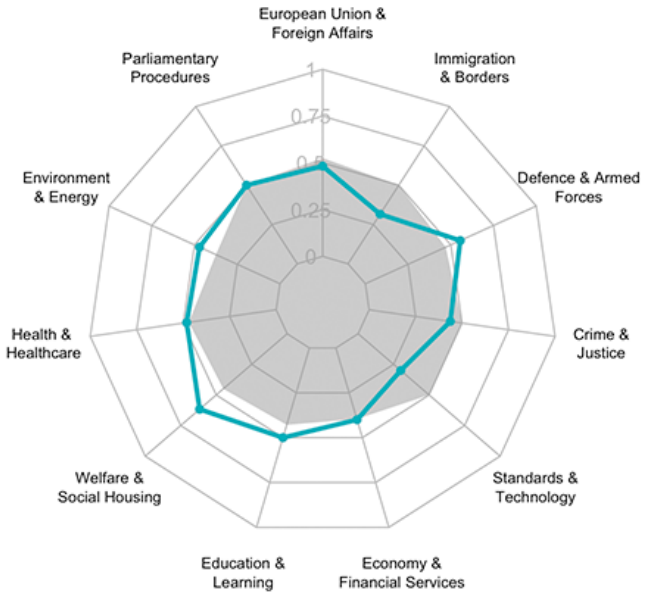
Keir Starmer 2006-2009



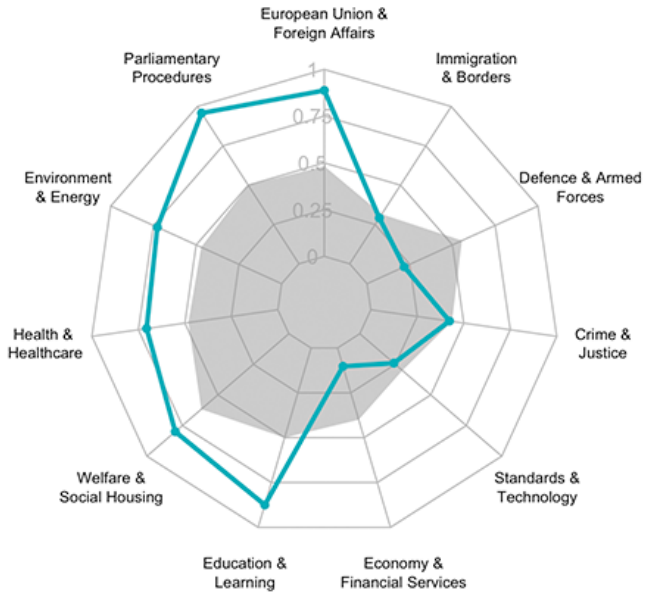
Keir Starmer 2009-2012



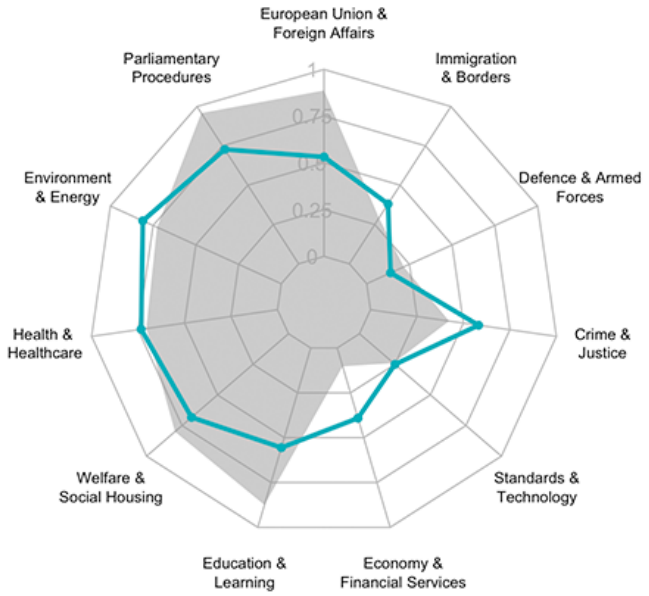
Keir Starmer 2012-2016

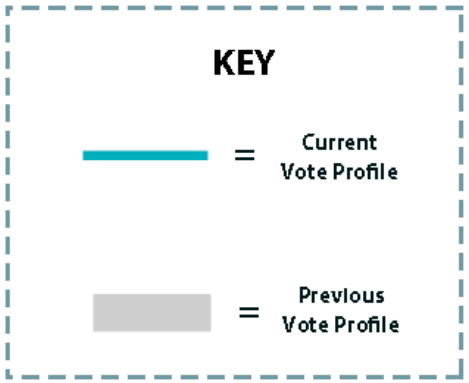


Keir Starmer 2016-2021



Keir Starmer 2021-2023





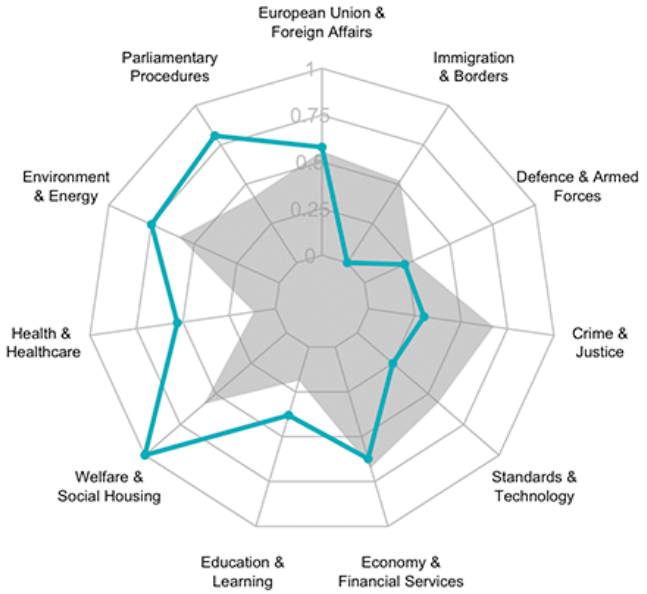
Sir Keir Starmer is similar to Boris in the early 00s & 10s, however clear polarisation in his views between 2016 - 2021 can be seen as the graph extends toward generally left wing issues.

Generally thought of as a more centrist leader for the labour party. It is worth pointing out the contraction in ‘Parliamentary Procedures’, ‘EU & Foreign Affairs’, ‘Welfare & Social Housing’ and ‘Education & Learning’ to within the 0.5 - 0.75 range between 2021 - 2023. This step away from polarisation and toward a more moderate stance will be revisited in the next section.

