The Journal 2020

From the Headmaster

“Rhetoric is no substitute for reality”, so said American economist and social commentator Thomas Sowell. Our School Value of Scholarship expounds the importance of intellectual curiosity, independence, creativity and habits of learning. And yet these qualities are by their very nature hard to pin down, hard to measure. Although we are justifiably proud of our philosophy of scholarship for all, how do we assess to what extent this love of learning, this desire to challenge and question are embedded within the culture of the RGS?

The context of the last few months has seen schools reinvent so much of the traditional approach to teaching and learning: rarely has the old-school chalk-and-talk mindset seemed such an anachronism. Physical isolation, remote learning, virtual lessons all by their very nature and phraseology place greater focus on individual’s independence and own innate sense of scholarship. Indeed, it is often said that at times of great adversity, true creativity and innovation come to the fore. I am delighted that this edition of The Journal illustrates that this has very much been the case throughout the RGS community. It is not only the sheer maturity, complexity of thought and conviction of these submissions which impress; equally, it is the energy, enthusiasm and passion which shine through each and every piece of work. Our students, having had a seed of inspiration sown, have put into practice all those elements of scholarship which allow quite extraordinary results to be realised.

I would like to take this opportunity to congratulate my Head of Scholarship, Mr Bradford, and all those students who have contributed to The Journal. I hope that all who read it are inspired and impressed in equal measure and appreciate that far from rhetoric true scholarship is certainly a reality at the RGS.

Dr Jon Cox
Headmaster

Editorial

Institutions across the world are redesigning curricula for the 21st century. Their aim, in an age of increasing technology, is to inculcate habits such as independence, imagination, collaboration and perseverance so that far from being obsolete, our students are ready to embrace the challenges the future will certainly bring. I am proud to say that here at the RGS our Independent Learning Assignment has, since its inception 12 years ago, been consistently fostering these dispositions within our students. Every year both the quality and variety of academic work produced is extraordinary and this year was no different. Published titles herein include an analysis of the biology of senescence, an argument over the validity of free-will, an analysis of the confederalist government of northern Syria and even a guide on how to build a jet engine in your shed complete with a working prototype. This cohort undertook some of the most ambitious projects to date and the results are very much worth the read.

Mr CS Bradford
Head of Scholarship

Forward

Whilst The RGS supports scholarship and celebrates intellectual achievement throughout the year, the pinnacle of academic diligence is undoubtedly the Independent Learning Assignment (ILA). Culminating in the Presentation Evening, the ILA is an extended research project, undertaken primarily in the break between Lower Sixth and Upper Sixth. Each student is given almost complete free reign in deciding their specialist topic, allowing the full range of students' niche interests to be given a spotlight. Every department then examines their assigned students' essays, narrowing down nearly 150 projects to the very best 11.

The Presentation Evening is an interdisciplinary affair in the truest of senses: following weeks of preparation, the students give five-minute speeches on their specialism, allowing for discussions of Chinese geography to blend seamlessly into analyses of sharks' lifespans and post-colonial literary theory.

The work undertaken not just by those shortlisted, but also by every sixth form student in creating an independent and impressive piece of work, cannot be understated. Outside of the eleven finalists, a further number of boys were awarded this year for their efforts in some of the most complex and impressive scholarship ever undertaken at the RGS. These included Connor Rajan's CREST Project understood only by a sole member of the Mathematics Department, Utkarsh Dandanayak's published research on financial technology in Africa, or Samuel Cherry's exploration of the biology of religious experiences.

I hope that you enjoy reading the projects as much as the students enjoyed writing them (even if you still don't understand how a 17 year-old built a jet engine in their shed). The ILA is truly a labour of love, and no-one has put more time and passion into the annual scheme than Mr. Bradford. On behalf of the Sixth Form, the finalists, and the Senior Scholars' Council, I would like to thank him for his persistent efforts to promote and celebrate scholarship in all levels of the school.

Alfie Cherry
Chair of the Senior Scholars' Council 2019

Contents

All references to appendices have been redacted from this publication but are available on request.

"I want a wife": a study into the value of unpaid household labour in the economy

Cameron Gardner

ABSTRACT

"Can't afford to pay your housekeeper anymore? Marry her! Then she'll do it for free."[1] The tragedy behind the joke is that for years mainstream economics has accepted this viewpoint. Today however this is being challenged by the field of feminist economics. Since 1953 the United Nations has issued guidelines on what must be included in Gross Domestic Product (GDP) statistics. This system is founded on a purely market based approach, and for the last 50 years has been the primary way in which we calculate both economic and social value. In 1999, however, the New Zealand politician Marilyn Waring published her seminal work 'If Women Counted'. This book - for the first time in the development of western economics - attempted to address the importance of unpaid work in enabling our economy. Today this issue has grown in importance. Called the "12 trillion-dollar question" by the consultancy giant McKinsey[2], the resolution to the issue could change the way we view our society. Economic value and social value have become intrinsically linked in our modern viewpoint. As such, a change in economic value will lead to a change in social value that could alter the way women, in particular, are treated around the world.

BACKGROUND TO THE ECONOMICS OF UNPAID LABOUR

Since the conceptual idea of an "economy", people have attempted to find a value for its size. In his 1920s work 'The Economics of Welfare' A.C.Pigou laid out the framework for what has become the theory used to calculate Gross Domestic Product (GDP). Pigou stated that national income was 'everything that people buy with money income, together with the services that a man obtains from a house owned and inhabited by himself'[3]. Since then, GDP has dominated modern economic thinking. However, it fails to consider unpaid work done: work for which there is no monetary transfer. This is normally work done inside the house that would be done by hired help if not done as unpaid work. Statistically this has been done by women, with the UN figures estimating that as much as 75% of all domestic work is done by women. The result of GDP ignoring this work means that not only is the societal value of work done by women largely ignored, but economic policy makers have not been able to see the true scale and nature of household output. This economic failure leads to a major societal failure in that women in particular are undervalued. This perpetuates the stereotyping of gender roles and the value they bring. Unless we reconsider the work done in the unpaid sector of the economy, economics will both fail to accurately measure the size of the economy and society will fail to value the work done by those who predominately do it: women.

The issue of unpaid work has been raised before. Phyllis Deane, a female employee of two British economists, Meade and Stone (who would later oversee the introduction of the 1953 United Nations standards for GDP) argued as early as 1938, while looking at economic output in a number of British colonies that, as such a large proportion of these countries output was driven by the household, it would be 'illogical' to exclude these components from GDP. Deane said that as so much time, particularly for women, was dedicated to activities such as collecting firewood or cooking, a government could not have a full picture of output unless it was included. Deane argued that these activities were being excluded because they were principally the work of women. The failure of the government of the day to not measure this data has meant that to this day the work done is not properly counted and recognised.

In her book, 'If Women Counted', Marilyn Waring drew attention to the fact that there was no recognition given to goods or services that were not sold for a monetary value. Going even further than Deane, Waring argued that if true societal change was to be possible there would have to be an economic one as well. The French economist Christine Delphy took this even further. She split the economy up into two modes of production: the industrial mode and the family mode. Delphy argued that the industrial mode was dominated by men and the family mode was dominated by women. However, only the industrial mode is recognised as work done. It is this failure to recognise the work done that lets the oppression of women continue.

The overall aim of my own primary research into the field of unpaid work is to look at the ignored household output done by the top 1% of households by income. Leading from Delphy's research into modes of production, in many households of the top 1% there appears to be one principle wage earner. Recent tabloid newspaper headlines have focused on large divorce settlements awarded to stay at home parents, labelling them as unfair or anti work. Within this backdrop it is important to look at how the work done inside the family home should be valued at using monetary terms. The old adage states 'behind every great man is an even greater woman': it must be considered whether in households with a high income it is necessary for one parent to carry out a large amount of unpaid work in order for the other partner to generate that income. The Italian philosopher Silvia Federici in particular has argued that the economic production we measure in GDP could not take place without 'non-economic' production. As such I wished to find a way of valuing this work carried out to give it the recognition it deserves in enabling our society to function.

PREVIOUS STUDIES IN CALCULATING THE VALUE OF UNPAID LABOUR

One of the major issues with measuring unpaid work is it is not obvious when it takes place. Furthermore, what is exceptionally difficult is to place a value on the work done. Both Waring and Deane suggested that work should be measured overall in time taken to complete. Waring has stated that time is 'the one investment we all have to make'. The principle idea behind economics is the basic economic problem: 'resources are scarce but wants are infinite'. Time is arguably the scarcest resource of all. Economic agents are forced to make choices on how they use their time, more so than how they spend their income. The way a person allocates their time influences their thinking more than how they spend their income.[4 5] Waring argued that time put into activities was the only fair way of estimating household inputs. By neoclassical standards, when no economic transaction has taken place during the production of unpaid labour the resource that is being given up is that of time.

As such the previous empirical attempts at valuing unpaid labour in the household have all focused on time usage. The Australian economist Duncan Ironmonger used time to calculate Gross Household Product (GHP) which he defined as 'the productive activities conducted by households using household capital and the unpaid labour of their own members to process goods and provide services for their own use.'[6] Principally, however, attempts to value unpaid labour have been carried out by the Bureau for Economic Analysis (BEA) which has published the Multinational Time Use Survey (MTUS) and American Time Use Survey (ATUS). The ATUS was first produced in 2004. The ATUS surveys 15,000-20,000 household and splits data into 7 categories: household production: housework, cooking, odd jobs, gardening, shopping, child care, and domestic travel. The BEA then uses this data to calculate a market value for the unpaid labour done. They have found that as household work has a low wage dispersion, by which there is not a large difference between the amounts charged by different people in the industry, the higher paid earners only earn a little more than the lower paid workers. Ironmonger has attempted to calculate the Gross Household Product using this data. Ironmonger used the BEA data to give a value for the inputs in hours, and then the market rate as the value of outputs. The total value of the outputs per household is then aggregated to form an estimate of GHP. Ironmonger then combined GHP with GDP to find a value for Gross Economic Product: the total produced by an economy when both the household and the markets sectors are included.

However, an alternative method to GHP to calculate the additional output produced is the opportunity cost method. This method looks at market value for the unpaid labour if the labour were in that person's chosen field: for example the salary a barrister would demand for an hour of their time if they did not choose to use to use that time for unpaid labour. This method of calculation requires far more data than that used by the MTUS and ATUS, so may pose more difficult on a larger scale.

Taking the theoretical background laid out by Deane as time being the most crucial resource that is devoted to any economic choice, the opportunity cost method is the most accurate representation of the value that unpaid work provides to society. The GHP product gives some value for the time spent carrying out unpaid work, but it does not show a truly representative value. It assumes uniformity across the spread of people's time, and that each labour hour is the same irrespective of background or household circumstances. To some extent this works in that we all are required to carry out the same unpaid work such as looking after children. However, this fails to recognise what people give up to carry out unpaid work. Different people have different constraints. When looking at GDP we do not assume everyone is equal and split up their time into 7 sectors and assign a monetary value per output for each one. Instead every person has a unique addition to GDP, in part based on their background, and work. This must also be done when looking at household product.

MY PRIMARY RESEARCH

Of the two methods outlined above I chose to use the opportunity cost method of valuing the additional output, based on the fact that it is a more theoretically accurate method of valuing the time given up. In order to gather data, I used a similar survey to that used in previous methods but then fed the data into an equation for opportunity cost.

Overall my results found that the value of the unpaid work done matched that of the paid work. In the UK the average household income required to be in the top 1% is approximately £166,000 for a household with two adults and two children. However, I found the average value of the total output (including both paid and unpaid work) done to be £335,496 per household that I measured. This would mean that the value of the two modes of production are nearly equal: total work is nearly double that of paid work. This means that the two adults of a household equal each other in total household output, whether it be predominately unpaid or predominately paid.

Total Value of Respondents Paid & Unpaid Annual Labour

Respondent Number	Paid and Unpaid work values		Respondent Number	Paid and Unpaid work values
1	£296,860		31	£96,570
2	161,130		32	£46,260
3	£219,786		33	£522,966
4	£339,215		34	£469,586
5	£422,782		35	£0
6	£165,440		36	£0
7	£266,352		37	£235,328
8	£110,760		38	£0
9	£174,756		39	£261,675
10	£136,180		40	£172,116
11	£255,200		41	£192,943
12	£118,536		42	£136,970
13	£154,619		43	£229,976
14	£39,121		44	£144,942
15	£128,409		45	£176,571
16	£0		46	£248,609
17	£51,694		47	£174,280
18	£133,216		48	£169,098
19	£122,169		49	£212,571
20	£155,067		50	£199,113
21	£860,900		51	£253,522
22	£76,352		52	£142,196
24	£210,811		53	£360,517
25	£58,339		54	£223,007
26	£38,730		55	£105,555
27	£56,339			
28	£97,144			
29	£126,703			
30	£88,139			

SURVEY DESIGN

In order to collect primary research data, it was necessary to design a time use survey like that used by the BEA. For the base categories of unpaid work I would measure, the Office for National Statistics (ONS) has GHP calculators that using the 3rd party method of valuation, split unpaid domestic work into 7 categories each with its own hourly wage: cooking, cleaning, child care, adult care, laundry, transport and volunteering. As such at a minimum level the survey would have to find estimates for weekly time that was dedicated to each activity, which could then be built into a look at how much time was spent overall. The data for the time use survey could then be directly fed into this input calculator to give a value for GHP.

To be able to use the opportunity cost method it was necessary to gather additional data as it depends if the subject is currently in full employment. For people who have remained in full employment their hourly wage can be calculated by looking at their annual salary. However, for people who have chosen to exit full time employment, often for unpaid work, it is much harder to calculate opportunity cost. The major factors that determine pay per hour are job sector, number of years of experience, education level and gender. As such it was necessary for the survey to collect data on each of those factors. This data can then either be used to compare to the salary earned by people still in employment or inputted into an equation to estimate a market wage value.

A further 3 questions related to data collection on the people who were replying to my survey. In order to make meaningful conclusions from the data I also needed to understand how much domestic work they actually paid for so that it could be taken into consideration when considering how much domestic work takes place both paid and unpaid. Finally, it was important to see how many respondents had children in independent schools. This could also be used alongside with the household income data to measure whether the respondents were in the top 1% of households if the respondent had chosen not to say the size of their household income. 630,000 children across the UK are privately educated. Assuming households in the top 1% are willing to privately educate their children, it may be assumed that the additional disposable income is spent on education.

CALCULATIONS

I used the following calculation for opportunity cost: (average sector wage)*(gender index)*(age index)*(education index)*(number of hours worked unpaid a week)*52

To use this calculation I had to gather information on estimated pay levels. The starting point for this was the average salary per sector. This wage value then would have to be modified using a gender, education and age index. An estimate can be provided for gender by using the average wages overall for a man combined to an average wage for a woman. In the United States for one dollar earned by a man the average woman earns 79 cents. According to research carried out by Pay Scale the difference between the earnings of men and women increases with age. As such my estimate for gender's effect on pay would have to be graduated based on each age band. 79 cents would be used to represent one dollar of male income in the 35-44 age range. For each age-band below that, the gap would close by 2 cents and for each age-band older, the gap would increase by 2 cents.[7]

Further to the difference that age makes on the gender pay gap, age is itself a factor in average earnings. The ONS issues the median taxable income for each age bracket. By taking the median value of taxable income in the 35-44 age band to have an index value of 1, an index value can be calculated for each age band: by dividing the median income for that age band against the median band.

Finally, the major determining factor was education level. The US government releases data on the unemployment rate and average median salary for different education levels. Taking a bachelor's degree to have an average index value of 1, the value can be manipulated to give an increase or decrease for each education band in the same way that gender can be.

As such this gives a final equation for an estimate of the salary of the respondent. (average sector wage)*(gender index)*(age index)*(education index). The ONS release survey data on the average weekly wages of 14 different industries which can be used to provide a value for average sector wage. This value for the weekly salary can then be used to calculate the opportunity cost. The value for weekly salary can be divided by hours worked (or if the respondent is not in work the hours worked in a 5 day 0900-1700) to give the value of an hour's work. This can then be multiplied by the hours of unpaid work done a week to give the opportunity cost for a week. Finally, this value can be multiplied by 52 to give the value of opportunity cost for the year.

In order to the give the values for the opportunity cost of the unpaid work the formula relies on knowledge of both the hours worked in a week and the industry that the respondent works in. However, this poses a significant issue when the person is either unemployed or has chosen to reduce their hours in order to meet the demands of unpaid work. This originally would mean that the unpaid work done by those unemployed would read as 0. In order for the formula to hold for those who are not in employment these two variables must be estimated. However, values for hourly wage can be assumed by using the index value for the person multiplied by the average personal income for that country. This value for hourly wage calculations can then be multiplied by the number of total hours done of unpaid work in order to give the opportunity cost for a year. As such a value can be calculated irrespective of whether the person is in employment or not. This is crucial. In some households one member of the household may choose to give up work or reduce their hours to carry out unpaid work.

There are some assumptions that have had to be made in order to form the equation and the calculation. Firstly, the equation assumes that the three different index elements of the equation can be split up into separate components and are not all correlated. If it were possible it would have been better to have created the index values based on data for all the three categories. Rather than having to create three separate values based on average taxable income for age, gender and education categories, it would have been better to have had the taxable income for say a woman aged 45-54 with a bachelor's degree.

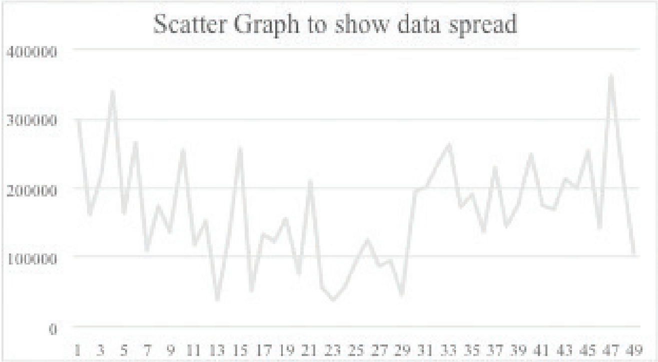
Furthermore, the calculations assume that it would be possible for the person not in employment to gain a job at their calculated 'market rate', with no barriers to re-enter the job market. While this may hold for those who choose to not be in work it does not hold for those who cannot find work. They have not chosen to exit employment, and such the formula may need further correction to take this into consideration. The opportunity cost will by necessity be lower, because this time cannot be spent in employment. It rather does not have an opportunity cost worth a salary but instead is free time. The time still has a value, as it could be spent be doing other activities but is not as valuable as if it could be spent working. As such this formula is far more accurate when looking at people who have chosen to reduce their paid workload than those who are unemployed.

The data was then inputted into the calculators to give three numerical values: the value based on the ONS calculator for GHP, the opportunity cost formula devised earlier and for those who were in employment an estimated value of their opportunity cost using salary data where available. This enables the opportunity cost method to be compared with Ironmonger's method using GHP. Furthermore, it enables comparison between the created formula and the opportunity cost for those who are in employment; we can see how accurate the formula is in comparison to actual recorded earnings.