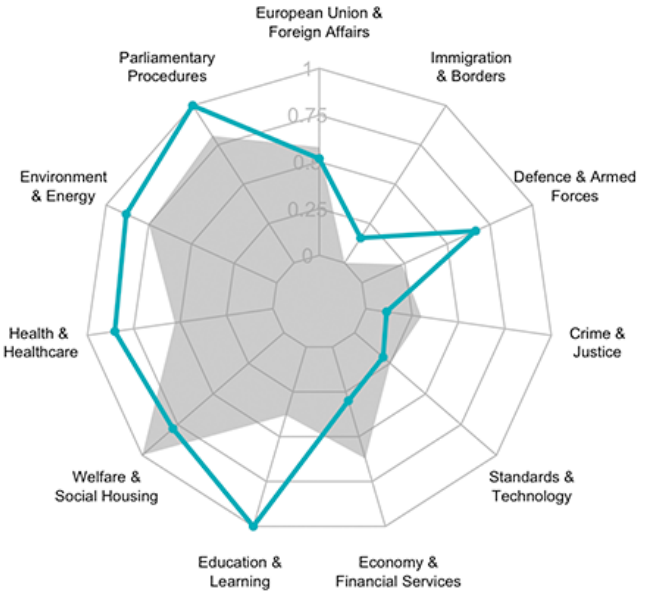
Vote Profile - Jeremy Corbyn

Jeremy Corbyn 2006-2009

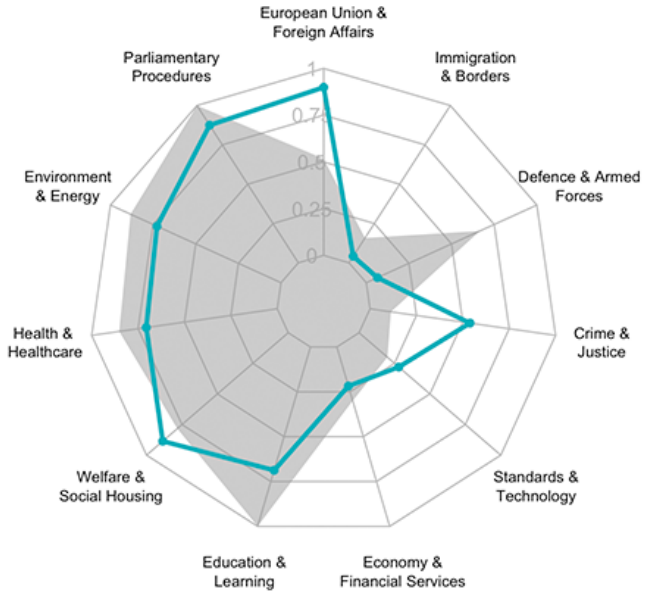
Jeremy Corbyn 2009-2012



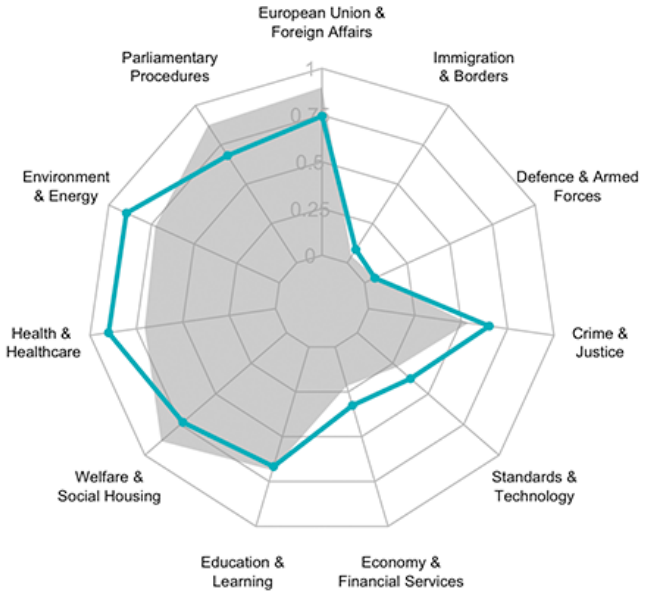
Jeremy Corbyn 2012-2016

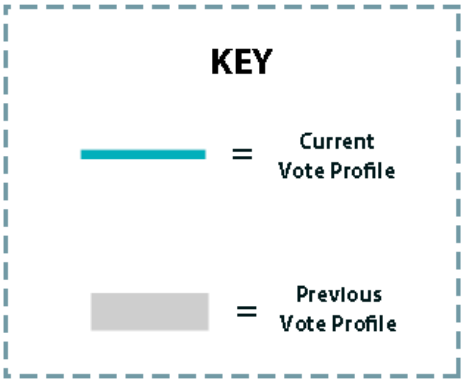


Jeremy Corbyn 2016-2021



Jeremy Corbyn 2021-2023





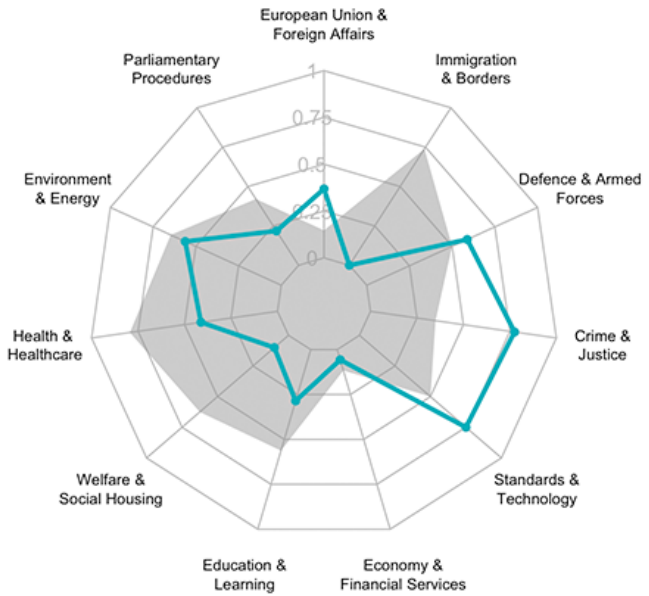
Jeremy Corbyn is useful to compare with Keir Starmmer as he is understood to be more radical in his views. This is reflected in his voting behaviour, matching the most opinionated MPs on ‘Immigration & Borders’, and ‘Welfare & Social Housing’ between 2009 - 2012, as well as ‘Parliamentary Procedures’, and ‘Education & learning’ between 2012 - 2016.

Mr Corbyn shows a clear and persistent shift towards a left wing political stance from 2012 until 2023. He generally takes a more extreme stance than Sir Keir Starmmer, despite both of them trending toward a similar shape in latter years. It is this similar shape within political groups that indicates polarisation in the House of Commons.

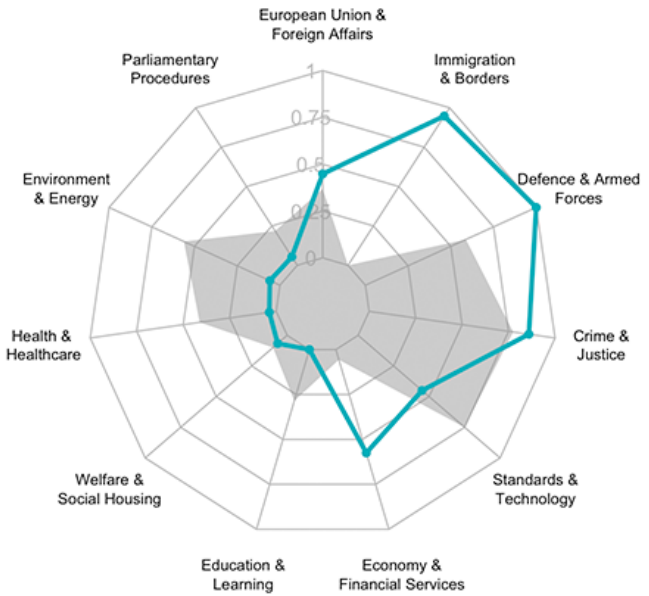
Vote Profile - John Redwood

John Redwood 2006-2009

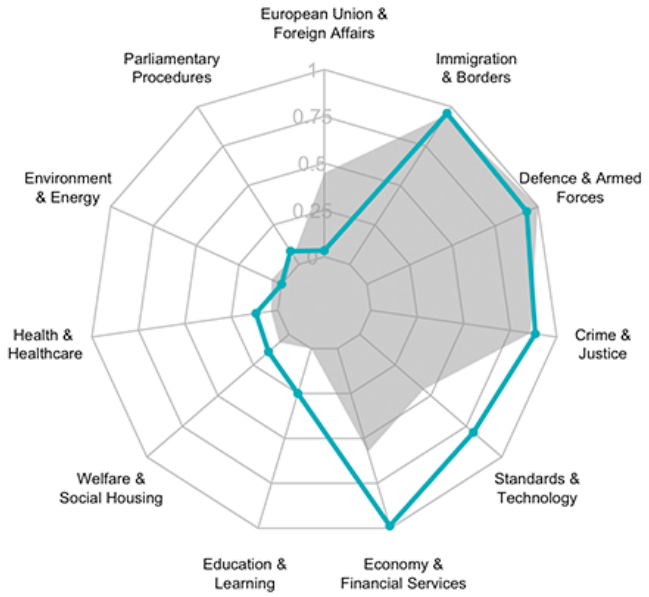
John Redwood 2009-2012



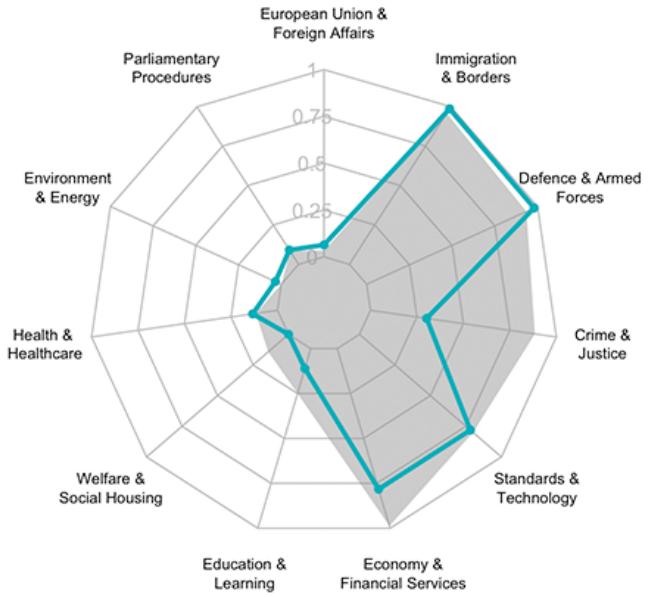
John Redwood 2012-2016

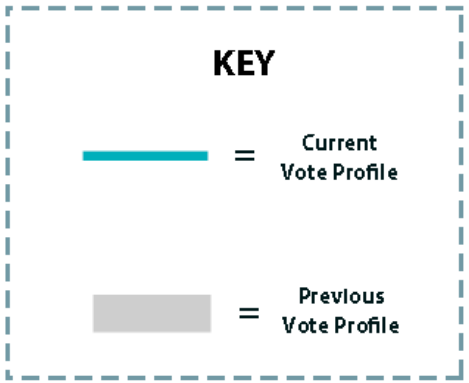


John Redwood 2016-2021



John Redwood 2021-2023





A final comparison to be made between Jermey Corbyn representing a staunchly liberal MP, and John Redwood a predominantly Conservative MP. Both show strongly opinionated graphs between 2006 - 2012, but no clearly similar shape emerges.

Beginning in 2012, a much more consistent set of values are demonstrated. Mr Redwood votes extremely conservative, especially on issues such as ‘Immigration & Borders’ and ‘Defence & Armed Forces’. Clearly taking opposing stances to Jeremy Corbyn, you can also see that the two vote against each other on ‘Crime & Justice’ between 2021 - 2023, reflecting their support or lack thereof for Boris Johnson.

The visualisations are extremely successful at capturing the attitudes of MPs, which should help constituents understand what they are voting for.

Polarisation

The data can also be explored more generally, looking for trends that might be insightful.

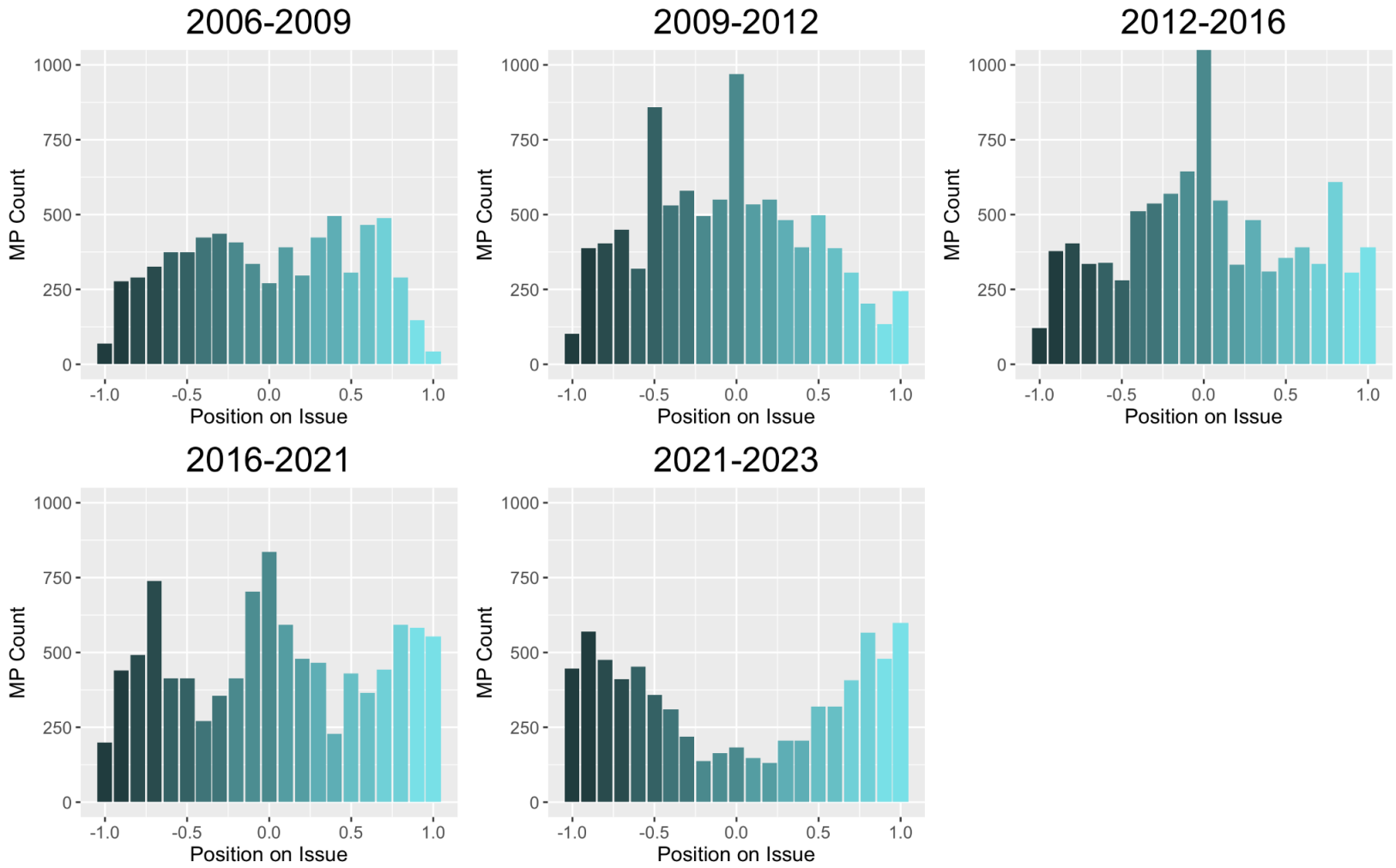
Across the individual MP graphs, a consistent trend of distortion in voter profiles can be identified. MP’s generally begin 2006 - 2016 as fairly moderate, before taking more extreme positions from 2016 - 2023. These latter dates coincide with Brexit, and Boris Johnson's election as PM, before ending with covid-19, Boris Johnson's resignation, and the Ukraine war.

This is a somewhat disconcerting finding. Polarisation can be difficult to articulate generally, but in these visualisations it can be clearly identified. Polarisation in the House of Commons is expressed as, similar voter profiles within, and opposing profiles between, party lines.

Because the voter profiles are normalised in the method above, we can count the number of MPs at each point on any issue between 0 & 1. During moderate periods the graph should be roughly normally distributed, with the bulk of politicians voting for or against policy roughly equally. During periods of polarisation, we can expect that distribution to invert, where the bulk of MPs will hold more radical perspectives on an issue.

By plotting the distributions on a bar chart, we should see this polarisation escalate between 2006 - 2023. I have renormalised the charts below around 0, with -1 and 1 now reflecting the extremes in position. It is irrelevant if the extreme view is liberal or conservative in this case, because I am only interested in the level of polarisation across the house.