WORKING EXAMPLE:

Respondent Number	4
Gender	Female
Age	45-54
RGS parent	No
Independent school parent	Yes
No. of children	2
Education level	Post graduate degree
Employment status	In full employment
No. of hours worked per week	40
Sector	Technology
Annual salary	£100,000+
Household income	£200,000+
No. of hours unpaid work per week	Oct - 15
No. of hours cleaning	6
No. of hours childcare	30
No. of hours unpaid work	5
No. of hours unpaid work	1.5
No. of hours unpaid work	6
No. of hours unpaid work	0
No. of hours unpaid work	0
No. of hours unpaid work	
Usual week?	No
Do you pay for domestic help?	Yes
How much?	2 hours every 2 weeks

To show the use of the formula I will take respondent number 4 as an example (see table above). Respondent number 4 is a woman aged between 45-54 who has a full-time job working in the technology sector after graduating with a post graduate degree. This gives her as a 45-54 woman a gender index value of 0.77, as well as an age index value of 2.53. These two values are then multiplied together to give value of 1.95 for both age and gender. There is then an educational level index value of 1.57 for her post graduate degree. All of this multiplies



to give an index value of 3.0615, which represents the effect on the average salary for this sector. The Technology sector pays an average salary of £964 per week, according to HMRC estimates, which would mean that respondent number 4 earns an estimated £2,951 per week to give an estimated yearly salary of £153,452. The respondent reports an estimated salary of £100,000+ which fits with this estimate. Respondent number 4 states she works 40 hours a week which means she is paid an estimated £77.74 per hour. She reported working 49 unpaid hours a week, meaning the opportunity cost of her unpaid work a week is valued at £3,614. This gives a total yearly value of unpaid work as £185,917.

When the two values of opportunity cost are compared overall using the entire dataset (one from the formula; one from the available salary data) it is apparent that in a number of cases the formula value is approximately between one to two times higher. However, in the data, 48% of the respondents to the survey reported earnings of over £100,000. This made it far harder to estimate opportunity cost based on salary, as the salary had to be assumed at £100,000. As household income data was also collected this can provide a further comparison. A 2005 US Census report[8] found that in the top 5% of households there was on average one income earner who earned 10 times the US median average and one income earner who earned 5 times the median income. This means that most households in the leading 1% tend to have a primary wage earner earning around 2/3 of household income. This means that for a respondent with personal income of over £100,000 and household income of over £200,000 it could be possible to assume a personal income of £133,000 or above. This means that it is likely that the formula will range from very similar to towards roughly double the salary value that is calculated.

I found the mean value of the dataset to be £182,205, with a standard deviation of £146,489. This standard deviation is to be expected, as there are some values for the data that are far higher than the rest, as well as some values that read as 0 (all of which are either in full-time education or unemployed and actively seeking work.) The value for the first quartile was £99,247 with the value for the third quartile being £228,233 giving an inter quartile range of £128,296. Looking at how this relates to the mean value the data is being skewed by outliers. As such I removed any values that were greater or less than the mean plus or minus one and a half times the interquartile range, as well as removing the 0 values. This meant I had to remove any value greater than £375,685, which was the four highest values. As a consequence, this reduced the average value to £167,748, as well as giving a lower quartile of £118,355 and an upper quartile of £219,786.

Overall what the data shows is that in the top 1% households the value of the opportunity cost is between 50% of the value of paid labour to 100% of the value of the paid labour. Some of the differential may be made up by using larger amounts of hired help to carry out some tasks which may have been done as unpaid work. In the UK a household income needs to be in excess of £160,000 to be considered in the top 1% of households. For respondents reporting personal income of greater than £100,000 there was an average opportunity cost value of £81,292. Numerous studies have shown that women do a larger percentage of unpaid work than men. This data set had roughly equal gender divide with a 51% male response, compared to a 48% female response (no gender specified accounted for 1%). Studies[9] show that unpaid labour is disproportionately undertaken by women in a household than men, in the same way that paid labour is also unequally shared within a household. As such, there may again be a two-thirds / one third split in unpaid work. In a two-parent household it can be assumed therefore that both paid and unpaid have an unequal distribution. However, the value of the proportion of work done by both members would be equal when considering both unpaid work and paid work. For example, one member could do 80% of unpaid work, 20% paid with the other carrying out 20% unpaid and 80% paid work. Because of imbalances of the spilt of how the work is done by different members of each household, when looking at household output what should be measured is the sum of both paid and unpaid work.

Therefore, we should look at the size of household output as opposed to personal output to compare the size of unpaid work done to the size of paid work done. To calculate individual personal output, we can sum the paid work done with the unpaid work. For my data set the mean value for all respondents for this is £167,748. Assuming my data set to represent some members of a household doing over half of all unpaid work and some less than half we can double the average value per person to give a total household value for all output produced. Overall, therefore, this gives an average value for my respondents of all of their household output – inclusive of both paid and unpaid work-to be £335,496. This means that GDP is failing to recognise additional household output by almost 100% of the recorded value.

CONSIDERATION OF TIME

Following the research of Waring and others the key component for measuring total household output was measuring time. Comparing the value of household output that I calculated to the value that they would be adding through GDP, it is clear that the household output value is nearly exactly double. The majority of respondents surveyed, therefore, are seeing nearly half of all their output going unmeasured. This means that the GDP is failing to capture all of what they are producing, and recognise its value. People all around the world have given up time for unpaid work; which if time is taken to have an intrinsic value means it must have value. This valuable output must be recognised by economic data.

Furthermore, the majority of respondents do not realise that they spend this much time carrying out unpaid work. One of the more unusual points raised by the data is that people tend to underestimate the amount of unpaid work that they carry out when thinking about unpaid work in total for a week compared to thinking about each task in isolation. The survey asked people to consider the total number of hours spent carrying out unpaid work in a week as well as the hours spent on each categories. In 67% of cases the number of summed hours from the 7 categories was higher than the total number they had worked out for the week. This would appear to show that people fail to grasp the true quantity of unpaid work that they actually carry out. This has two effects: firstly it means that when considering the amount of unpaid work done for imputation into a formula for opportunity cost it means that to increase accuracy it is necessary to add all the values for each individual category to give as the total time taken. Secondly it must be considered why respondents tend to view themselves as doing less work than they actually do. Leading back to research carried out by Waring it could be argued that we do not just underestimate the value of unpaid work but also the time we spend doing it. These two factors could be interlinked. When people consider the time they spend doing their job they consider the "9-5" for example: a fixed length of time for which they receive a salary or an hourly wage. Unpaid work does not have the same immediate feedback of what value you receive for time. The lack of a valuation for unpaid work means that people also fail to recognize the time they put towards it. Partly this may be because unpaid work is typically seen as something that is just carried out as a necessity rather than as a worthwhile activity. These two factors together mean that people do not view their unpaid work as valuable so they do not look at the amount of time they spend doing it and therefore do acknowledge its worth to society.

However, this societal viewpoint would be corrected if economic models took it into account. The unpaid work represents a major part of household output, enabling the society to function. Waring pointed to the basic economic problem arguing that time was a major scarce commodity. However, what is also true is that unpaid work can go towards 'fulfilling individual wants and needs'. The unpaid work households carry out enables the basic chores of life to be done; be it cooking or cleaning etc. If we did not carry out the unpaid work either we would all have to pay for these services to be done or they would not be done at all. This is obviously not practical. Households do not have the disposable income to pay for the thousands that this work would cost if done through the market. Typically, throughout history the responsibility for unpaid work has fallen on women. This has had crucial effects on society. In the same way the value created by this work has been ignored the value of those who have done it has also been ignored. Delphy has argued that as unpaid work is done more by women than by men, unless we look at its value in this way we also undervalue women. However, the value of this work has been shown to equal that of the paid work.

IMPLICATIONS & CONCLUSIONS

It would significantly affect public perception of this type of work if people realised that their unpaid work had as much value as it does. Comparing the total household output to the GDP figure created by the household changes the way the household should prioritize its resources. The value that is being measured is the time spent on the activity, not its fiscal remuneration. The work which is typically done as unpaid work is crucial for the functioning of most households, yet is seen as lacking in importance. If this valuation were to be used then the time taken for these tasks would be considered in equal importance, as an equally worthwhile use of the key resource of time. This has far reaching consequences. Going back to the research carried out by Deane, she argued that failure to use a method of output that calculates and includes the value of unpaid work is a failure to truly value the work done by women. Tragically, over the past 80 years very little has changed to challenge the fears she raised, yet the fears she raised are today as equally valid. The enduring newspaper publicity over divorces seeing 50% split of total assets to the lower wage earner continues to rage. However, this is justified in the time given up by the lower wage earner to carry out unpaid work. Looking back to the idea that the idea of 'behind every great man is an even greater woman' this quote is true in sentiment if not accurate in its gender labels. It is clear that paid work only makes up half of all household output. For one member of the household to devote their time to paid work means that one must carry out all the unpaid work. Overall their time and output would be roughly equal, whether they devote it to paid work or to unpaid work.

This conclusion would have a major effect on public policy. Politicians look at the GDP statistics under a microscope. A focus on household product would provide a more well-rounded viewpoint on not just the economy but on society. We use GDP to look at both national output as well as living standards per capita. However, as shown these figures only capture half of all the produced output. At the beginning of her work Waring argued that a refusal to look at unpaid work meant that politicians refused to consider policy that benefitted those who do the most of it: namely women. If politicians and policy makers were to pursue policy that aimed to increase GDP at the expense of unpaid work or to reward those who carried out paid work at the expense of those in unpaid work then those policies would have a negligible effect or even negative effect on total output. Instead, politicians should look at the combined product of unpaid work and paid work. This would achieve a number of things: they would enact policies that promote growth in both areas rather than just one.

In summary, over the past 80 years the study of GDP has seen economic value become intertwined with social value. However, in my data set households have seen almost half of all their output ignored by economic policy makers and politicians. This omission from policy makers has trickled down through society and led to a societal devaluation of this work, and the people who are more likely to carry it out: women. In order to achieve not only a holistic assessment of national output but also to achieve better gender equality, it is imperative that we look at the total household output inclusive of both paid and unpaid labour as the true contribution to the economy.

"Difficult but better than Pol Pot"[1] An assessment of Cambodia's recovery after the Khmer Rouge genocide

Charlie Kershaw

INTRODUCTION

Writing about the recovery of Cambodia from such a horrific period led by the Khmer Rouge poses something of a considerable challenge. Not only is the extent of recovery subjective but the wide-ranging areas of Cambodian society affected by the genocide means that there is a large variety of factors that must be analysed and evaluated in order to present the overall picture of Cambodian recovery. Nevertheless, I will attempt to meet such a challenge and have done so using two methods. The majority of this paper is written using secondary research, from books and numerous websites, and is then supported and occasionally challenged by my findings when I spent a month in the country in July 2019, 40 years after the fall of the Khmer Rouge. In many areas we will see recovery leading to progress but equally as visible are the areas of destruction and suffering that have improved little in the ensuing decades.

This paper is broken down into different aspects of Cambodian society to allow an analysis of each potential recovery in turn before pulling it all together at the end. I will first consider the impact of the Khmer Rouge and then move onto how the population numbers have improved before examining the extent of political recovery, a section which has numerous different components to consider. Next to be examined is the economic situation in Cambodia, an issue which underpins much of what is discussed in this essay, before moving onto the psychological and finally the social recovery. Each area has recovered to different extents and therefore I will conclude that Cambodia has only partly recovered from the genocide of Khmer Rouge regime. I will also suggest a potential reason for this and what I think needs to change for there to be a complete recovery, if such a thing is possible.

SECTION 1- THE IMPACT OF THE KHMER ROUGE

Year zero, the name applied to the beginning of Khmer Rouge control in Cambodia, provides a perfect representation of what four years of extreme and brutal communism inflicted on the country. The Khmer Rouge saw it as the beginning of a great new chapter in Cambodian history, but a combination of disastrous economic policies and extraordinary physical suffering meant that Cambodia suffered four years of extraordinary hardship which would endure from generation to generation, hampering the lives of millions of its citizens for decades to come.

Cambodia had previously been under French rule until 1953 and since then Prince Norodom Sihanouk had been their 'de facto' leader. Having won the civil war which started in 1970, the Khmer Rouge seized power in Phnom Penh on April 17th, 1975 (figure 1) and they immediately set about imposing their dictatorial communism on the country - something of an ideological paradox. There had always been a communist presence in Cambodia but any opposition that they posed had never been serious. It was not until Sihanouk declared his support for them during the civil war, potentially without knowledge of their true goals, that they began to gain popular support in the countryside. The new Khmer Rouge leaders forced hundreds of thousands of people to leave the cities and march into the countryside, commanding them to work on the land. Their ideology was an extreme form of communism in which they wanted to 'restart' the country, hence the name year zero, and bring about their communist utopia. It was to be an agrarian based society, hence the removal of all people from the cities, meaning that there would be no paid jobs. But for many the worst was still to come.

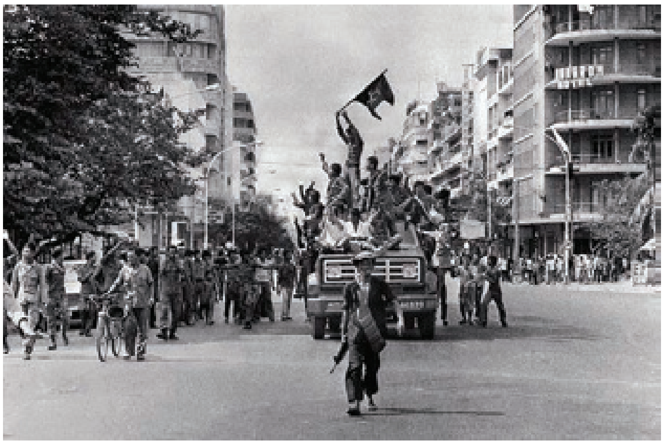


Figure 1: Khmer Rouge enter Phnom Penh 1975

The revolution was not merely a turn from the right-wing capitalism of the Lon Nol regime but was also an attempt to eliminate all evidence of the nation's political, cultural and religious history. Key parts of Cambodian society such as Buddhism and their temples were left destroyed just to name one example.

The Khmer Rouge were trying to create a system of government where everyone was equal, and everyone worked together to achieve common economic, agricultural and political goals. They made everyone work in the fields, dig canals and build bridges for 18 hours a day and often with very little food. However, the way in which this was implemented and then enforced to such extremes meant that it turned out to be the exact opposite of an equal society instead making it a country where its people were all treated equally badly.

To try and achieve their equal society, those that had been working professionals before the revolution such as judges, bankers, soldiers and politicians were all in danger of execution and by the end of the regime in 1979, 80% of all teachers and 95% of all doctors had been killed.

Extreme torture was used throughout the terror to locate this previous professional class and one prison alone in Phnom Penh was found to have killed almost 15000 people either through torture or immediate execution. When I visited this prison, S21, I was given a first-hand recount by our tour guide, Mrs Chithy, of her experience under the Khmer Rouge. Her father, sister and brother all died, and she decided to escape to Vietnam only returning after the Khmer Rouge had fallen. She told us that she used to cry every morning, showing how the scars of the Khmer Rouge remain visible to this day.

But suffering and death was not merely for the educated. The policies of the Khmer Rouge became more and more murderous until they were just executing people for not working hard enough or uttering a slight complaint. Since 1979, fields have been discovered with the bones and remains of hundreds and thousands of people, showing the signs of mass genocide against the ordinary working people, not just those killed by political motivation.There are hundreds of known "killing fields" across the country and I visited one just outside of Phnom Penh which has a large tower in the centre packed with human bones making it a very distressing sight.

Executions were only part of the danger for the Cambodian people. For many, their health and starvation was a much more prominent problem with there being no doctors to look after them and only being fed a maximum of two bowls of rice a day. Diseases, such as malaria, were common and thousands of people died each year because of starvation.

In total 2 million people, out of almost 8 million, died between 1975 and 1979.

The impact of their economic and political policies meant that even after the terror had ended there would be a similar level of suffering for the Cambodian people. There was no immediate relationship with another country from which to receive support and the agriculturally based economy, of which everyone was a part, meant that it would take time for other necessary services to be established that would drive the country forward. The death of so many people and the weakness of those that were left hampered any attempt to rebuild the country both physically and mentally, with thousands of people making the treacherous journey across the Thai border just to receive some sort of care.

All of Cambodia's previous ways of pushing the economy forward such as restaurants, markets and trade had all been shut down and the executions of virtually all educated people meant that it would be difficult to set up such institutions again. The destruction of the physical landscape also left a massive scar for Cambodians both psychologically with many important Buddhist temples destroyed and practically with the removal of roads and bridges, vital for transport and earning a living.

Many attempt to quantify the impact of the Khmer Rouge simply by the number of people killed and it is important that it is principally remembered in this way. But is also worth noting the considerable political, psychological and economic impact that they had that made the lives of so many Cambodians after the regime almost as bad as it had been during it.

SECTION 2 - POPULATION RECOVERY

Two million people, almost a quarter of the population, dying in the space of five years had a profound impact on all aspects of recovery after the Khmer Rouge regime. In this section I will look at how the number of people living in Cambodia has increased and the demographics of the people living in Cambodia now. It is worthwhile to analyse such figures as it shows what resources Cambodia as a nation possesses to be able to rebuild and progress in all aspects of society, making it an important building block for the rest of the essay.

The population of Cambodia in 1974, according to the World Bank, was 7.531 million, a figure that had grown from 5.7 million in 1960. This number tragically fell to 6.6 million in 1980 with the population only recovering to its previous number in 1985 as people struggled to fight diseases and cope with the continuing conflicts between the Vietnamese and Khmer Rouge, (The Vietnamese had invaded Cambodia in 1979 and driven out the Khmer Rouge). Since then, there has been a steady growth rate, with the number of people reaching 12.15 million in 2000 and 14.31 million in 2010 (see figure 2). The population now (2019) is at 16.4 million and is expected to continue to grow, with a current increase rate of 1.8%.

Of the current populace, 48.81% are male and 51.19% are female, with the male percentage having increased from 47.04% in 1980, representing the most female-biased sex ratio in the region, (see figure 3). In terms of the age of the population, 50% is under the age of 22

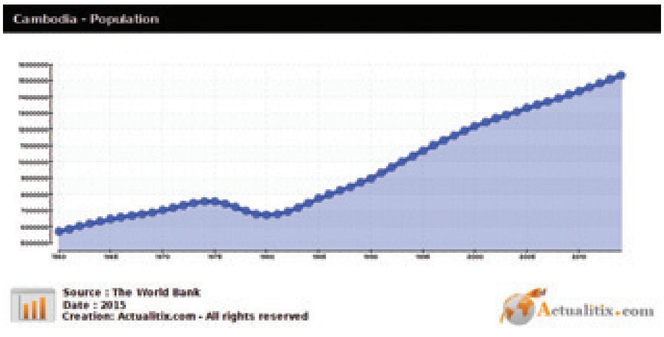


Figure 2: Cambodian population graph

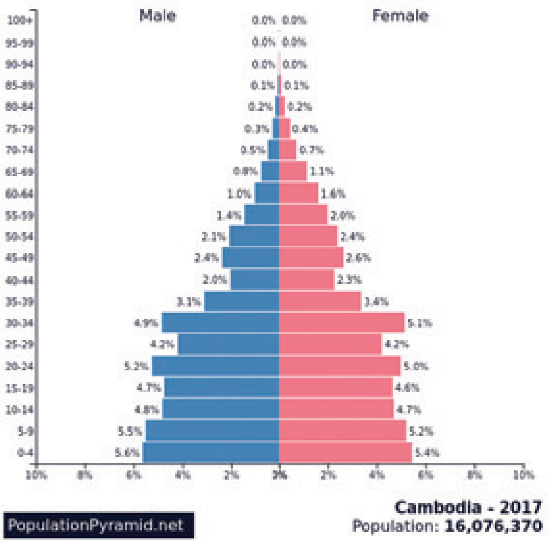


Figure 3: Male/Female and age population ratio

and in 2010, only 3% were over 65 years old. This last statistic is due not only to the sheer number of people killed by the Khmer Rouge that would have reached this age but it is also because of the poor health and social conditions that have been present over the past 3 decades in Cambodia, areas which I will come on to examine.

Much of the population is located in Phnom Penh, the capital, which boasts a head count of 2.2 million with the next largest city, Battambang having just 200,000. There is a clear divide between the welfare of Cambodians in urban areas versus those in rural settings with health care and education services, to name a few, being much more accessible in the cities. This is a theme that will make a constant reappearance throughout this paper and was a key point mentioned to me by many of the Cambodian people with whom I spoke.