Voting Distributions 2006 - 2023

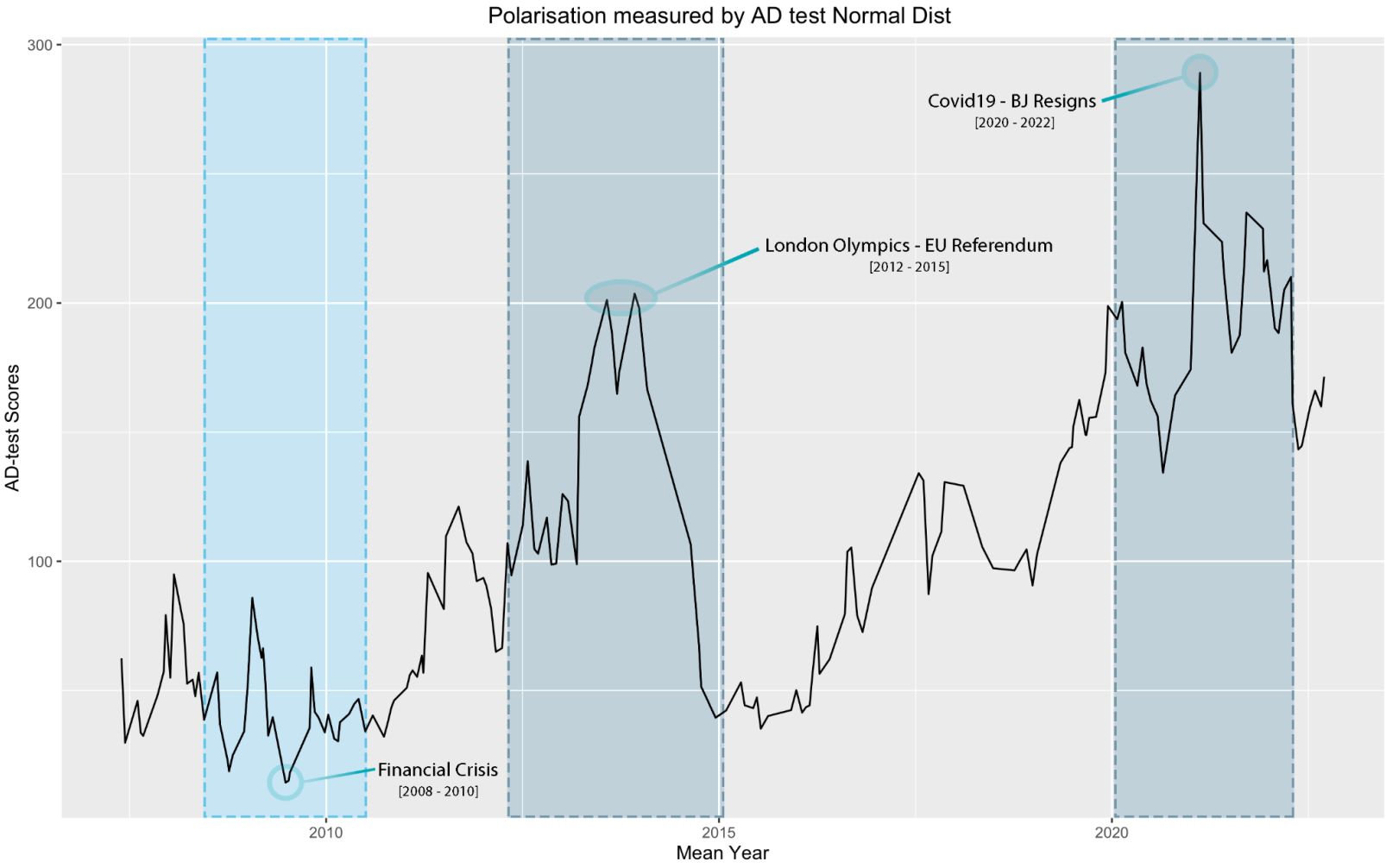


Some artefacts exist in the data because the number of MPs in each time frame changes when transitions between new parliaments occur. The number of divisions in each grouping is always 417 so I have taken no measures to clean the data and risk inducing bias. This explains the variance in scale between plots.

The effect is much easier to identify when animating the plot, where each frame equals 10 divisions, roughly a 1 month time frame. This animation is saved as a gif in the supporting documents.
Polarisation occurs over a period of a few years, first splitting into three distinct groups, seen most clearly in the 2016 - 2021 plot. In periods of extreme polarisation these 3 groups divide into two, creating an inverted normal distribution. This leads to the shape seen in the 2021 - 2023 plot.

Because we are measuring a normal distribution in this case, we can quantify the level of polarisation by taking a simple AD-test.

Instead of the 5 time frames above, I am going to split the divisions into approximately 2.3 year periods consisting of 287 divisions, approximately half the previous time frame. This is done for granularity. I am also going to jump 10 divisions at a time, testing the normality at each 2.3 year period. This is the same process used in the animated gifs. Finally I will plot the AD-test score on a line chart.



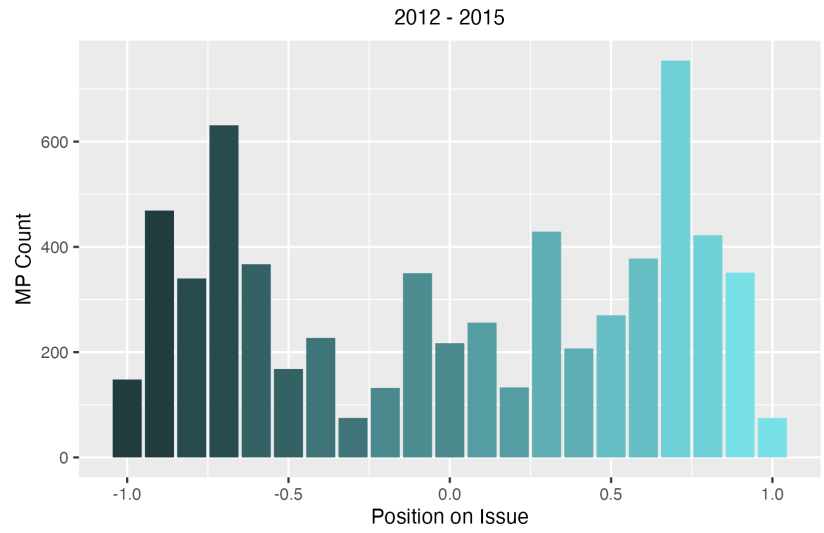
What we have created is an indicator for polarisation in the UK House of Commons. It shows two clear peaks, the first occurring in the range 2012 - 2015, with an AD-test of 203.6. Occurring during a coalition government starting with around the London olympics, ending in the build up to the EU referendum.
The second occurs in the date range 2020 - 2022, with an AD-test of 289.1. Starting with the covid-19 pandemic, and ending in Boris Johnson's resignation.

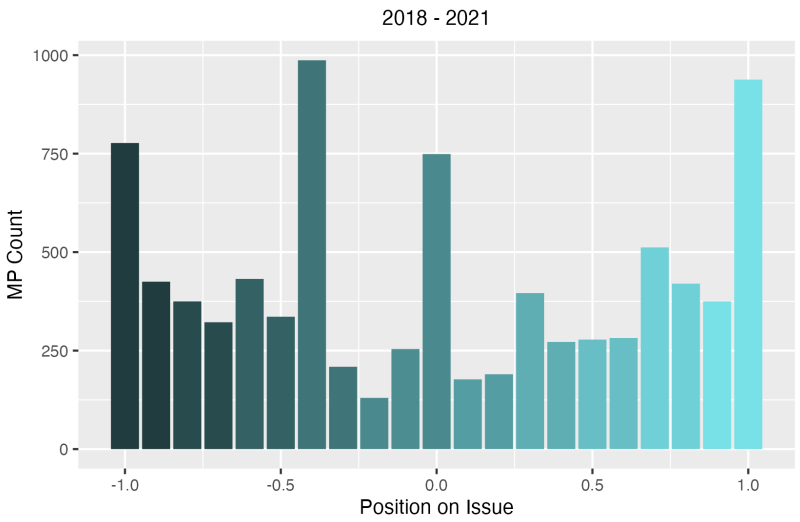
When animating the bar charts on the same 2 year resolution jumping 10 divisions at a time, the two periods of polarisation can be identified. The later peak in 2021 eventually splits into 2 distinct long tailed distributions, forming an inverted normal distribution which triggers the

extremely large AD-test score. The 2014 peak barely remains as three distinct groups, with their means around -0.7, -0.1, and 0.7 respectively.

A final observation is that ignoring the peaks, there appear to be periods of increased polarisation between 2011 and 2015, and a new period starting in the range with the mean year being 2017. This later period appears to be much greater and more persistent.

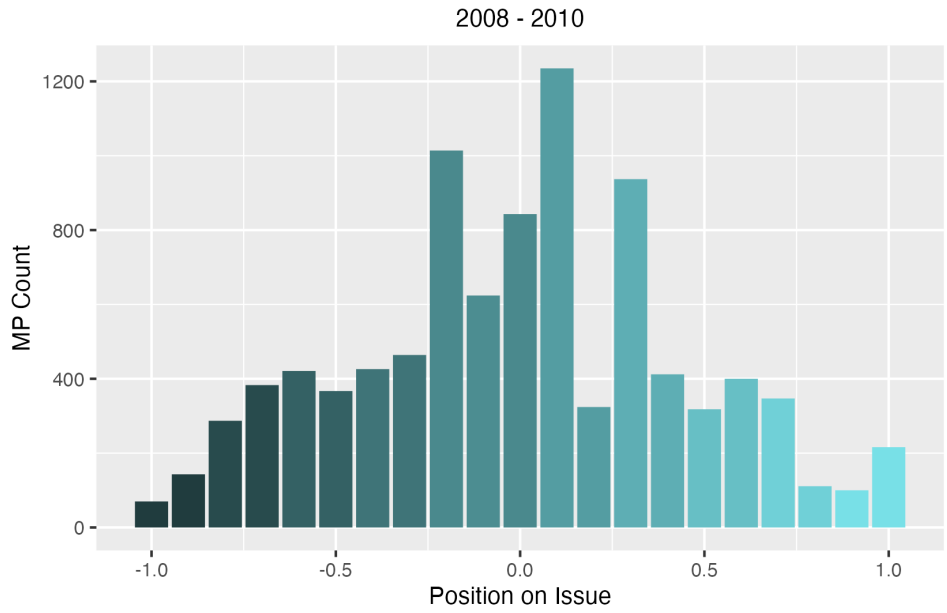
Peak Polarisation Distributions
(Mean Year 2014 Peak vs. Mean Year 2021 Peak)





For comparison, the minimum point of polarisation occurred between 2008 - 2010. This indicates politicians vote independently from each other, For, and Against policy in a way that is roughly normally distributed.

Minimum Polarisation Distributions
(Mean Year 2009 Minimum)



Discussion

Accountability & Transparency
The vast majority of the work done in this dissertation cannot be seen in the write up. Data and visualisation has been done for every politician that voted in the house of commons between 2006 and 2023.

To really make meaningful progress toward the aim of this dissertation the data should first be made publicly available for scrutiny, and others to work with. It should then be made available on a website in which people can access the visualisations and share across social media. Finally, steps should be taken to gather other useful profile data such as response rate, & rebel rate (the frequency a politician votes against the party line) in order to gather information in a single place for better decision making.

Polarisation
A few potential explanations for the polarisation of voting behaviour can be given, further research is required in order to determine which explanation is most likely. Here are some hypothesis to test using the existing data:

The mechanism for polarisation may initially start with individual politicians taking more extreme stances on an issue. Which causes both, opposition MPs to take opposing stances to balance votes tactically, and encourages MPs within their party to form similar views.

Individual groups may trigger polarisation, this could be checked by splitting the data on party affiliation and looking for lag or lead in certain groups. Also clustering similar voting MPs could identify new groups within parties.

Polarisation may be exaggerated by types of government, during the entire period of the Clegg-Cameron coalition (2010 - 2015), polarisation was high.

The types of politicians the party selects as leader may also have a considerable effect. Dynamics such as radical vs moderate, or populist vs traditional politician, should be tested.

The impact of a party whip should also be considered, this can be tested by measuring the ‘rebel rate’ which indicates how often a politician votes against the party majority. The party majority is often used, because party whip is privately held information.

Large Language Models
Despite human intervention in the form of testing, validation, trial and error. LLMs successfully outperformed traditional NLP approaches when identifying clusters. The training data, and intense human interaction helped to produce ‘relevant’ clusters which would normally have to be done by a human. The same cannot be saide for its ability to assign clusters due to a combination of model randomness, and poor prompt design.

The model effectively flipped the traditional roles of human and machine, where cluster assignment needed to be validated by hand, while cluster identification was done automatically.

GPT4 successfully classified divisions in relation to their given category. Although, in the future attempts should be made to identify the failure rate before any comparison can be made between Neural Nets, or traditional classification methods.
Evidence for strong classification capabilities can be seen in the final outputs. If significant misclassification took place, we could expect voter profile shapes to be generally circular, indicating high levels of randomness. Instead the graphs show labour party candidates typically shifted to the upper left of the graph, while conservatives are generally shifted toward the right of the diagram. This effectiveness may be exaggerated by the normalisation within the bounds of the maximum and minimum votes, but clearly enough divisions were correctly classified for the visualisations to be highly expressive.

Limitations

Identifying Polarisation
The evidence for polarisation was not the initial intention for this dissertation. When the trend appeared so clearly I wanted to confirm the finding.

At first I assumed that this occurred due to greater total MPs in earlier divisions, where an MP may vote on one or two divisions and leave all other dimensions as 0. Evidence for this can be seen in 3 of the 5 graphs when split into 417 divisions, however it does not appear to impact the finding generally. These artefacts do create outliers in the results.

I have some concerns that normalising the profiles may distort the graphs in situations where the house is unanimous on an issue. The minimum number of MPs being capped at 615 and the stance on an issue being averaged across all divisions in a given period largely mitigate this effect, but I would like to create synthetic data in order to stress test the methodology.

The more normally distributed 2006 profiles could also result from the LLM struggling to classify earlier divisions leading to high randomness. If the divisions were perfectly randomly classified, the distributions would approximate normality.
This too seems unlikely, because there is no reason to think the model would become less effective, especially as I became more adept at manipulating the model itself. If anything, I would expect greater errors in the later divisions as I gained experience working backward in time, but I painstakingly restarted the process several times in order to mitigate that exact threat. It is possible that the debate-as-text files contain less information or are poorly formatted in earlier debates, but this justification has limited evidence to support it.

Large Language models
Large Language models are still flawed. The methodology above was incredibly fragile, minor changes in the prompt had massive impacts on the accuracy and consistency of classification.

For example, the words ‘associated with’ policy x, was used in the final prompt because the model has been made incredibly cautious about making politically fueled judgments. If I asked directly if voting Aye in a division would tighten border control, it had a very high frequency of giving a C classification.

Classification C was introduced to the prompt as a catch all clause which prevented the model from guessing in very difficult divisions, this classification meant I would discard the divisions post-processing. Despite the sensitivity to the prompt, C classification only occurred in 102 out of 2189 divisions that made it to the classification task.
4.66% C-classification-rate, gives some idea of the model accuracy. Considering all the legitimate reasons C classification was used, including poor clustering, and irrelevant divisions (described below), using GPT4 for NLP classification should be considered an overwhelming success.