The main religion in Cambodia continues to be Theravada Buddhism, with the figure being at 95%, despite the Khmer Rouge killing thousands of monks and destroying hundreds of temples. There are estimated to be 300,000 Muslims and 1% of the population is Christian, showing a small degree of diversity.

Overall, Cambodia is a growing, young, female dominated population with a welfare gap between those living in the city and the countryside. In my view, when just looking at the figures,Cambodia has recovered as much as it could have been expected to, given the huge extent of loss of life and it could be said that it is this population growth which will help Cambodia in the future with more people being able to enter the workforce. Figures about the population are important because they not only show us the damage the Khmer Rouge did to the country, but they also help us understand the resources that Cambodia has that they can use to continue to improve their society in the future.

SECTION 3 - POLITICAL RECOVERY

Before analysing the steps that Cambodia has taken to recover politically from the Khmer Rouge regime, we must first decide upon what constitutes a political recovery. Unlike economic recovery (which will be considered in the next section) where one can either see a clear incline or decline in economic growth, political recovery is much harder to define. Therefore, we must find certain areas inside the political umbrella upon which judgements can made by comparing it to the situation before the Khmer Rouge and also immediately after it. The three areas of political recovery that I will look at are: how a government comes to power (i.e. elections or lack thereof); the stability of the government; and the corruption present in government, in the hope that I can paint the picture of the rise, the maintenance and the workings of the Cambodian government. An important point to remember is that people can often be guilty of judging countries similar to Cambodia by western standards especially when it comes to democracy. However, I do not apologise for doing so in this section due to the fact that Cambodia calls itself a democratic country and therefore should be judged accordingly by these standards.

In the 1950's and 60's, Cambodia was ruled by Prince Norodom Sihanouk who abdicated from his position as monarch in 1955 to run in elections in order to give himself the ability to engage in the politics of the nation. Sihanouk was seen as something of a God in Cambodia during this time, a point especially made by Haing Ngor in his book 'Survival in the Killing Fields' and is emphasised further by virtually every hotel, restaurant and school in Cambodia having his portrait on their walls. These elections were attributed with serious amounts of fraud and intimidation giving Sihanouk the victory with every seat in the National Assembly. Prince Sihanouk ruled the country until 1970, at which point he was abroad, and a coup led to a right-wing military dictator, Lon Nol, seizing power.

It is clear from this brief account of Cambodia in the period before the Khmer Rouge that it is a relatively low bar to set for Cambodia to return politically to how it was before and therefore we must place greater emphasis on what has or has not been done to improve their situation to a more democratic country.

3.1: ELECTIONS

At first it could be argued that Cambodia has recovered effectively due to the presence of elections every 5 years since 1993. Although it took 14 years to happen, the UN, who took control of Cambodia in 1992, managed to set them up and also draft a new constitution which made special provisions for human rights. Despite the period 1979-1992, we can see a clear political recovery due to the fact that regular elections take place. The UN can be seen as the catalyst for this recovery, sending in 16000 troops and, more importantly, 5000 civil administrators to help organise them. This was as much a victory for the UN as it was for Cambodian democracy, with the turnout being over 90%, and this number remaining consistently high for the elections to come. Furthermore, Cambodia held its first ever local commune elections in 2002 with numerous reforms to the electoral process being suggested currently.

However, whether we can call these elections "free and fair" is a different question and actually it is better to argue that they resembled many of the similar features that prevailed in the 1950's and 1960's.

Figure 4: Cambodian voting at first election in 1993

Political assassinations were commonplace, particularly carried out by Hun Sen who was trying to eliminate his political opponents in the Funcinpec party, led by Norodom Ranariddh. There was also serious intimidation from the remaining members of the Khmer Rouge who attacked numerous polling stations, adding weight to the view that the elections were neither "free" nor "fair". These types of action have continued since then, with political assassinations taking place before national elections in 1998, 2003 and fraudulent voting during 2008 and 2013 general elections. Even as recently as the 2018 general election, problems included: intimidation by CPP officials; a record number of invalid ballots; and a lack of opposition, due to the government having arrested Kem Sokha, the leader of an opposition party, for 'treason'. I found it is impossible to go anywhere in Cambodia and not see the big blue signs advertising Hun Sen's Cambodian People's Party. If I were to be voting in Cambodia not only would I not know that there are any alternatives, I would also be scared to not vote for the CPP. I spent a long time talking to a UWS (United World Schools) worker called Sem and he asked me whether there are signs like this everywhere in England. He was surprised when I said there wasn't and even more so when I told him that our Prime Minister often changes every 5 years. His reaction arguably shows how normal Cambodians may have a flawed view of democracy and therefore do not have the desire to potentially challenge what is in front of them as they do not know what their country could be like. Therefore, we can see that Cambodia has recovered to the extent that elections are held and that the turnout is relatively high, making it similar to the political situation before the Khmer Rouge but it is much more difficult to argue that Cambodia has improved the quality of such elections. The presence of violence, lack of opposition and fraud mean that Cambodia has still got a long way to go in ensuring that they are "free and fair". However, it is important to note that in comparison to neighbouring countries Cambodia is in a similar position with recent elections in Thailand having allegations of improper influence and elections in 2014 being named unconstitutional.

3.2: GOVERNMENT STABILITY

Elections are not the only way to gauge political recovery. We can also look at the stability of the governments that have been in power since the end of the Khmer Rouge and, ironically, we will see that the government has arguably been too stable, mainly due to the aforementioned weaknesses in the electoral system. We must start by seeing stability as a good thing and at face value, Cambodia has been a stable country with Hun Sen being the leader since 1985.

This is not to say however that there have not been destabilising events. A major point in Cambodian history was the grenade attack on March 30th, 1997, thought to have been carried out by Hun Sen[2], at Sam Rainsy's party convention, killing 16 people. Even more significant was the military coup in 1997, (see figure 5). The country had previously been led jointly by Hun Sen and Ranariddh but in 1997 fighting broke out between the two leaders after Ranariddh moved troops on the capital. The conflict lasted two days and Hun Sen won easily, leaving himself free to seize complete control over the country. The fact that such events took place must limit the extent of political recovery as Sen would not have carried out the aforementioned grenade attack if he felt stable and there would not have been a coup if both sides felt strong in their position.

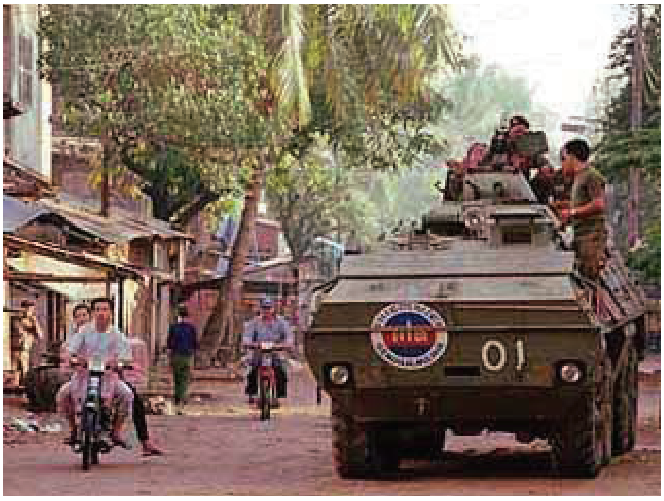


Figure 5:Tanks moving in on the capital in 1997

However, on the whole, Hun Sen has been able to keep himself in government for 34 years but the problem for the Cambodian people is his actions and in some areas lack of action that has not been in the best interests of the Cambodian people. Hun Sen rules as a virtually authoritarian dictator and it has been argued on many occasions that he is more interested in lining his pockets than providing services for the people. Clear examples of this include: building schools to say that he has built them but then not providing any teachers to fill them; a lack of an irrigation network that meant that rice production was 1/3 of what it had been in the 13th century; and not providing mental health services to deal with almost half of the population who had PTSD. There is also a clear difference in his policies between cities, especially Phnom Penh, and rural areas. There is huge investment in Phnom Penh, especially with infrastructure at the moment, whereas in the countryside most people live in raised wooden rooms with very few opportunities to better themselves. Here we can see why the stability of his reign (reign being an appropriate word for his 34 years in charge) is a problem for the Cambodian people, who despite protests such as 20,000 people taking to the streets in Phnom Penh in 2013, have been unable to remove him from office, hence the clear limits on Cambodia's political recovery.

3.3: CORRUPTION

A final, more obscure, way of analysing Cambodia's political recovery is by looking at the amount of corruption present in government. Corruption has always been a problem in Cambodian society with people facing the issue in areas such as obtaining medical services and getting fair verdicts in courts, with wealthier Cambodians being able to offer bigger bribes and, as a result, receiving preferential treatment. This was a key point explained to me by Ben, a tour guide in Kratie, who said that he had to give up school after his family could no longer afford to pay the bribes that were needed for him to pass exams.

The best way to analyse the extent of corruption in Cambodia is by looking at legislation that has or has not been passed that attempts to deal with the issue and whether this has made an impact. Hun Sen had been promising, since the UN left, that he would pass an anti-corruption law but had been making excuses for it failing in the National Assembly until it eventually passed years later. A law was eventually agreed in 2010 which stated that officials found guilty of corruption can face up to 15 years in prison and also established an anti-corruption unit which has investigatory and disciplinary powers. Despite such improvements, it has been found that this law has been not been enforced rigorously and Cambodia was recently ranked 156th in the world by Transparency International's 2016 Corruption Perceptions Index, having dropped six places since 2015. This is probably not helped by the fact that, as explained by Ben, there is still a large amount of bribery needed to gain the top jobs in the city, especially with workers coming from the countryside. However, one can argue that Cambodia has made significant improvements regarding corruption not only from the Khmer Rouge but even compared to where it was in the period before. But compared to the rest of the world there is still great deal left to be done, starting with a stricter enforcement of the anti-corruption law. If corruption is stopped at the highest level, then it should permeate beneath the surface into all aspects of Cambodian society.

Ninety percent of an iceberg is beneath the surface of the water and the same can be said, in my opinion, for the overall political situation in Cambodia. The presence of elections since 1993 and stable government since 1997 paints the picture on the surface that Cambodia is becoming a strong democratic country. It is certainly fair to say that Cambodia has democratic potential but the reality is that the political situation is not much better than it was in the 1950's and 60's. Vast amounts of corruption and political violence mean that there is still a long way to go before Cambodia can actually be the free democracy that it calls itself and therefore I would argue that there is still a long way to for Cambodia to achieve this goal.

SECTION 4 - ECONOMIC RECOVERY

In order to evaluate the extent of economic recovery, we must first look at what the economy of Cambodia was like before the Khmer Rouge to see, as with the political recovery, what will constitute economic recovery. During the rule of Prince Sihanouk, the economy made small steps with the Prince opting for unconditional aid from both the East and the West and there being a generally high agricultural production rate. The economy was understandably less prosperous during the civil war and the Lon Nol regime where there was a damaging effect on rice production mainly due to there being only one third of land under cultivation. Despite attempts at reform and foreign aid, the economy collapsed in 1975, leaving the country heavily reliant on the US even for food. I will look at the numerous ways in which Cambodia has recovered, all the while noting the limitations and I will conclude that Cambodia is a fast-growing economy with strong infrastructural potential but in relation to other countries is a long way off achieving economic stability.

4.1: MACRO-ECONOMIC SITUATION

There are different ways of measuring economic recovery and in this section, I will look at levels of GDP (Gross Domestic Product) since the Khmer Rouge; economic relations with other countries; and the economic well-being of the individual. A clear argument can be made that Cambodia has recovered in all of these areas and in many ways, we are just talking about normal economic development rather than any special form of recovery. This is firstly shown by an exponential rise of GDP, with it being $588.4 million in 1974, the year before the Khmer Rouge took over, and $22.16 billion recently in 2017. This growth is not as simple as just rising by the same amount every year. It had only risen to $2.534 billion in 1993 due to the lack of attention being given to the economy by the Vietnamese who were focusing on driving out the Khmer Rouge and therefore we can see the UN intervention in the early 1990's as the catalyst for economic growth in the 21st century, rising to 10.35 billion in 2008 and then $18.05 billion in 2015. The change by the government in 1995 to introduce a market driven system and the developing tourism and textiles industries have both been major causes of the rapid growth. It can be argued that this is not just a return to the status quo and is instead a clear sign of improvement from the pre-Khmer Rouge period. However, just because we can see signs of improvement it by no means shows that Cambodia has a strong economy. The increase in GDP is made to look substantially better by the incredibly low starting position and the figure for 2017 was considerably less than surrounding countries such as Vietnam which was $223.9 billion. This fact is something that Cambodian people are aware of as well. The UWS worker Sem said to me when we were talking about poverty in Cambodia, "If you look at Vietnam, Laos and Thailand they are much better". The Khmer Rouge period set Cambodia back such a long way compared to these countries and has still not caught up, something which ought to be at the forefront of the government's economic plans.

We can also evaluate Cambodia's economic performance by looking at the amount of foreign investment that they have received. Cambodia has only been attracting foreign investment since 1994 as foreign capital was not permitted before by the government, so we must look at Foreign Direct Investment (FDI) since then to see if they have been making good progress. In 1994, FDI made up 2.5% of the country's GDP and most recently in 2017 it was at 12.5% with the highest being at 14% in 2012. The increase in foreign investments shows us the rise in confidence that foreign investors have in the Cambodian markets and therefore, although it is difficult to call it a recovery because foreign investment was not permitted until 1994, we can certainly see economic development.

4.2: MICRO-ECONOMIC SITUATION

We have looked at the macro economic situation of the country, but it is also important to analyse how this looks for the average citizen and their day to day life. A responsible way of investigating the economic well-being of the population is by looking at the GDP per capita which looks at the country's economic output while accounting for its number of people, therefore making it a good way of analysing a country's standard of living. However, we must remember that there are couple of problems with relating GDP per capita to household disposable income. Firstly, not all income generated by production remains in the country and secondly some parts may be kept by the government or big corporations, with the latter certainly being the case in Cambodia. It can still be useful however and in 1974 it was $78.4 per head increasing to $745.79 in 2008 and then $1,384.42 in 2017, (see figure 6). Although we must treat it carefully, we can see an exponential rise in the average economic well-being of a Cambodian household.

However, this was not the message that I received from Ben, the Kratie tour guide, who said that there are around 2 million Cambodians working abroad especially in South Korea because the wages are much higher.

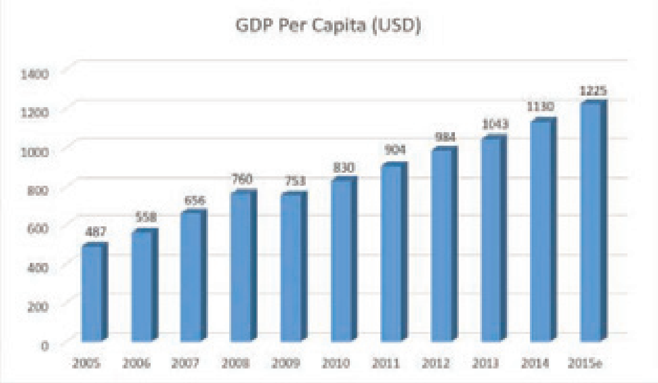


Figure 6: Graph showing increase in GDP per capita

Figure 7: Repairs on the Southern Line

One area where, at first, it seems as if Cambodia is progressing effectively is in terms of unemployment. Given that employment did not exist as a concept under the Khmer Rouge, unemployment levels such as 0.2% in 2016 are a remarkable achievement. However, once again this is not the full story and often the jobs are of poor quality and are only part time, meaning that people are struggling to make ends meet, as reported by the ILO in 2018. I saw this first-hand with our English-speaking guide, Dyna, in the Koh Kong jungle. He has three jobs at the same time: a jungle guide, working on the farm and some work with a film company in Phnom Penh. Although he is likely to be better off than most Cambodians living in the countryside, it still shows that work in Cambodia is sporadic and often not of the highest quality. Furthermore, the GDP per capita figures are substantially less than neighbouring countries, which also experienced communist rule, with Vietnam at $2,343 in 2017 and Laos at $2,457. Not only are they significantly higher than Cambodia, they have also risen faster, with all three having been virtually at the same level in 1985. Therefore, we must, despite the rise, place Cambodia in context with other countries and see the limitations of their economic growth.

4.3: INFRASTRUCTURE

Finally, in terms of the economy, it is important to look at the infrastructural recovery and the importance that this has on economic growth post Khmer Rouge. Normal economic development would not usually need to take into account almost wholesale physical rebuilding projects but given the extraordinary amount of damage inflicted by the Khmer Rouge and the surrounding civil wars it is an important part of analysing economic recovery as much of the economy depends on what infrastructure there is to support it.

A key part of any country is its transport networks and especially its roads which are necessary both for trade and human movement. The conflicts of the late 20th century destroyed almost all of its road networks, that weren't particularly numerous anyway, meaning that it was difficult for aid to be brought in, especially during the 1980's. The hard-pressed economic situation that the government found itself in in the 1990's meant that they went un-repaired and it was only in the 2000's that roads started to be rebuilt. There is currently a total of 47,263 km of road in the country, only a quarter of which is paved. It is important to note the role that China has played in the infrastructural recovery of Cambodia with their building of 70% of all Cambodia's roads. However, a lack of trained builders and supervision mean that roads that are built are often not up to standard. From my observations whilst in the country, the main roads in and going to the main cities are well maintained but the roads in rural areas are of poor quality and potentially dangerous at times, especially when it is wet.

The rail network, which peaked in 1970, has recovered slightly more effectively with the Southern line beginning passenger services again in 2016, (see figure 7) along with the transport of coal, fuel and cement. Furthermore, the Thai-Cambodianrail links were reconnected very recently in 2019, offering more opportunities for trade and tourism in the future. However, the Northern line is not so prosperous with flooding and a lack of maintenance meaning that the services listed above are not carried out. In the past two decades there has also been a rise in internet usage in the country with access figures having tripled since 2000 and having increased from 25% in 2015 to 50% in 2018. However, these figures are still below the global average as they are mostly only present in the cities, highlighting the work that still needs to be done in Cambodia to give everybody access to the internet. However, it can be argued that it is not quite as bad as it seems. The fact that infrastructure was all destroyed negates problems that other countries have with replacing old, inefficient systems and therefore puts Cambodia in a position where they can implement the newest technologies.

Overall, there is great potential for infrastructural and therefore economic growth, especially given the aid from China, but there is still a long way to go in terms of providing basic services for its citizens. In terms of the entire economy, impressive progress has been made, with limitations being that they have not yet reached the level of surrounding countries and that there seems to be a considerable urban rural wealth gap. But on the whole Cambodia has recovered fairly well economically and is poised at the next stage of improvement and development.

SECTION 5- PSYCHOLOGICAL RECOVERY

Recovery cannot simply be measured by political strength and economic growth. We must also look at the psychological recovery of the country, a task that poses something more of a challenge due to the lack of government published statistics on the issue and the varying level of figures that are available. This includes looking at the number of people suffering mental health problems and the quantity of services providing care for these people. However, it is important to remember that not all the mental health problems in Cambodia will have been caused by the Khmer Rouge regime but in this section, it is safe to assume that the vast number of mental health problems will have, in some way, arisen from the Khmer Rouge regime either directly or indirectly.

Before 1975, there was only one mental hospital, on the outskirts of Phnom Penh, which catered for the whole country. It was used as a prison by the Khmer Rouge and then not reinstated by the Vietnamese meaning that there were no mental health services in the country when they needed it the most. Experiencing and watching torture and seeing family members die are all huge causes of mental illnesses such as PTSD (Post Traumatic Stress Disorder), depression and anxiety and the lack of help available to the Cambodian people has meant that such illnesses have been allowed to go unchecked and in some cases, passed on to their children. This could be due to poverty because of their parent's potential inefficiency in the workplace or it could be because of their parents trying to mask their problems by turning to violence or drugs. Hence, we can see how much of a problem mental health is in Cambodian society as it affects all aspects of people's lives. The image is of the TPO centre in Cambodia, a leading institution in raising mental health awareness in the country (figure 8).