Testing was done in blocks of 10 repetitions for a given prompt. When a prompt was finished for a given issue, I would test between 5 and 15 examples to check accuracy and consistency. There were a few different outcomes that would occur:

> Consistent success - The vast majority of final prompts would classify a division in the same way, returning A or B for all 10 repetitions. It can be assumed that there is some error rate, however it was very low.
> Very rarely it would identify C consistently for all 10 divisions. This would usually occur if the division was incorrectly clustered, for example a parliamentary procedure about an MP’s conduct was incorrectly clustered into ‘Crime & Justice’.
>
> Consistent failure - This would occur due to poor prompt design, it would result in random classification 50% A and 50% B. Or C classification would occur at a rate exceeding 50% across different divisions being tested. This usually occurred due to spelling/code errors, or poorly differentiated policy positions. This could be avoided by working with the model to identify opposing perspectives on a given issue. I learnt to identify this through testing, and fix it by editing the prompt before beginning the classification task.
>
> Inconsistent failure - This would occur infrequently and was a limitation of the model. This problem occurred when the classification A or B, was really up to how you interpret the debate file. Often GPT4 would justify its decision by stating ‘In the long-term…’ for classification A, vs ‘The direct impact of..’ for classification B. The model would often be aware of its uncertainty, but would only classify the division as C about 20% of the time. Weirdly the model appeared to have a preference for long-term impacts, which it would classify ‘A’ 50% of the time, and ‘B’ 30% or vice versa depending on which class represented the long-term perspective.
> The fact GPT4 would classify these uncertain divisions either A or B 80% of the time means that there is certainly some degree of error in the data. It is very difficult to know how frequent this occurred during the classification task, in the limited testing it appeared

fairly infrequently. As such in the end misclassification just appears as noise on top of the majority of correctly classified debates.
To limit the impact in the final result, should a C classification occur, I would re-run any prompts that are associated with the same, or similar debate file, giving a C classification manually if any inconsistency was identified.

Irrelevant divisions - There are two failures here. Both are created when a division is correctly clustered based on the content of the Bill, however the division itself is not directly associated with A or B.
Occasionally GPT4 would solve this by classifying the division based on voting Aye for the bill, rather than voting Aye for the division. To prevent this confusion between bill and division, I repeated the division title 4 times in every prompt (tagging the text) which largely stopped the LLM from confusing the bill and the division.
The second failure occurred when irrelevant divisions were classified as A or B. The model could correctly classify the division as C, only about 80% of the time. I would attempt to catch this error manually by repeating similar divisions that occurred on the same day. If one of them returned C, I would then repeat the similar division, classifying it as C if 1 of the repetitions gave this result.

There are two important takeaways from this. Firstly, more powerful models would almost certainly improve accuracy, and consistency. It is not clear however that the safety measures implemented by Microsoft for Bing Chat improved its classification capability on sensitive topics.

The second point worth making is the opportunity for ensemble LLM modelling. During testing it became clear that simply removing, or taking the average of, uncertain divisions would have improved the accuracy of the final outputs.
At the time of writing, there is no LLM with all three requirements: Public availability, Document scanning, and a Functioning API. As such an ensemble approach was not possible due to time constraints with the project.

Conclusion

Polarisation
We can safely conclude that during no period between 2006 and 2023 has the House of Commons been more polarised than it is during this present era.

There are some signs that more moderate politicians, such as Keir Starmmer are beginning to return to the centre. This is promising and reflected in the indicator as we are down almost 30% from the all time high of an AD-test score of 289. This is not particularly reassuring considering the same indicator shows peak polarisation in 2016 roughly comparable to today with a score of 203 then vs 171 today.

Some level of disagreement should be considered healthy within the House of Commons. I am sure that further study into more authoritarian governments would reveal exceptionally low polarisation as politicians fear descent. Given more time and a larger dataset, I would like to explore more fluctuations in the House of Commons voting behaviour, in order to determine if the level of polarisation today is cause for concern or a natural part of a healthy democracy.

Digital Democracy
I found the visualisations on specific MP profiles exceptionally informative, I would like to investigate further visualisations and gather other key data that would be useful to the UK general public before launching any digital democracy initiatives.

I was extremely pleased with the expressiveness of the voter profiles, especially picking out key events in Boris Johnson's later career.

The ‘holy-grail’ for accountability and transparency, would be a heat map of the UK showing to what extent politicians vote in line with their constituents. Only half the data necessary for this visualisation is available (representative data). There is currently no voting data available to measure a constituencies voter profile. All proxy data is fundamentally flawed, such as local news paper perspectives which might reflect public opinion.
Ultimately the best way to produce this graphic would be to build an app which stimulates public deliberation and collects the data over time. This would be a second stage ‘Deliberative’ digital democracy project, which would take years to create. It could build a user base by sharing stage one digitally democratic visualisations, demonstrated in this dissertation.

New concerns over data privacy, & interference in results by motivated parties, become important barriers to overcome.

LLMs
I see massive similarity between the advantages of Neural Networks for image processing, and the Large Language models for NLP.
Parallel conclusions can be drawn about the fragility of traditional processing techniques which work effectively only if the data is in an ideal form. Significant changes to image context for

image processing tasks, is the same as significant changes in sentence structure for NLP tasks. These changes break models quickly even if measures are taken to prevent overfitting.

While pre-processing measures helped to improve NLP results, pre-processing had an equally beneficial effect when leveraging GPT4. It is incredibly valuable to understand the fundamental ideas behind traditional approaches, if only to better interpret what the LLM is attempting to do, and optimise its response. This would include sentiment analysis, which GPT4 appears to be vastly more adept than any traditional NLP method.

Document summarisation could even be used to generate inputs to some of the traditional methods, although Simmons (2023) suggests caution here. In this study I did not have time to experiment by re-running the original clustering algorithms on the GPT generated data. I am confident however that engineering the prompt specifically for this task could vastly improve all of the methods, DBSCAN, K-means and Hierarchical clustering.

GPT4 was excellent at identifying clusters, although it struggled to actually conduct the clustering task. This makes sense given its architecture, but with clever prompt engineering even this task could get an improved response. By committing the conversation to memory, it became vastly more consistent. Attempting to use GPT3.5 via API meant generating a new conversation each time, which was the primary cause of its failure.
I would like to experiment with clustering further. Using GPT4 to both generate novel input data for clustering algorithms, and using the clusters it is able to identify as labels to run classification algorithms or machine learning.

It was excellent at classifying unlabelled text data, especially considering my lack of expertise on the subject matter. I am sure that a Neural Net could outperform GPT4 in most classification tasks, because of the inherent randomness in the language models design. However the complexity of the debate file, and the confusing division logic means that for this specific task GPT4 performed extremely well.

GPT4 was able to summarise, classify, and explain its decision, on 17 years worth of parliamentary debate. This is a 12 month project, produced in a single working week. I am very excited to see these models become more developer friendly and expand their feature list, because the optimal run time with a human out of the loop is 16 hours. Even this could be reduced, because the models don’t run on the laptop hardware. Stacking the tasks, and running the prompts in parallel could bring the task down to a single prompt time, which took 30 seconds to produce.

Final remarks
I want to take this opportunity to end the dissertation with an open call to stakeholders in the government's transparency initiatives to come together. The individuals entering the data, publishing the data, and using the data are very different groups. Because of this the data is inconsistent, difficult to access, and poorly formatted for their final use cases.

From what I have read, members of parliament are open to improving this process, but we need to collectively agree on what formatting changes and additional information is required before they can help. This will require an openness to collaboration from inside the government, and external communities alike.

References

ALTMAN, S. 2023. Sam Altman: OpenAI CEO on GPT-4, ChatGPT, and the Future of AI | Lex Fridman Podcast #367. Lex Friedman [Online]. Available at: https://www.youtube.com/watch?v=L_Guz73e6fw [Date Accessed: March 2023]

ANTTIROIKO, A.-V. 2003. Building strong E-democracy - The role of technology in developing democracy for the information age. [Online]. Available at: https://www.researchgate.net/publication/220422679_Building_strong_E-democracy_-_The_role_of_technology_in_developing_democracy_for_the_information_age [Date Accessed: March 2023]

BENHABIB, S. 1998. Democracy and identity: In search of the civic polity. 24. [Online]. Available at: https://journals.sagepub.com/doi/abs/10.1177/019145379802400207?journalCode=pscb [Date Accessed: March 2023]

BARBER, B. 2003. STRONG DEMOCRACY Participatory Politics for a New Age [Online] Available at: http://ndl.ethernet.edu.et/bitstream/123456789/21022/1/3.pdf [Accessed March 2023].

BITTLE, S., HALLER, C. & KADLEC, A. 2009. Promising Practices in Online Engagement [Online]. Available at: https://publicpolicy.pepperdine.edu/davenport-institute/content/pa_cape_paper3_promising_mech2.pdf [Accessed March 2023].

BRATTBERG, E. & MAURER, T. 2018. Russian Election Interference: Europe's Counter To Fake News And Cyber Attacks [Online] Available at: https://carnegieendowment.org/files/CP_333_Brattberg_Maurer_Russia_Elections_Interference_Brief_FINAL.pdf [Accessed March 2023].

BERLINSKI, N., DOYLE, M., GUESS, A., LEVY, G., LYONS, B., MONTGOMERY, J., NYHAN, B. & REIFLER, J. 2021. The Effects of Unsubstantiated Claims of Voter Fraud on Confidence in Elections [Online]. Available at: https://www.cambridge.org/core/journals/journal-of-experimental-political-science/article/effects-of-unsubstantiated-claims-of-voter-fraud-on-trust-in-elections/9B4CE6DF2F573955071948B9F649DF7A [Accessed March 2023].

BENSON, T. 2023. GPT-4 - How does it work, and how do I build apps with it? - CS50 Tech Talk. CSP50. [Online]. Available at: https://www.youtube.com/watch?v=vw-KWfKwvTQ [Date Accessed: March 2023]

Barroso, G. 2023. All models are wrong, but some are useful [Online]. Available at: https://www.lacan.upc.edu/admoreWeb/2018/05/all-models-are-wrong-but-some-are-useful-george-e-p-box/ [Date Accessed: March 2023]

BROCKMAN, G. 2023. OpenAI's Greg Brockman: The Future of LLMs, Foundation & Generative Models (DALL·E 2 & GPT-3). [Online]. Available at: https://www.youtube.com/watch?v=Rp3A5q9L_bg&t=403s [Date Accessed: March 2023]
CHAPMAN, R. 2023. DEMOCRACY: Ancient vs Modern. Chapman, Ryan. [Online]. Available at: https://winstonchurchill.org/resources/quotes/the-worst-form-of-government/ [Date Accessed: March 2023]

CHURCHILL, W. 1901. ARMY REFORM. [Online]. Available at: https://winstonchurchill.org/resources/speeches/1901-1914-rising-star/army-reform/#:~:text=May%2013%2C%201901 [Date Accessed: March 2023]

CHURCHILL, W. 1947. THE WORST FORM OF GOVERNMENT [Online]. Available at: https://winstonchurchill.org/resources/quotes/the-worst-form-of-government/ [Date Accessed: March 2023]

CHAMANTH, V. 2023. Decoder-only Transformer model [Online] Available at: https://medium.com/data-driven-fiction/decoder-only-transformer-model-521ce97e47e2 [Accessed March 2023].

DEŽELAN, T. 2023. Young people'sparticipation in European democratic processes [Online]. Available at: https://www.europarl.europa.eu/RegData/etudes/STUD/2023/745820/IPOL_STU(2023)745820_EN.pdf [Accessed March 2023]

UK Government, 2015. Open Up! Report of the Speaker's Commission on Digital Democracy. [Online]. Available at: https://www.digitaldemocracy.parliament.uk/documents/Open-Up-Digital-Democracy-Report.pdf [Date Accessed: March 2023]

DILMEGANI, C. 2023. GPT4: In-depth Guide in 2023. AI Multiple [Online]. Available at: https://research.aimultiple.com/gpt4/ [Date Accessed: March 2023]

EDWARDS, A. & KOOL , D. 2016. Digital Democracy: Opportunities and Dilemmas [Online]. Available at: https://www.researchgate.net/publication/306380064_Digital_Democracy_Opportunities_and_Dilemmas_What_could_digital_citizen_involvement_mean_for_the_Dutch_parliament_Preferred_citation_title_Edwards_AR_D_de_Kool_Kansen_en_Dilemma%27s_van_digitale_democra [Accessed March 2023]

GOOGLE 2023. Introduction to large language models. [Online]. Available at: https://www.youtube.com/watch?v=zizonToFXDs. [Date Accessed: March 2023]

HOUSE OF COMMONS. SELECT COMMITTEE ON MODERNISATION OF THE HOUSE OF, C. 2004. Connecting parliament with the public, London, Stationery Office. [Online]. Available at: *https://publications.parliament.uk/pa/cm200304/cmselect/cmmodern/368/368.pdf* [Date Accessed: March 2023]

HILBERT, M. 2007. DIGITAL PROCESSES AND DIGITAL PROCESSES AND DEMOCRATIC TH DEMOCRATIC THEORY:
Dynamics, risks and opport Dynamics, risks and opportunities that nities that arise when democratic instit arise when democratic institutions meet tions meet digital information and communication digital information and communication technologies nologies. [Online]. Available at: *https://www.martinhilbert.net/DigitalDemocracy-eBook.pdf* [Date Accessed: March 2023]

HILBERT, M. 2009. The Maturing Concept of e-democracy: From e-Voting and Online Consultations, to Democratic Value Out of Jumbled Online Chatter. Journal of Information Technology & Politics. [Online]. Available at:

HUETLIN, J. 2016. The Rise and Fall of the Pirate Party. [Online]. Available at: https://newrepublic.com/article/137305/rise-fall-pirate-party [Date Accessed: March 2023]

HEEMSBERGEN, L. 2021. Proto-Institutions to Open Government: (In)forming Publics with the Transparency We Deserve [Online]. Available at: https://www.emerald.com/insight/content/doi/10.1108/978-1-80043-762-320211006/full/html?skipTracking=true [Accessed March 2023].

ITZHAK, I., STANOVSKY, G., ROSENFELD, N. & BELINKOV, Y. 2023. Instructed to Bias: Instruction-Tuned Language Models Exhibit Emergent Cognitive Bias [Online] Available at: https://arxiv.org/abs/2308.00225 [Accessed March 2023].

KIRSCHNER, P., SHUM, S. & CARR, C. 2005. Visualizing Argumentation: Software tools for collaborative and educational sense-making [Online] Available at: https://www.researchgate.net/publication/42796168_Visualizing_Argumentation_Software_tools_for_collaborative_and_educational_sense-making [Accessed March 2023].

LUBBAD, M. 2023. GPT-4 Parameters: Unlimited guide NLP's Game-Changer. Medium [Online]. Available at: https://medium.com/@mlubbad/the-ultimate-guide-to-gpt-4-parameters-everything-you-need-to-know-about-nlps-game-changer-109b8767855a#:~:text=GPT%2D4%20is%20the%20latest,2%20has%201.5%20billion%20parameters [Date Accessed: March 2023]

LEE, C.-P., CHANG, K. & BERRY, F. 2011. Testing the Development and Diffusion of E-Government and E-Democracy: A Global Perspective [Online]. Available at: https://www.jstor.org/stable/23017501?saml_data=eyJzYW1sVG9rZW4iOiI3NzA2MjY5Mi1jOGZmLTRlYzAtYWUyYi1jMTU3ZDNiODNiYzYiLCJpbnN0aXR1dGlvbklkcyI6WyIxN2VlMTcxNC1hNWEyLTRlOTUtYWY1ZS0zMjcyZGNkMDZjMjMiXX0 [Accessed March 2023].

MAHRER, H. & KRIMMER, R. 2005. Towards the Enhancement of e-Democracy: Identifying the Notion of the 'Middleman Paradox [Online] Available at: https://www.researchgate.net/publication/220356731_Towards_the_Enhancement_of_e-Democracy_Identifying_the_Notion_of_the_'Middleman_Paradox [Accessed March 2023].