The above highlights the threats that mental illnesses pose to society, but an argument can be made that since foreign intervention in 1993, there has been a significant improvement in terms of the services provided. Analysing the increase in mental health services is both much more accurate and much easier than looking at any potential decline in mental health problems. Data showing mental health problems is simply not present especially for the period 1979-1993 and it is only recently that thorough surveys have been carried out showing legitimate responses. In terms of services provided however, we are much better equipped to track their increase and consequently one could argue that we can infer from them that at least people are getting better treatment and potentially numbers are decreasing. The first training practices for psychiatrists were established in 1994, with mental health being at the forefront of the government's health plan, and by 2010, there were 18 health centres and 50 referral hospitals offering mental health services. Furthermore, although there is no proven link, we can see the recent trials of old Khmer Rouge members as a way of the country moving on and potentially helping with future symptoms of PTSD. A very high-profile example of this was the life sentence given to Kaing Guek Eav, also known as Comrade Duch, who was in charge of the prison mentioned previously where more than 14000 people were killed. An increase in mental health services and the beginning of prosecution for Khmer Rouge officials means that one could argue that mental health problems in Cambodia are improving.

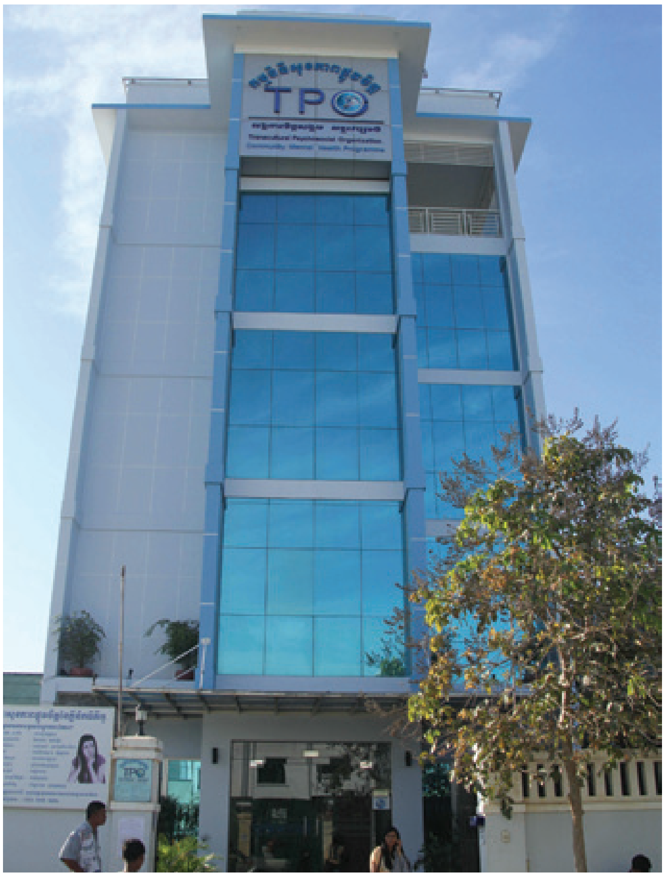


Figure 8: TPO centre in Cambodia

However, this argument is ultimately flawed and there are numerous statistics and surveys that show the mental health problem in Cambodia to be as bad as ever. Before we look at these stats, it is important to note the importance of speaking openly about terrible events such as the Khmer Rouge in coming to terms with it and consequently improving mental health. Talking to John, a tuk-tuk driver who lives in Battambang, he told me that the government doesn't like people talking about the Khmer Rouge and that you can get in trouble for speaking about it publicly. Mrs Chithy, the S21 tour guide, told me that when she became a tour guide and started to talk about what had happened, she stopped crying every morning. This shows the significance of discussing the Khmer Rouge purely from a point of view that it will improve people's mental health, something that the government doesn't seem to value as much as other areas.

Coming back to the figures, they show from 2009 and 2010 that almost 40% of Cambodians have suffered some sort of mental illness since the Khmer Rouge regime. Whether all this is because of the regime is another question but other studies have shown that 14% of all those aged 18 and older have PTSD, something that can be more directly traced to the suffering under the Khmer Rouge. It is not just those still living in Cambodia that are affected as the U.S. National Institute of Mental Health in 2005 found that 60% of all Cambodian immigrants suffer from PTSD. This is 17 times higher than the rest of the American population, consequently leading the institute to link it to the Khmer Rouge regime. The problem is considerably exacerbated by the significant lack of services to care for those with mental problems, with the statistics above not showing the whole picture. The 18 health centres that provided service in 2010 were only 2% of the total number of health centres and it was 50 out of 84 total referral hospitals showing how a vast number of Cambodian medical centres are incapable of proving mental health care. Furthermore, there were only 80 trained psychiatrists, most of them in Phnom Penh, in the whole country which, in 2010, had a population of 15 million. The main problem however is not the lack of psychiatrists nor the small number of medical centres, it is the lack of understanding and knowledge about mental health in Cambodia. There is a general taboo around the subject meaning that many people just hide it away and the problem is not dealt with. This is a top down change that needs to happen and for Cambodia to be able to deal with the issues that they have, education must be provided on the subject so that more and more people are able to recognise what they are going through and can go to get the help they need. Education, as we will come onto in the next section, is vital for recovery.

SECTION 6- SOCIAL RECOVERY

There are a number of different aspects that could constitute a social recovery after such a traumatic experience like the Khmer Rouge regime. In this section, I will focus on two key areas. Firstly education, looking at numbers of schools and teachers etc, and how education can lead to recovery for the whole society. Secondly, having already looked at the significant mental health problem affecting the country, I will also examine the physical health of the Cambodian people and what the government and other organisations are doing to help. These areas are linked, for example we can see education as a means for creating new doctors which are then in place to deal with the physical health of an increasing population. Furthermore, the social recovery is reliant on the political and economic situation of the country as the government needs to have the money and be willing to improve education and health services. We will analyse and evaluate both sections and conclude that recoveries and improvements have been made but considerable work needs to be carried out so that Cambodia can transform into a society where its people can flourish rather than just simply survive.

6.1: EDUCATION

Cambodia had made impressive strides in terms of education since gaining independence in 1953, with government departments setting standards and targets, marking a huge improvement from the period before where, for example, only 7 high school students graduated in 1931. The arrival of the Khmer Rouge took the country even further back however, killing 80% of all teachers and destroying virtually all the schools. Along with this, there was an extermination of all educated classes, such as doctors, lawyers and bankers, meaning that not only did a whole generation of children grow up with no education but there was no one present to teach the next generation. Given the incredibly low level that Cambodian education was reduced to, it is accurate to say that there has been an impressive recovery, especially after 1993. Before we look at the figures, it is important to consider the importance of education to Cambodian society. Speaking to Danny (another UWS worker) he said “I am very happy watching the children learn. The knowledge will help them”. Even if most of them end up becoming farmers, which is what Danny said most children do in rural areas, the fact that they have an education gives them the opportunity to better themselves if they so wish, something which was not possible during and immediately after the Khmer Rouge. This impressive recovery is firstly shown by the continual number of schools that have been built over the past three decades. The amount reached 12,889 in 2017, a figure that was 1,519 higher than in 2013, emphasising that improvement has not slowed down and there is a willingness from the government for these schools to be built. It is not just the government that are building these schools. Organisations such as UWS who have built over 100 schools since 2008, including Kam and Tiem Kram schools which I visited, are also vitally important in improving education services across the country.

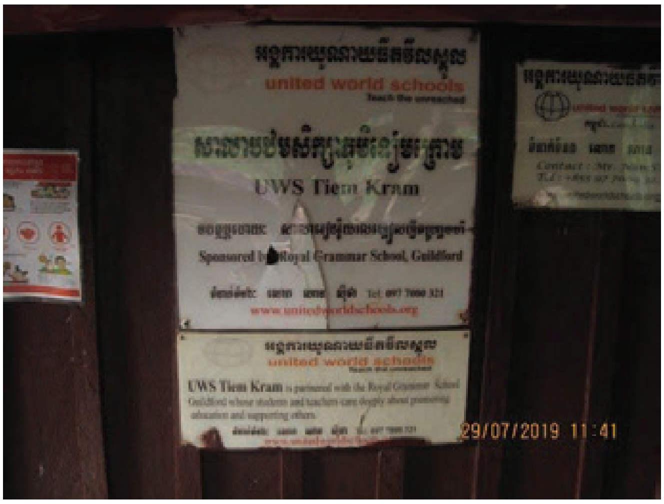


Figure 9: UWS Tiem Kram- sponsored by RGS

Linked to the economic recovery, there has been an increase in teacher salaries as well with the average minimum wage for teachers rising from $80 a month in 2014 to $230 in 2017, portraying the improved economic situation in Cambodia which then relates to a better education service. Furthermore, there has been an increase in the number of children enrolling in education, with the number being 82% for primary education in 1997 compared to 97% in 2018. Dyna suggested to me that this increase is because it is now seen in Cambodia as socially wrong for children not to go to school, a significant change in attitude for the better. This increased presence of children in schools has been the main factor in improved literacy rates for young people aged 15-24, where the rate was at 92% in 2015, increased from 88% in 2008. However, from spending a week in two Cambodian schools, although the literacy rate may be more accurate the attendance figures do not show the whole picture. Many of the children, especially in the countryside, do not come to school on a regular basis. I saw this on one of the days where it rained heavily and only half as many children turned up compared to the day before. Danny explained that when it rains the children are needed on the farms to help cultivate the rice. More needs to be done by the government so that families are not dependant on their children’s help so that they can be free to get an education. However, the improvement in not only the number of children attending schools but also the amount of schools being built shows a considerable recovery from the Khmer Rouge period.

These increased numbers, however, can often hide what is actually going on in the schools themselves and can mask whether any real progress is being made. While it is impressive that so many schools have been built, reports show that many of them are under equipped, for example it was found that 53% of all secondary schools have no water, consequently decreasing the standard of education that Cambodian children are getting. Furthermore, despite there being such high numbers of children in primary education, many of them have not been making the progress that they should, with only 27% of 3 to 5 years olds being on track with literacy and numeracy. This, added to a lack of knowledge about the importance of education from parents, means that by the time these children are 17, over 55% of them have dropped out of any form of education. It is at the top end of education where the biggest problem is. The high dropout rate and poor higher education programmes on offer mean that very few people come out of Cambodian education with skills that they can apply to the real world. Danny told me that he dropped out after secondary school and that very few people go on to university while Sem revealed that he pays for his son to have private education as well because the normal education sometimes isn’t good enough.

Figure 10: Children at UWS Kam School

In order for Cambodian society to progress, more money needs to be spent on higher education services so that the population isn't just a group of people who can read and write but are people who can think creatively and independently to push their country forward. It is excellent that there are increased numbers of children in schools and this has been an area of success so far, but the next step needs to be taken to increase the quality of education being provided, not just the quantity. This is being done by organisations such as UWS with workers like Danny, who is a teacher trainer, and he told me that he "gives ideas to the Cambodian teachers and they give them to the children", something which needs to be matched by the government.

6.2: HEALTH

mould as education with improvements being made in terms of the physical wellbeing of its people but there is still a lot of work to be done on providing a quality health service all over the country. Advancements in the health of Cambodians is primarily shown by the increase of life expectancy since the Khmer Rouge period where it dropped to 14. It rose to 45 in 1980, 56 in 2000 and reached a high of 67 in 2015 paying testament to better health care being developed over the decades following 1979. An example of this development was the strong work of getting HIV and AIDS under control in the early 21st century, two potent diseases that were major causes of early Cambodian deaths. Furthermore, maternal mortality rates have decreased significantly, going down from 409 deaths per 100,000 in 1990 to only 161 per 100,000 in 2015, showing the upgrade quality of birth care in the country. This has all been helped by enhanced sanitation in the country with the WHO (World Health Organisation) reporting in 2010 that 64% of households were now able to receive safe drinking water, consequently reducing the spread of diseases and improving the efficiency of medical practices.

However, there still remains significant problems with the Public Health Service meaning that Cambodia falls short of numerous international standards and lags behind many of its neighbours when it comes to health care. This is firstly because of a lack of quality health care workers. In 2011 the WHO published a report saying that there were 1.3 health workers per 1000 people, a figure considerably less than neighbouring countries such as Thailand. Another figure that shows Cambodia to be in much worse position compared to Thailand is the number of hospital beds per 1000 people where the figure is 2.2 in Thailand and only 0.7 in Cambodia. This was particularly emphasised by Sem who said that his mother was not able to go to a hospital to have him as it was too far away. Sem is one of 8 children. This shows us that there needs to be a significant increase in investment from the 6% of the Cambodia's GDP that is currently spent by the government in order to pay for both more health care workers and better facilities for them to work in. The inadequacy of health care in the public health sector has led to over 60% of patients opting for private health care, an option where there are fewer regulations and therefore higher risks for patients. Consequently, it is vital that the government spends enough money to improve their public health to encourage people to opt for it instead while the government starts to issue regulations to private health care companies.

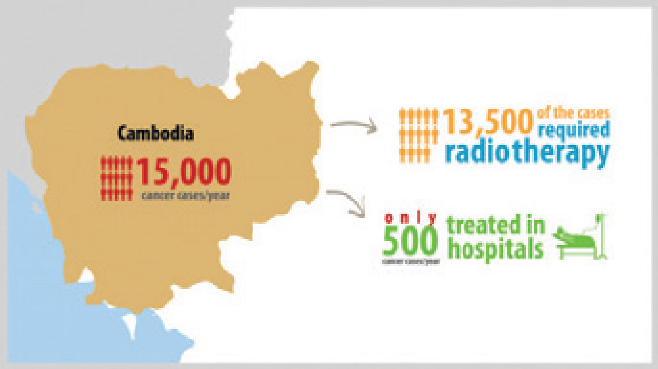


Figure 11: Cancer problem in Cambodia

Overall, given the high risk disease area and decades of conflict that Cambodian people have lived through, they have made an impressive recovery, improving both life expectancy and maternal mortality rates, however as with many aspects of Cambodian society, steps need to be taken to improve health care for all Cambodians by providing better facilities and more health professionals, especially in rural areas.

CONCLUSION

There are many different things that we can conclude from analysing Cambodia's recovery over the past 40 years. In answering the question "to what extent has Cambodia recovered from the Khmer Rouge regime" the answer has to be only 'partly'. There are numerous areas of recovery and improvement such as the establishment of elections in 1993 and annual rise in GDP in the 21st century. The trials of former Khmer Rouge officers and huge increase of the number of children in primary education all add to the argument for Cambodian recovery. However, we can still see elements of Cambodian society that have not recovered from the Khmer Rouge period and will take considerable time and money to sort them out. The lack of quality teachers in the education system and the small numbers of mental health facilities in the country are just two examples of areas where Cambodia falls short and these combined with unstable economic conditions compared to other countries and small numbers of hospitals throughout the country mean that we cannot say that Cambodia has recovered completely from the Khmer Rouge regime.

Overall, Cambodia's most successful area of recovery can be found economically with most of the criticisms just being that their figures are not quite as high as surrounding countries. The fact that Cambodia boasts a very high GDP increase rate means their economy ought to continue to grow towards these nations and it should not be the biggest area of concern for Cambodian governments. In terms of education and healthcare there is quantity but not enough quality and therefore in the future money must be spent to train up teachers and doctors to fill the gaps in these services. A much more serious problem is posed by mental health with the sheer number of people suffering from such problems and the huge lack of facilities to support them. Before lots of money is spent, the government itself must be educated on what the problem is and how it should be dealt with.

It is interesting to assess why Cambodia has been able to recover in some areas very effectively but has not been able to in other areas. From my research, the reason why I think this is the case is due to which areas the government sees to be important and which areas it does not. For example, it is in the government's interests to try and maintain political stability because it keeps them in charge and the slow progress made with anti-corruption was due to over half of the government being involved in such practices. Furthermore, a lack of progress on mental health and psychological issues is due to mental health being something of a taboo subject in Cambodia with the government seeing it as a relatively unimportant issue. The government, the one that has been the same since the Vietnamese left in 1989, is very reluctant to change and has showed over the past 30 years that the welfare of its citizens comes second to lining its own pockets and keeping its power. In order for Cambodia to make the substantial changes it needs to make to recover fully from the Khmer Rouge, a change of leadership is needed from Hun Sen and his Cambodian People's Party to one that cares about the well-being of the Cambodian people. Mrs Chithy said to me, "Life is difficult, but better than Pol Pot". The government should focus on removing the first four words of that phrase.

Biomimicry and Adhesion: The natural revolution of synthetic adhesives

Ben de la Court-Wakeling

INTRODUCTION

The mainstream market of adhesives has changed little in the past 50 years, despite the need for greater versatility and functionality. The problem hindering change is that there is no catch-all solution to the problems with conventional adhesives. Adhesives work through different chemical and physical mechanisms and attempting to collect them and explain their issues under one umbrella would oversimplify a complex topic. There are a wide range of adhesives used in the world for good reason: sticky notes require a radically different criteria for functionality than polyvinyl acetate as the former must be easily detachable and the latter a semi-permanent attachment method.

However, most current adhesives fall drastically short of the standard set by nature in their relative areas of application. In the 540 million years of random mutation since the Cambrian explosion, natural examples of adhesion have become highly optimised and thus frequently vastly superior to human designed alternatives. No synthetic adhesive can rival the strength of a mussel clinging to surfaces underwater (1,2); no reusable adhesive can self-clean or provide such a vast difference in attachment and detachment force like a gecko's toes (8), and no liquid adhesive can rival the humidity tolerance of a honey bee's (3). Through these biomimetic case studies new ground is being broken in developing the next generation of smarter, more proficient adhesives.

The applications of these and other bioadhesive technologies are extensive. Programmable, on/off adhesives could become a reality, providing traction for mobile search and rescue robots (10) or even allowing spiderman-esque climbing of buildings. Possibilities for new surgical adhesives, modelled on mussel adhesion, may allow implants to be directly attached to living tissue. (42). Even temporary mechanical fixings such as screws could be replaced by gecko based adhesives (8).

The aim of this paper is to analyse adhesion methods unique to the natural world which separate certain biological adhesives from their closest synthetic counterparts, focusing on the paradoxical adhesion of the Tokay gecko, and interpret the state of synthetic recreations.

SECTION 1: ADHESIVE DESIGN – THE HUMAN WORLD COMPARED TO NATURE

THEORY OF ADHESION

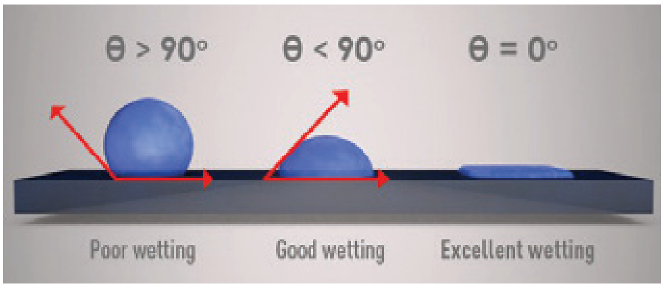
To begin with, an understanding is required of the principles of adhesion. For an adhesive acting between two substrates, the fundamental idea is of cohesive and adhesive forces. Cohesive forces are the internal forces acting to maintain the structure of an adhesive, opposing external forces which attempt to pull the mass apart. For water these are the intermolecular forces, primarily the hydrogen bonds and permanent dipole forces between the individual water molecules. If the cohesive forces are weak, the adhesive can fracture between substrates.

Adhesive forces are the opposite of cohesive forces as they dictate the strength of attachment between the adhesive and the substrate. There is no basic theory of adhesion, but instead four primary theories, which combine in varying amounts to account for the strength of each adhesive. (23,24,25) The first is adsorption, where Van der Waals attractions between the adhesive, usually liquid, and the substrate bind the two together, provided the liquid wets the surface to a significant degree. For porous surfaces, mechanical adhesion can play a major role as the adhesive seeps into cracks and solidifies, binding the two surfaces like Velcro. The third method is diffusion, found in polymeric adhesives where polymer chains interdiffuse between the two materials, seen commonly in plastic welding. The final primary method is chemisorption where reactions occur at the interface forming strong chemical bonds between the substrates.

The nature of an adhesive is determined by the relationship between the cohesive and adhesive forces. Two extreme examples, water and iron, are both extremely poor adhesives for different reasons. (23) Water exhibits relatively strong adhesive forces compared to its weak cohesive forces. This can be shown by the ability to stick wet paper to a wall and yet peel it off with ease: this is due to the water adhering to the wall and then splitting apart leaving a layer of water molecules on both surfaces. On the other hand, iron has immense cohesive forces due to the metallic bonding, yet minimal adhesive forces, relying on weak Van der Waals interactions.

The ratio of cohesive and adhesive forces is also strongly related to the physical state of the adhesive. The rough principle is that the greater the ratio of cohesive to adhesive forces the more solid behaviour a substance exhibits, and the smaller the ratio the more liquid behaviour a substance exhibits. This theory of adhesion, while greatly simplified, can describe much of adhesive design until now. Adhesives used for different functions utilise the ratio of cohesive to adhesive forces in varying ways. Tapes tend to use greater cohesive forces than adhesive so they can be peeled off and stuck back on without undue marking on the surfaces. On the other end of the scale, capillary action adhesives rely on the bonding between the molecules of the adhesive and the substrate to be stronger than the intermolecular forces. This allows the adhesive to creep up the surface of materials as it is energetically favourable to form as many bonds with the substrate as possible. As such these adhesives are thin liquids.

The principle of surface wettability is a major factor in maximising the strength of an adhesive. On a macro scale, two surfaces can look flat and flush with one another when on a micro scale the surfaces are irregular and rough. Therefore, if an adhesive is viscous, exhibiting solid-like behaviour, there will be minimal contact area between the adhesive and the substrate due to the inability to conform to the surface topography, reducing adhesive forces from the maximum. Surface wettability is defined as the ability for a liquid to spread out over a solid surface. The greater the surface energy of the substrate compared to the liquid the more the wetting will occur. This wetting is measured by the contact angle of the liquid with the substrate, where $\theta < 90°$ equates to partial wetting of the surface. The importance of this is the greater the wetting of an adhesive the more it can fill the irregular surface topography of the substrate, increasing contact area and thus the strength of the bond (28,31). Surface wetting is directly linked to the strength of the adhesive and cohesive forces of the adhesive, with greater wetting corresponding to reduced viscosity. Whilst the principle of wetting only applies to liquid adhesives a similar idea can be presented for solid adhesives, with more malleable adhesives contacting a greater proportion of the surface (47). This limits viability for conventional solid adhesives due to the reduced mechanical strength of softer adhesives. (26) The requirement of pressure sensitive adhesives (PSAs), such as blu-tack, to have Young's moduli inside the Dahlquist zone (i.e. to be tacky enough to maximise contact area), reduces durability and resistance to opposing forces, leaving them vulnerable to creep, degradation and fouling.



1 Surface wetting of a liquid droplet demonstrating contact angle θ
Source:https://www.masterbond.com/techtips/surface-wetting

METHODS OF CONVENTIONAL ADHESIVES

To maximise benefits of solid and liquid adhesion, the most common form of conventional adhesives are polymer-based adhesives. The mechanism of these involves a collection of monomers which react in a process called curing, sometimes activated by contact with air, sometimes with a reagent, to from long chain polymers which set solid. Existing in a liquid state before setting to a solid state enables them to maximise surface wetting so the adhesive fills all gaps in the surface of the substrate before setting solid. When solid, the cohesive forces increase greatly, improving the mechanical properties of the adhesive between the two substrates, such that far greater force is required to break the joint between substrates. In the case of polyvinyl acetate (PVA), the adhesive often becomes the strongest part of the joint.

Epoxy Resin is an example of a strong thermosetting polymeric adhesive. To make epoxy resin, a pre-polymer diepoxy is mixed with diamine, which sets off a reaction that results in the ends of the diepoxy molecules bonding to the diamine molecules and forming a crosslinked network of the molecules. Diepoxy and Diamine are liquids at room temperature due to the small size of the molecules, however the crosslinked network epoxy resin formed sets as a hard plastic between the substrates.

The adhesion energy of polymers is actually very low due to the low surface energy of polymeric adhesives as a result of the lack of polar functional groups in the polymers (30). As such, they only adhere by non-specific dispersion forces¬¬. This appears inherently paradoxical, due to polymeric adhesives making up some of the strongest structural adhesives humans have created. However, a study into macroscopic adhesion between rough surfaces shows that the 'van der Waals interactions between surface atoms produce attractive pressures that are orders of magnitude larger than atmospheric pressure', which can be utilised providing the material can maximise contact area (39).

PROBLEMS WITH CONVENTIONAL ADHESIVES

Creep is the deformation of adhesives over an extended period of time when a constant load is applied. Considering a lap joint, if a constant shear force with great enough magnitude is applied to one end the shear strain will remain constant for an initial period of time before increasing linearly with log time. This exponentially increasing shear strain is a result of the creep which continually weakens the mechanical properties of the material due to the flow of material in a complex microstructure. If the effect of creep is great enough then it can lead to failure of the adhesive at the joint.

The effect of creep can be greatly increased by temperature and humidity. Increasing temperature softens the material, reducing stiffness and making it behave more like a liquid. This means there will be an increased flow rate of material due to creep. There is also a relative critical humidity below which adhesive failure is unlikely to occur and yet when reached causes sudden adhesive failure. This is due to enhancement of the creep by fluids due to fluid degradation of the polymer. Reductions in shear strength of between 25% and 80% were noted for epoxies 'attributable to the effects of solvents' (4).

The market for multi-use adhesives is limited currently to pressure-sensitive adhesives, often in the form of tape. These adhesives utilise peeling for re-use as the force required for peeling is often many times smaller than the overall adhesive force of the tape. However, these adhesives have many flaws. Firstly, their scope for re-use is limited due to self-adhesion where the tape will stick to itself and reduce contact area, and adhesion to micro-particles which causes the tape to accumulate dirt across its surface with every new use again reducing adhesive contact area. This means they are only reusable a small number of times. In addition, the stronger the attachment force of the PSA the stronger the detachment force required, reducing the versatility of more powerful adhesive tapes due to the effort required for re-use. Thus, there is currently no such thing as a completely reusable adhesive.

Finally, there is a very limited scope for adhesives that can function underwater, with a small number of epoxy resins able to set in this environment (34). The vast majority of conventional adhesives experience failure in adhesion and mechanical properties underwater.

THE EXAMPLE OF NATURE

Honey bees have become a new source of inspiration for bio-inspired synthetic adhesives due to the remarkable rate tunability and humidity tolerance of the bee pollen adhesive they use to collect pollen, the major food source for any hive of honey bees. As they fly from flower to flower, they collect pollen and pack it into pollen pellets on each of their hind legs using the salivary secretion nectar as a glue. Honey bees carry a pollen pellet on each of their two hind legs with a combined weight of 25% of the mass of the bee, at speeds ranging from 1-20ms^{-1}, across months where relative humidity and temperature can change greatly. The way this is achieved is through a combination of the bee produced nectar and the plant produced pollenkitt which combine to create a two-phase liquid adhesive. The aqueous phase is the nectar, made up of glucose and fructose, which exhibits strong rate-dependent adhesion. Alone, the aqueous phase is highly susceptible to changes in humidity as water is taken up to dilute the adhesive or evaporated to concentrate it. In a study published in Nature Communications in March, researchers found a 40% decrease in maximum adhesion strength of a droplet of the aqueous phase alone when relative humidity was increased from 57% to 75%. However, when combined with the oily phase, the plant secretion pollenkitt which contains a complex mixture of saturated and unsaturated lipids, adhesion loss was halved at both 15% and 75% RH (3). These results lead to the current understanding that bee pollen adhesive is made up of the rate-dependent adhesive aqueous phase coated with a thin layer of an oily phase, pollenkitt, which does not contribute to adhesion however strongly mitigates the effects of changing humidity.

To create permanent adhesives which can function underwater without degradation, scientists are studying the adhesive secretions of mussels and barnacles which allow them to stick to the rough and often slimy surfaces of rocks, under the turbulent forces of oceans, without fail. Mussels secrete adhesive proteins into the byssal groove which then harden in a similar way to injection moulding. This forms a byssus, a bundle of hair-like structures with adhesive pads on the ends which attach to the surface they need to stick onto (1). The strength of these byssal threads is created similarly to synthetic polymer adhesives through curing, as smaller polymer chains of adhesive proteins crosslink to from a connected structure. The thread and plaque formation of the byssal threads allow them to function like guy-lines for the mussel, a result of the aligned biopolymer making the threads strong, highly extensible and yet hard, and even self-healing, whilst the adhesive plaque on the end provides the strongest known adhesion underwater (2). This adhesive pad can stick to biotic and abiotic substrates underwater in a way no synthetic adhesive can manage. Five adhesive proteins have been identified in the byssal plaques of *M.edulis*, Mefp-1,-2,-3,-4 and -5. Mefp-1 has been shown to bond to glass, plastic, wood, concrete and Teflon, and is most comparable to epoxy resins or synthetic cyanoacrylate due to its versatility in substrates it can bond to, quick polymerisation and high bond strength. Many of the proteins have high proportions of a compound called DOPA, which is believed to cause the moisture-resistance, as well as playing a major role in bonding to glass and rocks as it can complex with metal ions, oxides and semimetals like silicon.

Climbing animals and insects exhibit a remarkable case of convergent evolution. Geckos, stick insects, spiders, ants and other insects, through very different evolutionary paths, have achieved a similar solution to the need to climb on vertical and inverted surfaces with controlled, rapid attachment and detachment. Adhesion is largely achieved through Van der Waals interactions between the pad and the substrate. Therefore, all examples use a mechanism similar to tacky pressure sensitive adhesives by preloading and deforming the adhesive pad to increase surface area of contact, then mechanically modulating adhesive forces through application of shear force proximally for attachment and distally for detachment (7). In ants this takes the form of the claw flexor muscle pulling the pads towards the body and the elastic recoil of the stretched exocuticle doing the reverse for detachment. Originally, the wet adhesive pads of insects were considered to function very differently from the dry pads of geckos, controlling adhesion by viscous and capillary forces in the fluid. However, research from the Cambridge University Department of Zoology (5,6) discovered instead the fluid secretions act as a release layer, minimising viscous dissipation and time dependence of the pads. The behaviour of the wet pads was more akin to dry elastomers, and therefore not so different from dry adhesive pads.

The functionality of these dry adhesion methods, particularly the one used by Tokay geckos, is revolutionary to our concept of adhesives. The pads of a gecko's feet are non-adhering and therefore self-cleaning when unloaded, to maintain maximum contact area with the surface they are on. A gecko manages to detach its foot in the space of 15 milliseconds with no measurable detachment force, despite generating up to 40 times its weight in shear force when 4 feet are attached (8). Finally, the gecko's feet retain function despite countless attachment detachment cycles over its 10-year lifespan (9).

- -

Each popular conventional adhesive typically retains functionality over a narrow range of situations, with the lack of usability of strong permanent adhesives underwater and the unsatisfying trade-offs of PSAs between easy detachment and strong attachment, and durability and wettability. By analysing biological counter-examples (35), in this case honey bees, mussels and geckos, and breaking down the systems in use to understand the mechanisms and form reproducible models of their adhesion, scientists can look to improve the world of conventional adhesives. In some instances that is through improving existing ideas and methods by studying examples like mussel adhesion which utilises comparable methods to the curing of polymeric adhesives. In this case there are reports of synthetic proteins based on mussel foot proteins with increased adhesion strength underwater (37). However, examples like the Tokay gecko present alien methods of adhesion when compared to conventional adhesives with radically different properties, suggesting many problems with conventional adhesives could be solved by researching, understanding and adopting new biological methods of adhesion.

SECTION 2: THE MANY PARADOXES OF THE TOKAY GECKO

The adhesive pads on a Tokay gecko's feet bear little resemblance in makeup to any modern synthetic adhesive. Instead of fluid adhesive secretion the gecko utilises dry adhesion enabled by the complex hierarchical structure of its pads, detailed down to the nanoscale with evolutionary adaptations that make the gecko one of the most prolific climbing animals of its size. The ease of movement of the gecko, with their ability to run at a sprint vertically and traverse upside down on any natural surface they encounter, even when wet, has led to their species being studied for thousands of years. However, in the past 200 years

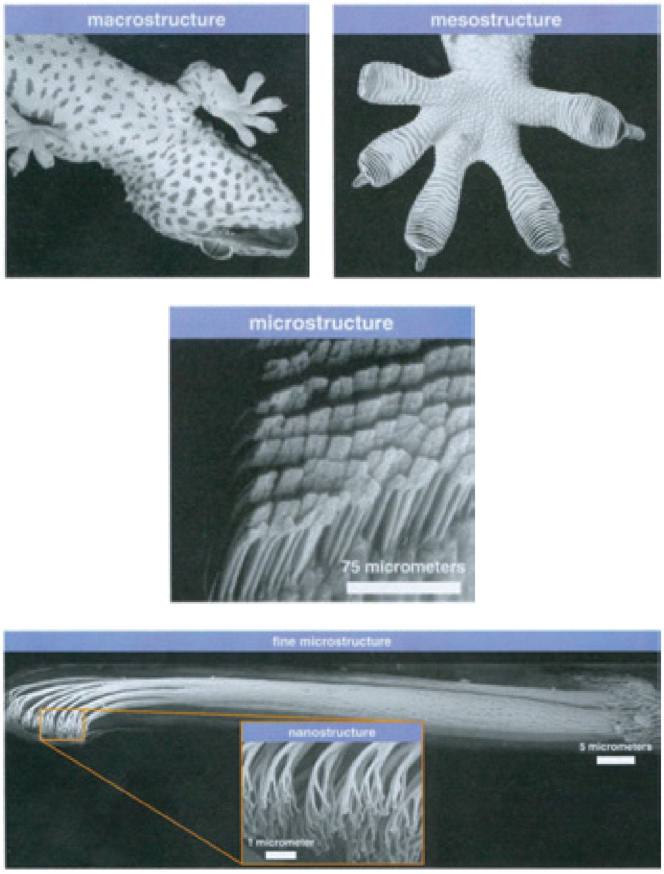


Figure 2: a) Tokay gecko resting on glass pane
b) Gecko toes displaying ridged scansors
c) setae arranged in diamond clusters d) single seta,showing branching into nanoscale fibrils of β-keratin (spatulae) towards curved end.
Source: (8)

scientists have been able to reveal the layered structure of the toes and come closer to explaining exactly how the gecko achieves its feats of movement.

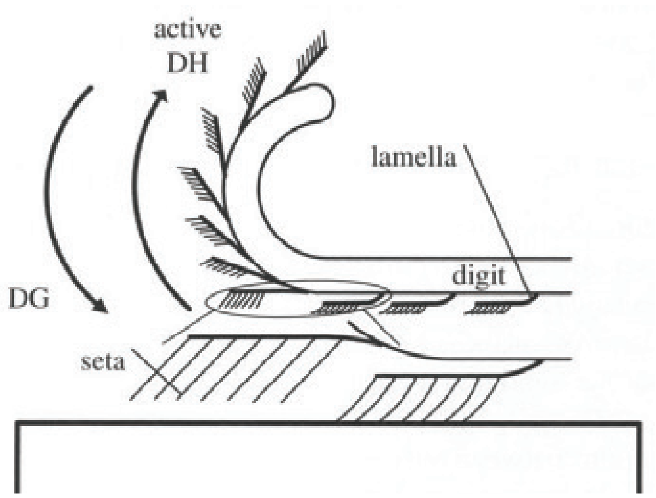
Unlike tree and torrent frogs, the Tokay gecko does not tend to use wrap-around gripping of surfaces and does not require the use of adhesive surfaces on the belly and upper limbs, utilising only its feet. The feet are covered in tiny ridges called scansors (adhesive lamellae). Each scansor has uniform micro arrays of hair-like structures called setae, comprised of fibrils of β-keratin, with each seta at roughly 110 μm long and 4.2 μm wide. The seta group into fours to form a diamond shape on the end. Branching of the seta into thousands of nanoscale filaments of β-keratin towards the ends, known as spatulae, can be seen in Fig. 2d, with flattened tips of 0.2 μm x 0.2 μm forming a precise mould to the surface. In the unloaded (unattached) state, the setae are recurved proximally with the left edge of Fig 2d approximate to a vertical surface a gecko is about to step onto whilst climbing (16).

Findings by Autumn et al (2002) demonstrate whole animal measurements of the maximum parallel force exhibited by Tokay geckos as 20.1N, thus averaging 6μN per seta. However, singular seta were found to be non-adhering unless a normal preload force and slight rearwards drag of 5μm were applied at which point the shear force was measured at 200μN. This was 32 times the whole animal measurement, enough force that theoretically if all 6.2 million spatula functioned at full capacity a gecko could support the weight of two medium sized humans (130kg). Therefore, only a minimum of 2% of the setae need to be attached to support the maximum shear force demonstrated in whole animal measurements. This safety factor of 3900% will not be this high in practice due to the setae not all engaging upon foot attachment and micro particles like dirt blocking adhesion. However, a gecko is still able to catch itself from a fall or fend off predators with only 1 foot attached. The preload and drag required to reach maximum adhesion mimics the foot placement of the gecko whilst climbing, pulling the setal shaft into tension and pointing the spatulae distally to bring them flush with the substrate (16).

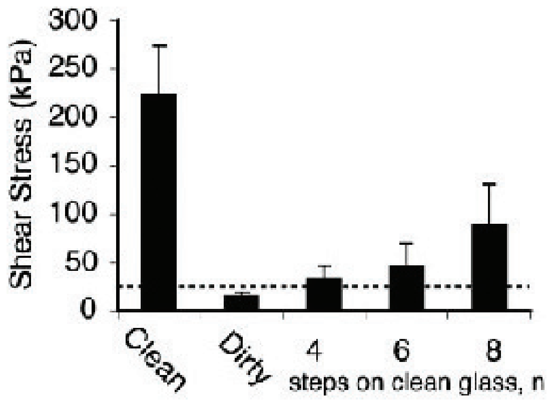
Fluid adhesive (glue); suction cups; mechanical interlocking (hooks); friction; electrostatic forces; capillary forces and van der Waals adhesion have all been suggested as mechanisms for the gecko's adhesion. All of these explanations have been rejected bar van der Waals adhesion in a number of studies over the last hundred years, with the final decisive study undertaken by Autumn & Peattie in 2002 where they demonstrated that the adhesive forces of gecko setae do not increase on a hydrophilic material, as would be expected if capillary adhesion was a major factor. They recorded a 2% increase at the single seta level for adhesion forces with hydrophobic SiO_2 and with hydrophilic Si semiconductors. Indeed, gecko setae are super hydrophobic with a water contact angle of 160.9°. The only force that can allow two hydrophobic materials to adhere in air is van der Waals.

This method of sticking for gecko feet means they prove extremely versatile in the materials they can adhere to. The spatula nanoarray provides a very large contact area, and the super hydrophobicity of the setae indicates they are highly nonpolar, therefore ensuring very little difference in adhesion to polar and non-polar surfaces. Van der Waals forces are largely independent of surface chemistry, however are greatly dependent on distance between substrates thus the adhesion is more dependent on surface geometry.