MUELLER, R. 2019. Report On The Investigation Into Russian Interference In The 2016 Presidential Election [Online] Available at: https://copblaster.com/uploads/files/mueller-report_compressed.pdf [Accessed March 2023].

OECD, ORGANISATION FOR ECONOMIC, C.-O., DEVELOPMENT, ORGANISATION FOR ECONOMIC, C.-O. & DEVELOPMENT 2020. Innovative citizen participation and new democratic institutions : catching the deliberative wave, Paris, OECD Publishing. [Online]. Available at: *https://www.oecd.org/gov/innovative-citizen-participation-and-new-democratic-institutions-339306da-en.htm* [Date Accessed: March 2023]

OPENAI. 2022. Introducing ChatGPT. [Online]. Available at: https://openai.com/blog/chatgpt [Date Accessed: March 2023]

OPENAI 2023. GPT-4 Technical Report. [Online]. Available at: https://cdn.openai.com/papers/gpt-4.pdf [Date Accessed: March 2023]

PAUTZ, H. 2009. The Internet, Political Participation and Election Turnout: A Case Study of Germany's www.abgeordnetenwatch.de [Online] Available at: https://www-jstor-org.uoelibrary.idm.oclc.org/stable/23744507 [Accessed March 2023].

PAGE, S. E. 2017. The diversity bonus : how great teams pay off in the knowledge economy [Online]. Prince Frederick: Recorded Books, Inc. Available at: https://www.jstor.org/stable/j.ctvc77c0h [Accessed March 2023].

PRAKASH, A. 2023. GPT-4 early impressions and how it compares to GPT-3.5 [Online] Available at: https://www.thoughtspot.com/data-trends/ai/gpt-4-vs-gpt-3-5 [Accessed March 2023].

RADFORD, A., JOZEFOWICZ, R. & SUTSKEVER, I. 2017. Learning to Generate Reviews and Discovering Sentiment. [Online]. Available at: *https://www.martinhilbert.net/e-democracyHilbertJITP.pdf*

https://winstonchurchill.org/resources/speeches/1901-1914-rising-star/army-reform/#:~:text=May%2013%2C%201901 [Date Accessed: March 2023]

DALIO, R. 2021. Principles for dealing with the changing world order : why nations succeed and fail, New York, NY ; London, Avid Reader Press. [Online] Available at: https://books.google.co.uk/books?hl=en&lr=&id=fljfDwAAQBAJ&oi=fnd&pg=PT5&dq=DALIO,+R.+2021.+Principles+for+dealing+with+the+changing+world+order+:+why+nations+succeed+and+fail,+New+York,+NY+%3B+London,+Avid+Reader+Press.&ots=zD3Ujfr7pm&sig=LF4OmrnG1mRgoI-tyuWmzp8eFtQ&redir_esc=y#v=onepage&q&f=false [Date Accessed: March 2023

ŞENDAĞ, S. 2010. Pre-service teachers' perceptions about e-democracy: A case in Turkey [Online]. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0360131510002046 [Accessed March 2023].

SIMON, J., BASS, T. & BOELMAN, V. 2017. Digital Democracy
A summary report for practitioners [Online]. Available at: https://ecas.issuelab.org/resources/29642/29642.pdf [Accessed March 2023]

SPAKOVSKY, H. 2015. The Dangers of Internet Voting. BACKGROUNDER: The Heritage Foundation. [Online]. Available at: *https://www.sos.state.co.us/pubs/rule_making/written_comments/2015/20150714Spakovsky.pdf* [Date Accessed: March 2023]

SUNSTEIN, C. R. & EBOOK CENTRAL ALL SUBSCRIBED, T. 2006. Infotopia : how many minds produce knowledge, Oxford ; New York, Oxford University Press. [Online]. Available at: https://www.audible.co.uk/pd/Infotopia-Audiobook/B004EVF56K?qid=1692271955&sr=1-1&ref=a_search_c3_lProduct_1_1&pf_rd_p=c6e316b8-14da-418d-8f91-b3cad83c5183&pf_rd_r=EMJJ [Date Accessed: March 2023]

SUROWIECKI, J. 2005. The wisdom of crowds, New York, Anchor. [Online]. Available at: https://www.audible.co.uk/pd/The-Wisdom-of-Crowds-Audiobook/B004FTKOAI?qid=1692272058&sr=1-2&ref=a_search_c3_lProduct_1_2&pf_rd_p=c6e316b8-14da-418d-8f91-b3cad83c5183&pf_rd_r=ZAP4N9R65Z6BNMW75C38&pageLoadId=HxoyekcM2DkDT7ks&creativeId=41e85e98-10b8-40e2-907d-6b663f04a42d [Date Accessed: March 2023]

SHEVCHUK, V. 2023. GPT-4 Parameters Explained: Everything You Need to Know. Medium [Online]. Available at: https://levelup.gitconnected.com/gpt-4-parameters-explained-everything-you-need-to-know-e210c20576ca [Date Accessed: March 2023]

SIMMONS, A. & VASA, R. 2023. Garbage in, garbage out: Zero-shot detection of crime using Large Language Models. [Online]. Available at: https://arxiv.org/pdf/2307.06844.pdf [Date Accessed: March 2023]

TSAGAROUSIANOU, R., TAMBINI, D. & BRYAN, C. 1998. Cyberdemocracy Technology, Cities and Civic Networks [Online] Available at: https://www.routledge.com/Cyberdemocracy-Technology-Cities-and-Civic-Networks/Bryan-Tambini-Tsagarousianou/p/book/9780415171359 [Accessed March 2023].

The Economist Intelligence Unit, 2016. Democracy Index 2015 Democracy in an age of anxiety [Online]. Available: https://dataspace.princeton.edu/bitstream/88435/dsp017p88ck01w/1 [Accessed March 2023].

The Democratic Society, 2023. Why. [Online] Available at: *https://www.demsoc.org/theory-of-change* [Date Accessed: March 2023]

UBEROI, E. & JOHNSTON, N. 2021. Political disengagement in the UK: who is disengaged?, UK, House of Commons Library. [Online]. Available at: *https://researchbriefings.files.parliament.uk/documents/CBP-7501/CBP-7501.pdf* [Date Accessed: March 2023]

VEDEL, T. 2006. The Idea of Electronic Democracy: Origins, Visions and Questions [Online] Available: https://academic.oup.com/pa/article-abstract/59/2/226/1555923 [Accessed March 2023].

World Bank. 2021. Military expenditure (current USD) - United States, Russian Federation, Germany, United Kingdom, China [Online]. Available: https://data.worldbank.org/indicator/MS.MIL.XPND.CD?end=2021&locations=US-RU-DE-GB-CN&start=1960 [Accessed March 2023].

World Bank. 2021. Military expenditure (current USD) - United States, Russian Federation, Germany, United Kingdom, China [Online]. Available: https://data.worldbank.org/indicator/MS.MIL.XPND.CD?end=2021&locations=CN-US-RU-DE-GB&start=1963 [Accessed March 2023].

WEISS, D., WOLMER, J. & VATSA, A. 2022. Blockchain-based Electronic Voting System for Modern Democracy: A Review. IEEE Integrated STEM Education Conference Princeton. [Online]. Available at: https://ieeexplore.ieee.org/document/10025096 [Date Accessed: March 2023]

WILSON, E. 2023. The real problem of humanity is the following: We have Paleolithic emotions, medieval institutions and godlike technology. And it is terrifically dangerous, and it is now approaching a point of crisis overall. [Online]. Available at: https://www.goodreads.com/quotes/9770741-the-real-problem-of-humanity-is-the-following-we-have [Date Accessed: March 2023]

WOLFERS, J., ZITZEWITZ, E. & NATIONAL BUREAU OF ECONOMIC, R. 2004. Prediction markets, Cambridge, Mass, National Bureau of Economic Research. [Online]. Available at: *https://pubs.aeaweb.org/doi/pdfplus/10.1257/0895330041371321* [Date Accessed: March 2023]