The question then lies as to how geckos can detach their feet within 15 milliseconds despite this immense adhesive force. As noted with the single seta attachment, the adhesive force is created only after an initial preload and drag. In essence, the chemical adhesion is 'switched on' by mechanical modulation. With detachment this principle is even more sharply noted since increasing the angle between the setal shaft and the substrate to 30°, the critical angle, causes immediate detachment with little force required. The adhesion of a gecko's foot, therefore, is programmable. Initial theories to explain this detachment linked the detachment to the peeling of a pressure sensitive adhesive. Increased stress at the trailing edge of the seta breaks the bonds between seta and substrate, causing the seta to return to the unloaded default state. This was supported by the macroscale hyperextension of the toes and apparent peeling (20). However, this suggestion fails to explain how geckos are able to remain inverted when weight should cause the toes to begin to peel, and also predicts the detachment angle to decrease with increasing peeling force. A study by Autumn et al (2006) (20) found the detachment angle to be independent of applied force and therefore rejected peeling. They presented an alternative model based on a new concept of frictional adhesion, a model unlike conventional adhesion wherein friction is a function of the normal force. Instead, this model proposes that the adhesion is a function of the shear force with a linear relationship. Thus, gecko's uses opposing feet to generate friction (shear forces) on inverted surfaces, to maintain a setal angle less than the critical angle therefore allowing fine control of detachment. Wu et Al (2015) (13) took this model further, incorporating it into a hierarchical model of the adhesion that explains the digital gripping (DG) and digital hyperextension (DH) of the gecko's toes. Attaching the setae to the digit of the gecko are the lamellae, the individual rows arranged across the digit which can be directly controlled. If the setae were attached directly to the digit friction would have to be lowered as far as possible to keep detachment energy efficient, due to the relationship between shear force and adhesion. On attachment to the surface the gecko curls it's toes down (DG), bending the lamellae inwards and pushing the setae down to act as the preload and drag required for maximum adhesion. For detachment the toes are hyperextended, bending the lamella outwards and releasing the pressure on the setae to reduce adhesion.



3 Figure 3: Simplified diagram of gecko toe, demonstrating digital gripping (DG) and digital hyperextension (DH). Source:11

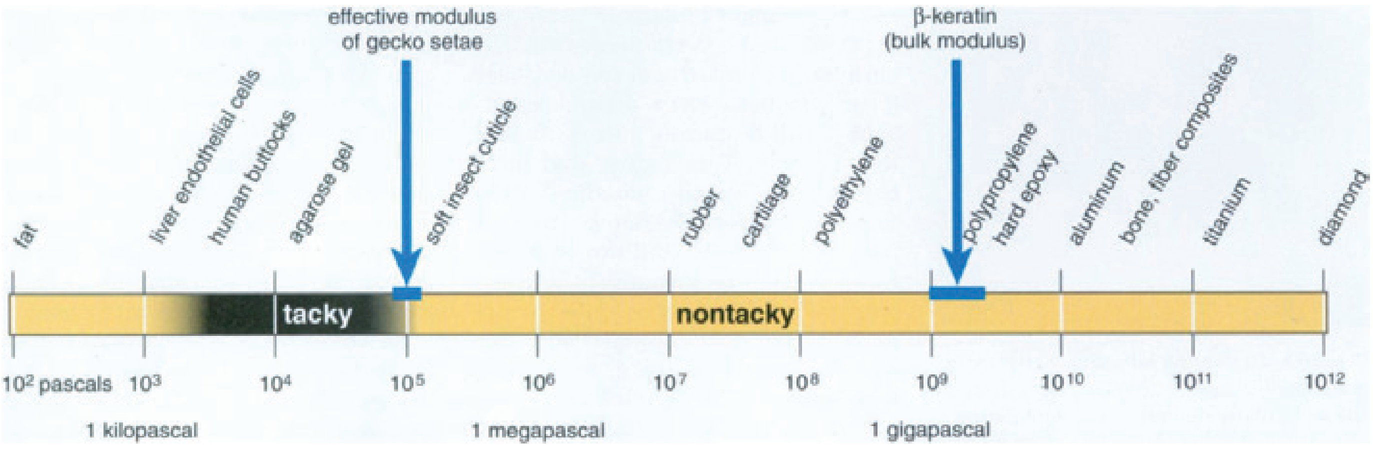


2Figure 4: Kellar Autumn's investigation into self-cleaning by dusting gecko feet with microspheres. Dotted line indicates minimum force required to support its bodyweight with one toe. Source: (8)

Another unique feature of gecko adhesion is that, over a lifetime of use, their adhesive pads remain clean without grooming or fluid secretion. Unlike conventional adhesives it appears that gecko feet get cleaner with repeated use, shedding the micro-particulate contaminants which interfere with adhesion (21). The reason for this is that, paradoxically, gecko setae are strongly anti-adhesive. The super-hydrophobicity of the setae, resulting in no polar forces or hydrogen bonding, lowers their adhesion energy such that the adhesion energy of all spatula touching a particle is less than the adhesion energy between the particle and the surface. When combined with the minimal surface area of the setae in their recurved default state, the base of the gecko's foot will shed micro-particles as it walks. If polar forces and hydrogen bonding were allowed, individual setae attachment may be stronger however they would lose the self-cleaning properties, increasing the number of contaminated spatula and reducing adhesive force of the array (21,28). In addition, the digital hyperextension serves an active self-cleaning function by rapidly expelling dirt particles as the toes peel back (10).

The closest comparable conventional adhesives are pressure sensitive adhesives (PSA). Like PSAs, the gecko adhesive pads can be reused, albeit indefinitely due to their self-cleaning properties, and can deform to create maximum contact area with the morphology of the surface. However, PSAs deform plastically upon applied attachment force and are prone to fouling, degradation and creep over time. Gecko toe pads deform elastically, returning to the default unloaded state after detachment, and do not foul, degrade or creep over the lifetime of a gecko. This is due to the setae consisting of filaments of β-keratin, a very hard material capable of withstanding significant stress. However, β-keratin has a Young's modulus of $2.6x10^9$ Pa, 4 orders of magnitude higher than the boundary for the Dahlquist criterion, the zone which defines tacky materials (materials which spontaneously deform to increase surface contact and can be detached without leaving residue). Through the hierarchical structure of the digits the effective Young's modulus is reduced to 100kPa, the upper end of the Dahlquist criterion, thus allowing the hard and strong setae to behave like a tacky adhesive (8). Early models explain this by relating the setae to cantilever beams which behave as spring.

These remarkable features serve as a reminder of how far synthetic adhesives have to go in comparison to their natural counterparts. Gecko adhesion throws up a number of paradoxes, the solving of which provide insight into novel ways to approach adhesion and engineer smarter synthetic adhesives for the future: how the feet are both non-adhering and yet can theoretically provide enough force to carry the weight of two humans; how the feet stick to surfaces and



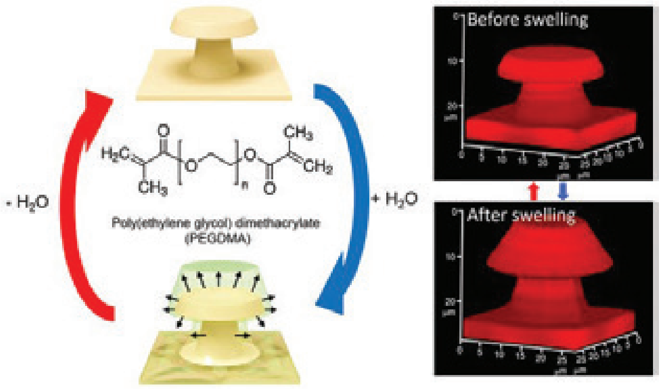
not to micro-particles; how the feet are made of stiff beta-keratin and yet wet the surface, and finally how the feet can detach with so little force compared to the adhesive force. It is only by understanding the dynamics of the entire gecko system, using an integrative approach to every level of the adhesive structure that it can be accurately mimicked (11).

SYNTHETIC RECREATION

As understanding of Tokay gecko adhesion has developed since the turning point of Kellar Autumn's study in 2000, there have been many attempts to synthetically recreate a comparably proficient dry adhesive. The challenges posed by this are tremendous due to the great complexity of the hierarchical structure of the gecko adhesive pad, including the need to create nano-scale replicas of the setae and spatula. Initial attempts (8) had adhesion coefficients (ratio of adhesive force to preload) of half to one percent of gecko setae. More recent attempts have provided increasingly more accurate representations, with a Korean team in 2009 (15) measuring 26N/cm^2 of shear attachment in the angled direction of their nanohair material, and 2.2N/cm^2 in the opposite direction. However, this feat represents more of an understanding of reproducing the basic mechanics of 'hairy' adhesives, and lacks the defining criteria of gecko adhesion such as the

A different team from the University of Massachusetts Amherst bypassed the challenges of manufacturing nanoscale hairs, focusing instead on the larger scales of gecko adhesion by utilising commodity materials and draping adhesion to gain many of the benefits of gecko adhesion. By integrating soft elastomers such as polyurethane with stiff fabrics like Kevlar, along with a synthetic tendon, the dual benefits of surface contact and elastic stiffness provide powerful adhesion in the determined load-bearing direction which surpasses Tokay geckos, whilst allowing easy detachment. An index card sized piece of their material Geckskin could hold over 300 kilos in a deadlift. (39,40)

A more human inspired approach to reversible adhesion has been pioneered by a Korean team (38), who aimed to find a method of reversible adhesion under aqueous conditions. This is highly desirable as current attempts at solving underwater adhesives use synthetic mimics of mussel foot proteins which are not suitable for all purposes due to the irreversibility of the adhesion. This new idea uses interlocking hydrogel micropillars which absorb water and swell, fixing the pillars together to provide strong mechanical adhesion of 79N after 20 hours of swelling. When dried, the swelling is reversed, reducing mechanical adhesion and allowing the two sheets to be pulled apart with a force of 9N. The structure is built from polyethylene glycol dimethacrylate (PEGDMA), which traps water in the three-dimensional polymer networks resulting in significant volume expansion, with adhesion force found to increase over time left in aqueous conditions.



4 Figure 5: Swelling of the hydrogel in aqueous conditions. Source (38)

With innovations such as these appearing out of worldwide research labs, the emerging world of synthetic adhesives promises to look far different to the conventional mainstream adhesives that have dominated the market. Many of these innovations require not only creative human minds, but the examples set by the ultimate designer of Evolution. By breaking down these examples, entirely new theories of adhesion can be constructed, demonstrated by the frictional adhesion of the Tokay gecko, which open new possibilities for adhesive design. Whilst there can be no such thing as generally applicable adhesives, versatility can be massively improved by programmable adhesives, and function in specific areas improved by adopting optimized biomimicked versions of conventional adhesives. Following the example of the bee, the mussel, the gecko and many others, in the words of Fakley: "the designers of the future will have adhesives which do considerably more than just stick".

Looking at Grime through a postcolonial lens

Billy Jordan

Grime is an angry music genre, it takes the aggressive beats of Hip-Hop and introduces lyrics of protest and defiance; but it has evolved to include songs that reflect on home and family. It has also moved from a fringe movement into the mainstream and, consequently, many artists have become highly successful. I have been listening to UK rap for the last two years and, as my appreciation of it has grown, I have expanded my listening tastes from the well-known artists such as Stormzy and Dizzee Rascal to some more remote rappers such as SBK, Novelist and Pa Salieu.

GRIME'S ORIGINS

Grime is a primarily black, English form of rap. Dan Hancox, social commentator, describes its musical origins as 'mostly second and third generation black Britons who were just estranged enough from their cultural roots in the Caribbean, or Africa, or both, and far enough along the lineage of unique British dance styles – acid house, jungle, drum 'n' bass, UK garage – that they could draw from them all.' Unlike British Hip-Hop (the British response to American rap, where people would mimic entirely the accents and mannerisms) Grime is purposefully English, with no attempt to mask the heavy London accent. The accent in Grime is so strong that when Dizzee Rascal toured America in 2004 he had to verbally spell out the title of his album to a radio presenter after being asked to repeat himself three times.

Before Grime there was a movement in deprived areas of London called 'garage', where social housing communities would gather on rooftops and other locations to rap over beats. This was different from the preceding movement of American Hip-Hop: each MC would only have a short amount of time on the mic before passing it on, leading to a much faster tempo with people attempting to create a lasting impression. This was broadcast through pirate radio from changing locations as, with a growing popularity, politicians openly condemned the violence and viscerality portrayed in the verses. This is comparable to the American media's condemnation of the rap group NWA, as they believed it incited violence in black communities such as Compton. In response to this criticism Eazy-E (one of the founding members of NWA) said that the group was just describing how it really was on the street. This is similar to Krept, Konan, Skengdo and AM's defence of Drill (a form of road rap) when they were recently called to parliament: they said that these kids were just rapping the reality of the streets and Drill was a tool to escape this. We have seen through Dizzee Rascal and more recently SL that this is the case and that, despite some of its shortcomings, Drill and Grime has given an escape from poverty to many people, and a sense of identity to those that listen to it.

Early prominent 'garage' artists included Wiley (often considered the godfather of Grime) and Dizzee Rascal who started making an impact from the age of 16. These musicians would describe their anger at the middle-class occupation of their areas. They felt that pubs which were popular meeting areas were starting to charge extortionate amounts for mediocre food, appealing to the growing middle-class population of London, but making them too expensive for the residential population. Many believe that this was, at least indirectly, caused by New Labour's social housing projects, and resulted in the mainly immigrant communities, living in council flats, being unable to move home and disperse as they were not able to afford relocating beyond their community. The deterioration of the social housing communities is reflected in Dizzee Rascal's 2003 track 'Sittin Here' – 'And it's the same old story, benefits claims and cheques in false names. And it's the same old story, students truant learn the streets fluent. Yeah it's the same old story, strange, there's no sign of positive change. 'Cause it was only yesterday we were standing firmly on our feet.' Here Dizzee comments on how, as the areas they live in have become more expensive, young people have resorted to crime and fraud just to live and how this is worsening at an alarming rate.

Looking at early Grime lyrics it is clear that the government's perceived disregard for the poor made people angry enough to express it in music. This anger has remained in the same areas through poverty's perpetuation: children born into poverty have had far less access to good education, and their parents would have had far less time, education and cultural capital themselves to help. Alongside this there have been many reported cases of racism in both further education and employers; as Dave said in his song 'Black' – 'Working twice as hard as the people you know you're better than, because you need to do double what they so you can level them'. While this has begun to change for the better recently, with reports of increased diversity in universities and workplaces, black people and other minorities still feel that in order to achieve the same positions as their white, male counterparts they must significantly surpass them, showing that we still have a long way to go to achieve equality.

The few members of the social housing communities who managed to become rich were not able to gain any sort of social standing, due to our societal fascination with 'old money' over 'new money'. This has resulted in a lack of inspirational figures for people to look up to and to motivate these communities, again perpetuating their poverty. This poverty trap is probably what is lying behind the waves of people beginning to enter the Grime scene through the free music streaming website – Soundcloud – where you can both publish and listen to music. Indeed this is similar to Brazil where many children in poorer areas devote their free time to football training in hopes of replicating the riches of the football players we are exposed to in the media.

As Grime has developed, however, the political message at its core has begun to diminish, with only rappers such as Dave continuing to create political songs. In recent popular Grime songs such as Hardy Caprio and DigDat's 'Guten Tag' and Jay1's 'Mocking It' the lyrics only portray the rapper's superior lifestyle, whether that be their cars, clothes or girlfriends. I believe this to be an expression of insecurity, which could perhaps stem from disenfranchisement and a desire for status. In Britain traditionally status has been derived from inherited wealth for many generations. For most immigrants this is unachievable, leading to a new form of status which is reliant solely on current material possessions over any sort of wealth. This is highlighted blatantly in the American rap group Migos' hit single 'Bad and Boujee' which opens with 'You know so we ain't really never had no old money, we got a whole lotta new money though'.

One value that has persisted throughout Grime, however, is the inflated sense of pride. Rappers would immediately exchange so called 'diss tracks' (songs designed to insult someone) as soon as they were offended in any way. The most notable example of this is Stormzy's song 'Shut Up' which is a three minute irate response to someone commenting that he looked like a backup dancer and, while this is quite recent, we can see the exact same message in Wiley's much older diss track on his former unofficial brother Skepta named 'Flippin Tables'. In both these instances the offence was seemingly inconsequential (in Wiley's case it was as small as creating a song with a rapper he had fallen out with) however, the offended person responds with far more aggression in order to ensure they do not lose any status that might occur from being seen to let someone get away with the insult.

POSTCOLONIAL THEORIES OF THE 'OTHER' AND THE 'EXOTIC'

Postcolonial theory is the study of how former colonies have been affected by their occupation and, in turn, the influences on culture and identity in the UK. Due to Britain's huge empire in the Victorian era, the British culture has been super-imposed upon communities such as Nigeria, Kenya and Ghana. There has also been mass scale immigration to Britain from former colonies. Despite increasing numbers of different nationalities immigrating to England, however, mainstream media has portrayed the image of the British person as predominately white. While this has changed recently, with efforts from large corporations to increase diversity, our historic denial of non-white people being British has had a profound impact on the black communities' identity and comfort in this country.

The 'other' is a postcolonial theory of how a dominant group stigmatises the differences of another group. This was originally used during slavery to justify our treatment of racial minorities: by classifying different groups as others, dominant groups are able to impose views that other groups are 'barbarians' or 'savages', therefore minimising any perceived humanity. Furthermore, the idea that a community is inhabited by a primitive other has proven to be the basis of much of the western desire for control and domination. Even missionaries would attempt to bring God into the hearts of Africa, in order to 'enlighten' the perceived 'savages'. The postcolonial theorist Syed Manzurul Islam pointed out in his book 'The Ethics of Travel' that many 14th Century explorers such as Marco Polo were 'machines of othering', where they would describe any difference in the cultures they came across as morally wrong.

This is heavily linked to another postcolonial theory of the 'exotic', which describes the way in which we have traditionally perceived, and to some extent continue to perceive, many Eastern and African communities as interesting, similar to an attraction in a museum. We can see countless cases of this in history, with one notable example being King Leopold, of Belgium, who brought back war prisoners from the Congo to put in parks as exhibits for the Belgians to see (the Congolese 'exhibits', unused to the cold winter subsequently died). More recently Western cultures have appropriated the exotic words of post-colonial native languages such as namaste and chai. As Nikesh Shukla writes:

'Namaste means hello.

Namaste means I'm bowing to you.

It's a customary greeting.

It's a respectful salutation.

It has become a bastardised metaphor for spiritualism. It's white people doing yoga, throwing up prayer hands chanting "AUM" and saying "namaste" like their third eyes are being opened and they can peer directly into the nucleus of spirituality.'

The main difference between these two concepts, however, is how we perceive them. Whilst exoticising other races can be degrading, and undermines their legitimacy, it is when we look at the other that we feel intimidated, threatened and distinct. I believe that Grime uses this 'other' to protest and to defy becoming the 'exotic.' Indeed, the intent to scare is a part of the protest. Some of Grime's most notable songs such as Skepta's 'Shutdown', Stormzy's 'Big for your boots' and Bugzy Malone's 'Warning' are deliberately designed to be aggressive, almost scaring the listener. Skepta's 'Shutdown' includes an excerpt from a news programme where a posh young female voice describes his show as 'A bunch of young men all dressed in black, dancing extremely aggressively on stage. It made me feel so intimidated and it's just not what I expect to see on primetime TV'. Skepta purposefully includes this in the song as a mockery of the white middle-class which he is proud to have 'intimidated'. It is also notable that the delivery of this line has undertones of racial profiling, where the omission of the descriptor black for the 'young men', while taking offence at the 'black' colour of their clothes, implies that it is actually their skin colour which scares her. Skepta is simultaneously boasting about making his music mainstream and contradictorily mocking his mainstream audience, in order to ensure that people know his music is still aimed at those who have experienced the same hardships. He tells us that we cannot appropriate his music into the mainstream. He will always be the other - distinct and feared.

Similar behavioural traits can still be observed today, showing that these concepts cannot be simply disregarded as history. The French postcolonial artist Kader Attia linked The Western tradition of taxidermy to these themes, stating that this reflects the 'need for control and domination, where wild animals are used as trophies and as evidence of our species' successful mastery' over other species. In 'Measure and Control' he places two stuffed monkeys alongside an African tribal mask to signify how our attitudes towards the two are alarmingly similar, both just interesting objects.

LOOKING AT GRIME IN THE LIGHT OF THESE THEORIES

Another thing I have noticed through my research is how much of a problem racism still is. Dan Hancox says in his grim biography 'Inner City Pressure' – 'young black men have struggled to be treated with dignity in public space in Britain for as long as they have been in public space in Britain'. However, the component of racism that has been so angrily represented throughout Grime's history is that of police discrimination. There have been decades of racial profiling from the police, with 'routine checks' and 'sus laws' (stop and search). Hancox comments that 'this kind of toxic innuendo – the implication of a kind of innate black criminality – reared its head again as late as 1995, when Metropolitan Police chief Paul Condon made controversial high-profile comments linking mugging to young black men specifically'; this theory was immediately debunked by criminologists. It is no wonder that so many songs in Grime ask for their black friends to be freed from prison, for example at the end of Mostack's recent Daily Duppy he says '3 Jugga J, DK, Blade, Mikes, DB, all the mandem' ('3' being slang for free from prison.) The police have historically treated this community so unfairly that the Grime generation have developed a resentment for their authority.

In its early stages UK rap has faced many challenges from the authorities with many artists censored by the police. One of the most notable examples of this is with the drill rapper Digga D who has been banned by the police from rapping as they say his lyrics 'incite violence'. However, despite being in and out of jail frequently, he continues making songs in protest as he considers music his way out of poverty. Before that Giggs, often seen as the person who started 'road rap', was often prevented from doing shows- the main source of income for rappers in this genre- by the police. Given this apparent censorship of certain predominately black communities' music the predominate view of a racist authority seems plausible, contextualising the aggression that people in this genre have towards the police.

In a broader context, one of the more problematic issues that has arisen from slavery, according to Toni Morrison, is the emasculation of men and the ensuing sexism. Male working slaves had no real power, or the independence that we would have normally associated with masculinity. Therefore, in the aftermath of the abolition of slavery black women were generally more submissive, in order to try and rebuild a sense of pride in their partner, which many believe has led to the alarming amounts of sexism in black culture today. We don't need to look much further than a generic rap song to see the way that women are routinely objectified in a way that conflates with a desire for pride as rappers brag about their girlfriend's physical appearance. One notable example of this is Jay1's recent single 'Your Mrs' – 'I'm taking your Mrs, nah brudda, I'm joking… My girl with the back looks lovely' ('back' being slang for her bottom).

As Varaidzo pointed out in her essay 'A Guide to Being Black' there are a multitude of factors that must be adapted for black people to immerse themselves in British culture and not be seen as other. Even the natural African 'afro' hairstyle can be othered as 'the hairstyle for black radicals', and to fit in black women must have their hair weaved painfully tight just to align themselves with the media representation of how black women should look. She comments that the majority of rap songs can draw a whole party's attention to the only black people present through the use of 'That Which Cannot Be Spoken' (the n word). When Kanye West's 'Gold Digger' comes on, for example, the entire room's attention would be drawn to the only black people in the room to observe whether they would sing along with the n word. Varaidzo points out that white kids singing along outside the company of any black people is 'a tree falling in a forest conundrum… is it still racist?'. Just the existence of such a racially charged word alerts people to race, therefore, its continual use in Grime can be seen as another instance of attempting to stand out as other - to own their blackness unapologetically.

Rappers have also adapted the English language for themselves, in a similar way to American 'ebonics' – American black English that is regarded as a language in its own right. Words have been adapted and, in some cases, even invented in a way that aligns English with Caribbean and African languages. Examples of this include 'ting' (thing), 'man' (someone else or yourself) and 'yute' (kid). In creating a new form of English Grime MCs have managed to leave their own mark on English in a similar way to how Shakespeare invented words. In the face of disenfranchisement the Grime generation have created their own culture, refusing to conform and instead choosing proud defiance at being treated as other.

The problem for so many of the Grime generation, such as Loyle Carner, is that they don't have a defined culture of their own. Loyle Carner said that as mixed race he was neither accepted by the white kids at school nor the black kids. This is very similar to Stuart Hall's concept of the 'Familiar Stranger' where immigrants would often form bonds with each other as they both had a common trait of loneliness and disenfranchisement. Grime has given many unrepresented individuals the ability to group together and start their own community, where they are able to control how they themselves are represented. They are able to bring shape, identity and politics to the other that they represent.

Overall, I believe that Grime has been ground-breaking in merging three very different cultures – colonial African, Caribbean and British - whilst giving young black men a representation of themselves in the media. We saw recently with Stormzy's number one single 'Vossi Bop' that the mainstream media has started to take Grime seriously, which in turn has resulted in greater black representation in the media. Despite this, we have seen plenty of downsides where some rappers seem to glamorise violence and drugs; and normalise sexism. In my opinion, Grime is the community's defiant celebration of themselves as 'other': a community that does not want to fit in with a culture they feel has routinely forgotten them.

How to build a jet engine in your shed

ILA STEM Winner 2019

Yuntian Fang

INTRODUCTION

The main problems with modern jet engines is the enormous stresses and heats that the internals of the engine experience. The parts must withstand pressures of up to 140psi and heats of 1100°C with the metals used in construction melting at around 1300°C for long periods sometimes to the tens of hours all the while operating reliably.

This mix of very high-end performance and reliability results in commercial jet engines costing exorbitant amounts with one jet engine in the 747 costing $12.2million in 2015[1].

However, with the rise of hacker/DIY culture, there have been many designs for working small gas-powered turbine engines using off-the-shelf components that are easily attainable. In addition, 3D printing which was once very expensive and usually restricted to large companies has now become cheap enough that it is possible to purchase a high-quality printer for under £200 and even cheaper if you buy a kit and assemble it yourself. This allows individuals to create high precision parts such as the front compressor and aerodynamic surfaces.

DESIGN CALCULATIONS:

TURBINE SHAFT (1)

For small model turbine engines, the shaft is the part that poses the greatest challenge as well as being the main limiting factor. This problem is unavoidable due to the effect of gravity as at a critical speed, even gravity will cause a perfectly balanced shaft to begin to oscillate. This can cause catastrophic problems if the shaft begins to oscillate at its first natural frequency which will quickly lead it to its destruction as well as of the entire engine.

(See Figure 1.1)

By balancing the shaft on callipers, I have determined that the centre of gravity of the shaft is 105.25mm from the tip with the shaft (overall 203mm) weighing 125.9g.

Firstly, we calculate the second moment area which describes how mass is distributed along a arbitrary axis (how even a shape is).

For calculations, the largest diameter is 11mm and the smallest is 6mm. However, the mean diameter come to 9.82mm by finding the area of the shaft then dividing by its length.

$$I = \tfrac{1}{4}\,\pi \cdot r^4 \qquad (1.2)$$

$$I = 4.56 \times 10^{-10}\ m^4$$

Now the stiffness of the shaft can be calculated. The young's modulus of silver steel is around $2.07\text{x}10^{11}$Pa. (Taken from source 1.1)

$$C_q = \frac{3 \cdot E \cdot I \cdot l}{a^2 \cdot b^2} \qquad (1.3a)$$

$$C_q = 5.437 \times 10^5\ kgs^{-1}$$

a = from tip to center of mass

l = total length = $a + b$

Finally, the critical speed is calculated.

$$n_c = \sqrt{\frac{C_q}{m}} \qquad (1.3b)$$

$$n_c = 2106s^{-1} = 126000rpm$$

The critical speed also varies with other design variables such as shaft unbalance, length and diameter of shaft and bearing support. Also due to it being very hard to determine without damaging the shaft or with no access to specialised equipment. Thus, to keep it safe, I will set a maximum running speed of 75% its resonance or 94krpm or $1572s^{-1}$. Limited by this, we can carry on the calculations.

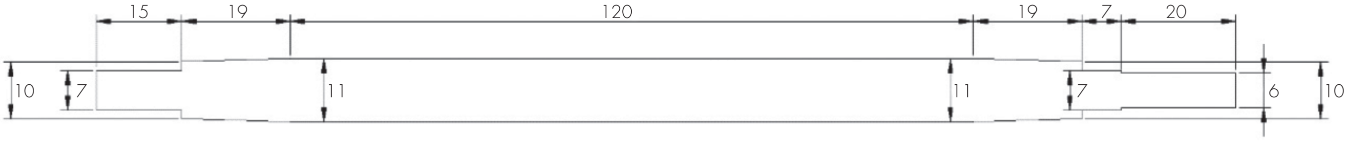


Figure 1.1: Diagram of the turbine shaft

TURBINE FANS (2)

The second major limiting factor are the fans at the back of the engine used to extract energy from the heated air. This is due to the immense centrifugal forces and heat they experience.

Firstly, we consider the maximum peripheral speed that they can spin at. Looking at the graph, we can see that the strength of the steel decreases with temperature. Looking at figure 2.1, we can find a estimate for the ultimate tensile stress that stainless 304 steel can take at around 500°C.

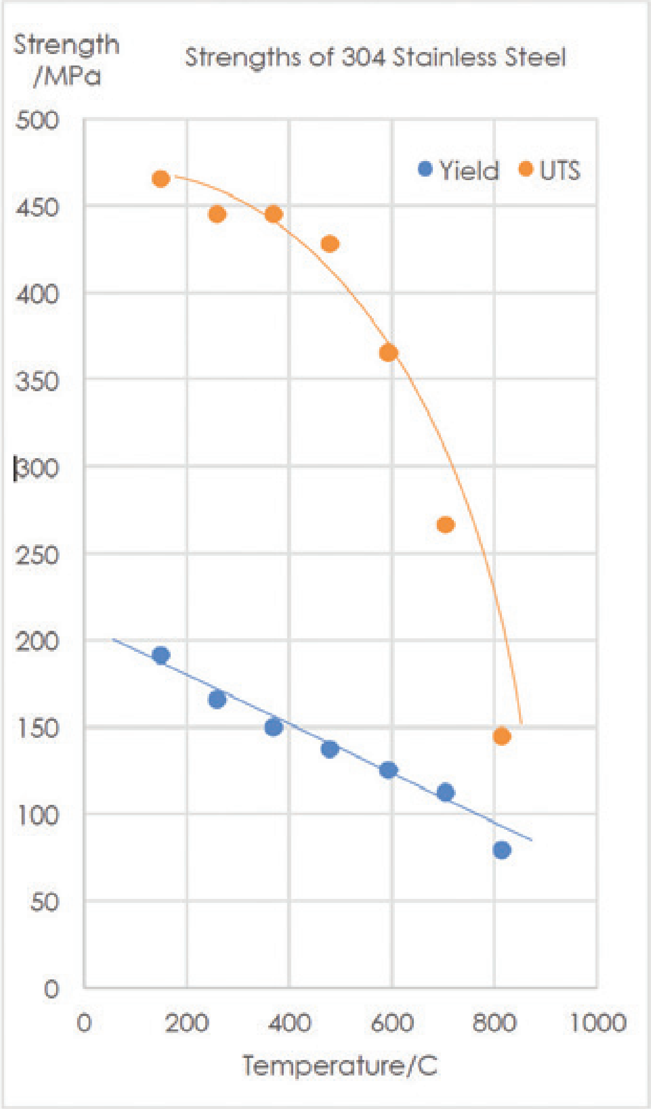


Figure 2.1: Graph showing the various strengths of stainless steel at different temperatures. (Based on data taken from source 2.1).

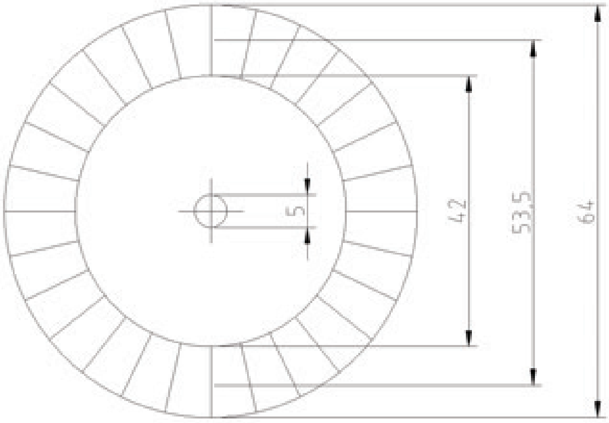


Figure 2.2: Diagram of the turbine wheel. Dimensions in mm.

Now, to calculate the maximum speed that the turbine can turn at is described in this equation from source 2.2.

$$u_{max} = \sqrt{\frac{3\sigma \cdot r_{blade}^{2}}{p(r_{blade}^{2} + r_{blade}\, r_{bore} + r_{bore}^{2})}} \qquad (2.2)$$

$u_{max} = 217.7ms^{-1}$

p = *density of stainless steel* $8000kgm^{-3}$

σ = *elastic limit at* 480ºC = $137MPa$

Now to convert it into rotations per second.

$$n_{max} = \frac{u_{max}}{d_{outer} \cdot \pi}$$

$n_{max} = 1083s^{-1} = 65krpm$

Turns out the limiting factor was the turbine wheel and not the main shaft. By using this speed, the mean airspeed at the back of the turbine can then be calculated.

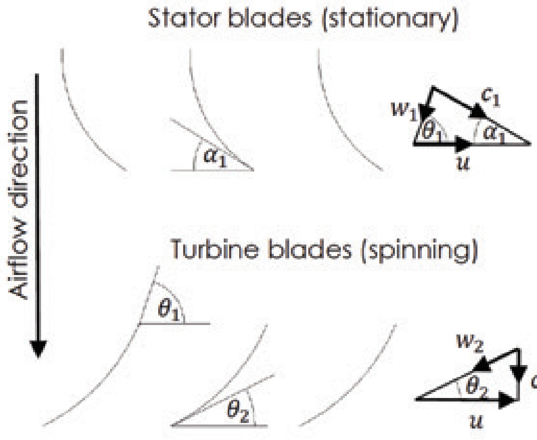


Figure 2.3: Diagram showing the blade angles and speed triangles of the fans at the back of the engine. (Reproduced from page 44 of source 2.3)

$u_{mean} = 182.0ms^{-1}$

w = *speed of airflow referanced to the blade*

c = *speed as seen by a outside observer*

u = *peripheral speed of blades*

From these blades, we can draw triangles that represent the vector flow of air. The vectors W show the direction the air moves in from our point of view. Vectors C show the airflow as seen by the spinning fan blades.

The purpose of the diffuser (stationary) is to redirect the airflow so the maximum amount of energy can be extracted to spin the turbine.

Now, we must find the optimum blade angle of θ_2. To do this, we can firstly find the relationship between blade angle, temperature and thrust.

Firstly, we can calculate the exhaust velocity c_2.

$$c_2 = u_{mean} \cdot tan\theta_2$$

The final pressure change is also evident from the law of pressures.

$$p_{exhaust} = p_{normal} \cdot \frac{T_{ambient}}{T_{exhaust}}$$

Finally, the mass flow is needed.

$$\dot{m} = \frac{\pi\,(d_{blades}{}^2 - d_{inner}{}^2\,)}{4} \cdot c_2 \cdot p_{exhaust}$$

Now, the thrust can be found.

$$F = \frac{\pi\,(d_{blades}{}^2 - d_{inner}{}^2\,)}{4} \cdot u_{mean}{}^2 \cdot tan^2\theta_2 \cdot p_{exhaust}$$

For temperature, we have to perform a two-part process which also takes into account the pressure changes.

POWER AND PRESSURE (3)

The source of power in a turbine engine is when the gas from the diffuser section is deflected by the turbine section, creating a moment on the shaft.

$$P_{shaft} = \dot{m} \cdot u_{mean}{}^2$$

From this, we know how much power the compressor can use for compression as described in this equation taken from page 46 of source 2.3.

$$P_{shaft} \cdot \eta_V = \dot{m} \cdot C_p \cdot T_1\,(\lambda^{\left(\frac{\gamma-1}{\gamma}\right)} - 1)$$

η_V = *efficiency of compressor stage* ~70%

$\dot{m}$ = *mass flow / kgs*$^{-1}$

C_p = *specific heat of air* 1000$Jkg^{-1}\ K^{-1}$

T_1 = *temperature of inlet air* 293 / K

λ = *overall pressure ratio*

γ = *ratio of specific heats* ~1.4

We can substitute $P = \dot{m} \cdot u_{mean}{}^2$ equation from before into this new one.

$$\lambda = \frac{u_{mean}{}^2 \cdot \eta_V}{C_p \cdot T_1} + 1^{\left(\frac{\gamma}{\gamma-1}\right)}$$

$\lambda = 1.24bar$

This means that the pressure directly after the compressor wheel will be 1.24bar or 17.9psi.

From this, we can also calculate the final exhaust temperature using conservation of energy with this equation taken from page 47 of source 2.3.

$$P_{in} = C_p \cdot T_{exhaust}\left(1 - \frac{1}{\lambda^{\left(\frac{\gamma-1}{\gamma}\right)}}\right) \cdot \dot{m}$$

C_p = *specific heat of air* 1000$Jkg^{-1}\ K^{-1}$

λ = *overall pressure ratio*

γ =*ratio of specific heats* ~1.4

$$T_{exhaust} = \frac{u_{mean}{}^2\left(1 + \frac{tan^2\theta_2}{2}\right)}{C_p \cdot \eta_T \cdot \left(1 - \frac{1}{\frac{u_{mean}{}^2 \cdot \eta_V}{C_p \cdot T_1} + 1}\right)}$$

Now, we can plot the graph described before.

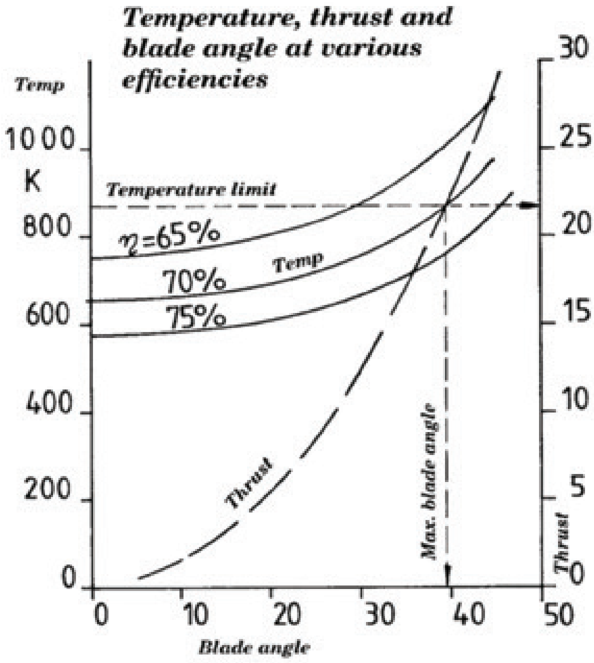


Figure 3.1: Graph showing the relationship between exhaust temperature, thrust , efficiency of the systems (η) and blade angle. Taken from page 47 of source 2.3.

Looking at the graph, the blade angle that maximises thrust is around $\theta_2 = 39$. This allows us to complete the previously mentioned calculations.

$$c_2 = u_{mean} \cdot tan\theta_2$$

$c_2 = 147.4\ ms^{-1}$

Temperature is now calculated before thrust.

$T_{exhaust} = 829K = 556^oC$

Now for the exhaust gas pressure.

$$p_{exhaust} = p_{normal} \cdot \frac{T_{ambient}}{T_{exhaust}}$$

$p_{exhaust} = 0.435\ kgm^{-3}$

$$F = \frac{\pi\,(d_{blades}{}^2 - d_{inner}{}^2\,)}{4} \cdot u_{mean}{}^2 \cdot tan^2\theta_2 \cdot p_{exhaust}$$

$F = 17.3N$

CLEARANCES, HEAT AND FUEL (4)

Due to the high temperatures involved as well as the stresses, we must also factor in the fact that the turbine wheel will expand when running.

$$\Delta l = l \cdot \Delta T \cdot \propto$$

$$\propto \ = \textit{thermal extansion coefficient}$$

Stainless steel's coefficient is 1.73x10-5. (Data taken from source 4.1). Assuming operating at 556°C.

$$\Delta diameter = 64 \cdot (556 - 15) \cdot 1.73 \times 10^{-5}$$

$$\Delta diameter = 0.599\text{mm}$$

This thermal expansion is also accounted for since the housing for the blades will also expand but at a different temperature (~500°C).

$$\Delta radius = 65 \cdot (500 - 15) \cdot 1.72 \times 10^{-5}$$

$$\Delta radius = 0.545\text{mm}$$

Additionally, the blades also experience centrifugal force which will also cause expansion. However, this is insignificant as is shown. Data for the 304 stainless steel taken from source 4.2. This is done again using the equation from source 2.2.

$$\sigma = p\frac{u_{max}{}^2}{r_{blade}{}^2} \cdot \frac{(r_{blade}{}^2 + r_{blade}r_{bore} + r_{bore}{}^2)}{3} \qquad (2.2)$$

$$\Delta l = \frac{p \cdot u_{max}{}^2 \cdot (r_{blade}{}^2 + r_{blade}r_{bore} + r_{bore}{}^2) \cdot l}{3E \cdot r_{blade}{}^2}$$

$$\Delta l = 2.19 \times 10^{-5}\,\text{m} = 0.0219\text{mm } \textit{(insignificant)}$$

To estimate fuel consumption, it is relatively simple by using specific heat capacity.

$$Q = mc\Delta T \quad thus \quad \dot{Q} = \dot{m} \cdot C_p \cdot \Delta T$$

$$C_p = \textit{specific heat of air } 1000Jkg^{-1}\,K^{-1}$$

$$\dot{Q} = 0.104 \cdot 1000 \cdot 536$$

$$\dot{Q} = 62900W$$

Since the fuel source is propane which releases 50.33MJkg-1 of heat, (from source 4.3) we can calculate how much propane we need.

$$\dot{Q} = E \cdot m_{\dot{propane}}$$

$$E = \textit{energy released per kg of burnt propane}$$

$$m_{\dot{propane}} = \frac{\dot{Q}}{E}$$

$$\textit{Propane needed} = 1.25 \times 10^{-3}\ kgs^{-1} = 1.25gs^{-1}$$

Or the engine can run at full throttle for a few minutes on a can that should be used for camping.

COMPRESSOR SHAPE (5)

Firstly, we must consider what type of shape the compressor will be. Either an axial one (many fans in series), an impellor which consists of a flat disk with fins on it or something in-between.

To decide, we must firstly calculate the running factor of the compressor from the equation taken from page 51 of source 2.3. This allows us to find the shape of compressor that will give us the highest efficiency.

$$\sigma = \frac{2n\sqrt{\frac{\dot{m}}{\lambda \cdot p_{normal}}\pi}}{(2 \cdot u_{mean}{}^2)^{3/4}} \qquad (2.3)$$

$$\sigma = 0.252$$

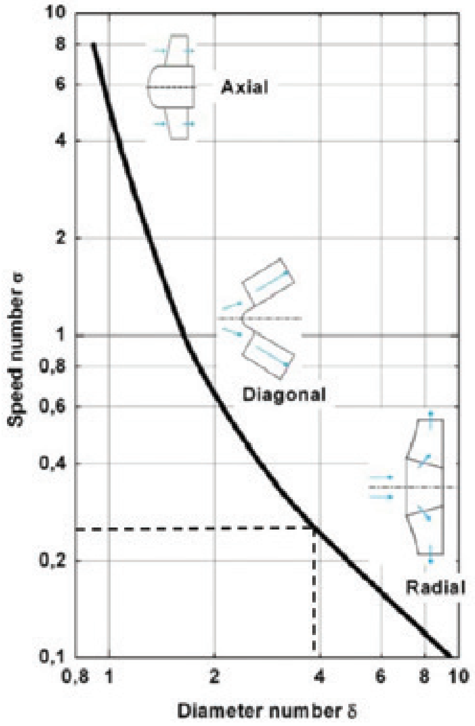


Figure 5.1: A graph showing the relationship between the speed and diameter number for a compressor. Taken from page 355 of source 5.1

From the graph above (Cordier diagram), we can then calculate the optimum dimeter for the fan from the equation taken from page 354 of source 5.1.

$$D = \frac{2\delta}{\sqrt[4]{\dfrac{2 \cdot u_{mean}{}^2 \cdot \pi^2}{\left(\dfrac{\dot{m}}{\lambda \cdot p_{normal}}\right)^2}}} \qquad (5.1)$$

$$D = 0.0743\text{m} = 74.3\text{mm}$$

$$\delta = \textit{diameter number}$$

This means that the compressor shape will be somewhere in-between a diagonal and radial compressor around 74mm in diameter.

Again, we turn to case studies that are documented in page 54 to page 55 in source 2.3 that have shown that oversizing the compressor can cause many problems. One case had a compressor with a diameter 90% of the turbine wheel. This engine ran but had a very high exhaust temperature, stopping it from being used at full throttle.

Another compressor with a diameter of 85% of the turbine's ran very smoothly with an exhaust temperature below 500°C but with reduced thrust.

From it is important to not oversize the compressor. Thus I will be using a compressor wheel 90% of the turbine diameter (compressor diameter = 58mm)

COMPRESSOR DESIGN (6)

Symbol	Formula	Used
D_2	58mm	58mm
D_1	0.5 D_2	29mm
B_1	0.2 D_2	11.6mm
B_2	0.1 D_2	5.8mm
θ_1	34^0	34^0
θ_2	45^0	45^0
R	$\frac{D_2^{\,2} - D_1^{\,2}}{4\,(D_2 \cdot cos\theta_2 - D_1 \cdot cos\theta_1)}$	37.2mm
ρ	$\sqrt{R^2 + \frac{D_1^{\,2}}{4} - R \cdot D_1 \cdot cos\theta_1}$	26.4mm

Figure 6.1: A table showing the rough calculations for all the dimensions in figure 6.2. Replicated from page 54 of source 2.3.

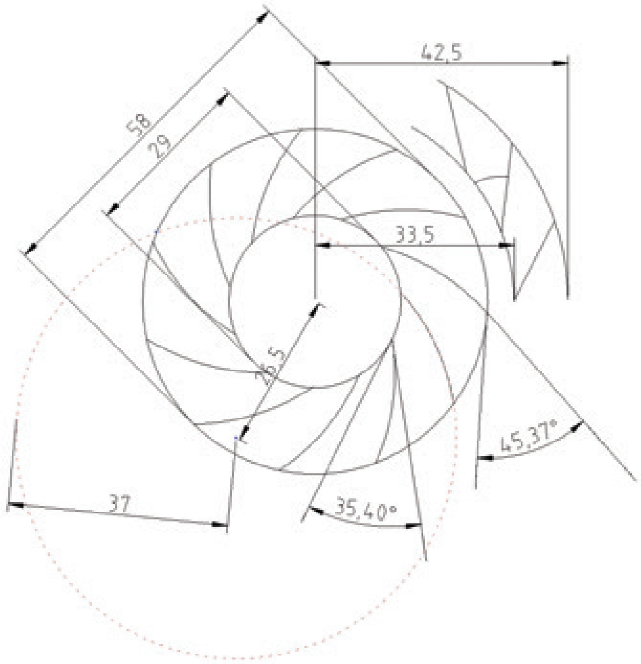


Figure 6.2: Diagram showing the used dimensions in designing the front stator section (front view). Replicated from page 52 of source 2.3.

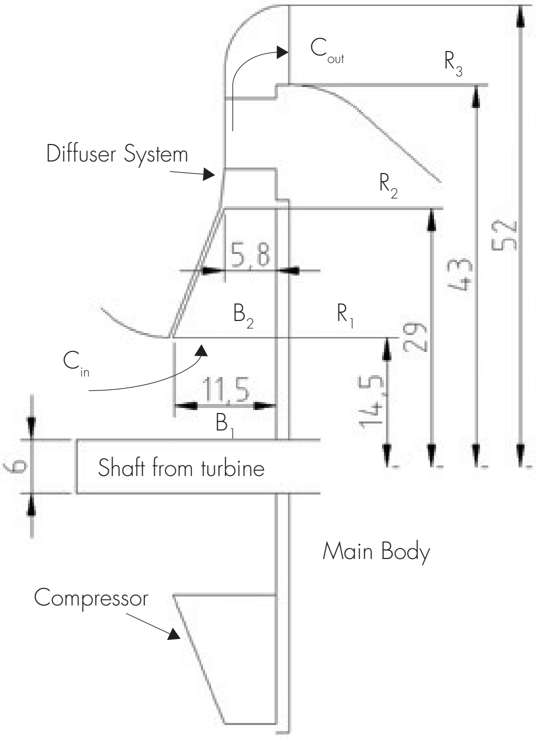


Figure 6.3: Diagram showing the used dimensions in designing the front stator section (side view). Replicated from page 52 of source 2.3.

COMPRESSOR WHEEL STRENGTH (7)

Now, we need to calculate the stress acting on the blades of the compressor wheel. Equation taken from page 56 of source 2.3. The yield point of PLA used to print the compressor is around 35.9MPa as stated in source 7.1. However, this will vary with other factors such as printing temperature, infill type and percentage, layer height and others while printing. The value for density is also taken from source 7.1.

$$\sigma_{inlet} = \frac{p \cdot u_1^{\,2} \cdot B_1^{\,2} \cdot cos\theta_1}{D_1 \cdot s_{inlet}} \qquad (2.3)$$

Rearranged, we can calculate the blade thickness needed at the speeds expected.

$$ss_{inlet} = \frac{2 \cdot p \cdot R_1 \cdot \pi^2 \cdot n_{max}^{\;\;2} \cdot B_1^{\,2} \cdot cos\theta_1}{\sigma}$$

$$s_{inlet} = 1.373\text{mm}$$

$$s_{outlet} = \frac{2 \cdot p \cdot R_2 \cdot \pi^2 \cdot n_{max}^{\;\;2} \cdot B_2^{\,2} \cdot cos\theta_2}{\sigma}$$

$$s_{outlet} = 0.591\text{mm}$$

Additionally, we need to consider the maximum stress the compressor well will experience as approximated in this equation taken from page 56 of source 2.3.

$$n_{max} = \sqrt{\frac{\sigma_{max}}{0.83 \cdot p \cdot 4 \cdot R_2^{\,2} \cdot \pi^2}}$$

$$n_{max} = 987s^{-1} = 7780Krpm$$

This shows that the strength of the compressor wheel does not need to be considered since it is so much higher than all the other limiting factors.

COMBUSTION CHAMBER (8)

The purpose of this is to slow down the incoming fast-moving air so that it does not blow out the flame. This is due to the finite speed of flame propagation. This is achieved by placing three sets of holes that restrict airflow using the equations described in source 8.1.

$$A_{first} \approx A_{inlet} \cdot 0.35$$

$$A_{second} \approx A_{inlet} \cdot 0.6$$

$$A_{third} \approx A_{inlet} \cdot 1.5$$

$$A_{inlet} = \pi \cdot \frac{D_1^2}{4} = 661\text{mm}^2$$

$$A_{first} = 231\text{mm}^2 \quad A_{second} = 397\text{mm}^2 \quad A_{third} = 992\text{mm}^2$$

Using a 3.5mm drill bit for the first set, we need to make 24 holes.

For the second row we need 18 holes using a 5.5mm drill.

The last row of holes uses a 7.5mm drill bit resulting in 22 holes needed.

SUMMARY

Due to the fact that much of aerodynamics is about minimising losses and increasing efficiencies in a system, it is very hard to determine if a engine design will be successful and begin to self-sustain.

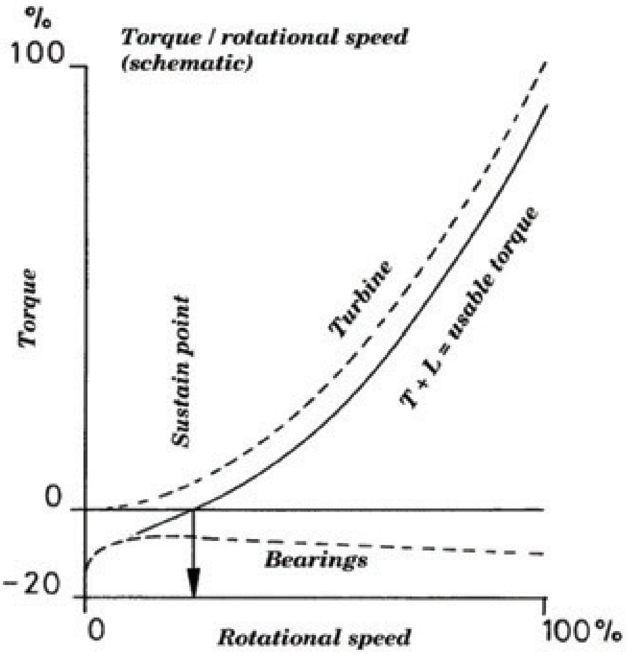


Figure 9.1: A graph showing the relationship between the resistance in the bearings and the torque from the turbine. Taken from p98 of source 2.3.

As seen in figure 9.1, the bearings provide a mostly constant friction which the torque from the turbine must overcome. At the self-sustain point where the torque is equal to the rolling resistance, the turbine does not need external power and can now run unaided.

This rotational speed for an engine similar size is at 8000rpm (page 97 of source 2.3), around 11% of its maximum rotational speed (75000rpm). Correctly lubricated, the rolling resistance will decrease, further lowering the self-sustain speed needed.

Finally, due to inevitable inaccuracies with the manufacturing of these parts, it even harder to see if it will work and can only be seen by running a full test. However, I have been able to spin the turbine shaft up to speeds of 11000rpm and thus I have confidence it will self sustain.

BUILDING THE ENGINE

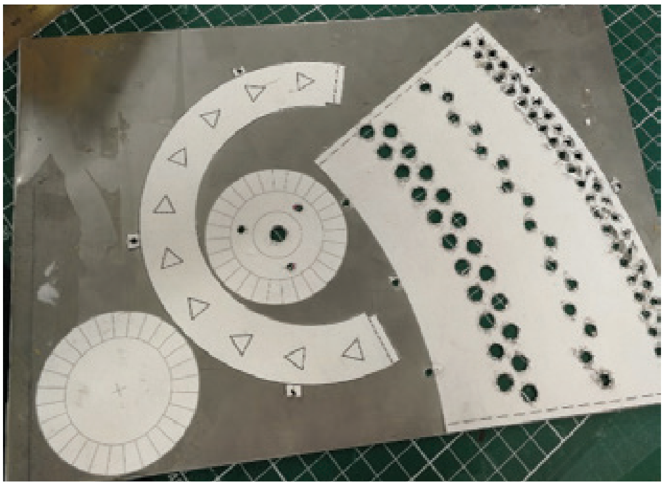


Figure 10.1: Paper templates are printed and stuck onto 0.5mm thick 304 stainless steel sheet. Holes are drilled according to previous calculations.

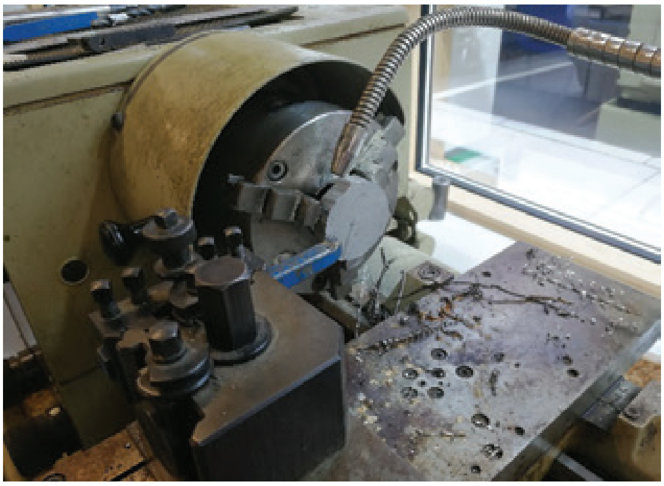


Figure 10.2: Structural component being turned on the lathe out of 12mm thick aluminium.

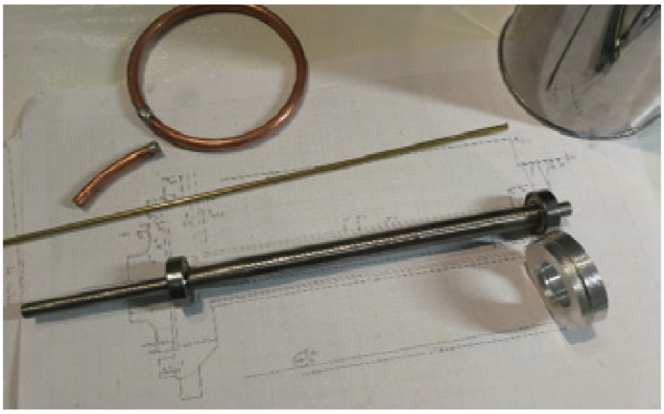


Figure 10.3: Main turbine shaft is also turned on the lathe. Also see the copper combustion, the brass lubricant tube, and the finished structural component.

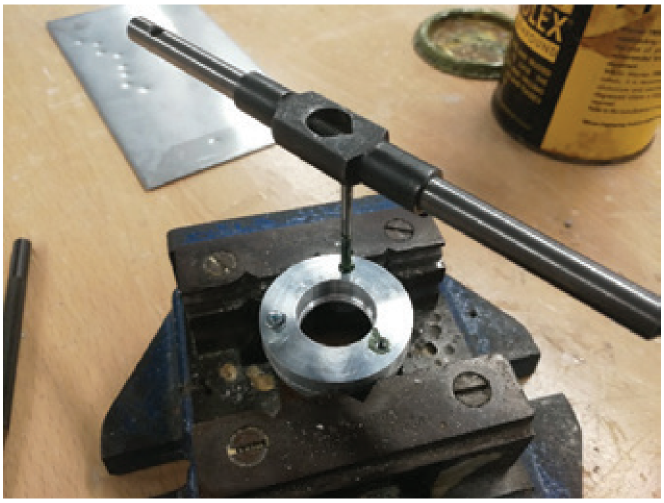


Figure 10.4: Tapping the structural component to allow screws to be inserted.

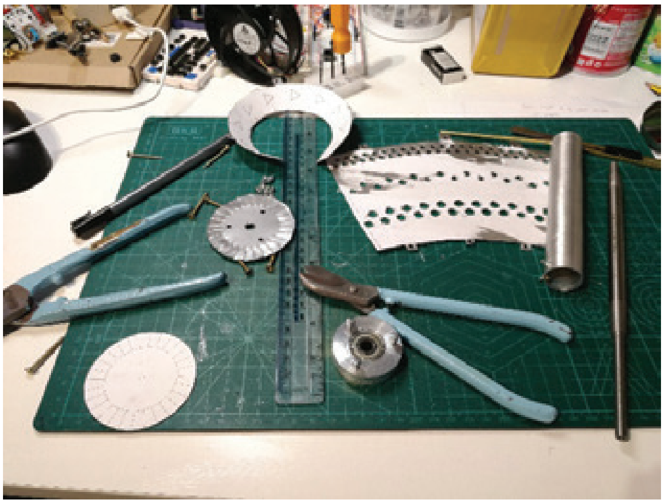


Figure 10.5: Cutting out the sheets from figure 10.1. See bottom middle for the finished structural component and right middle for the structural tube.

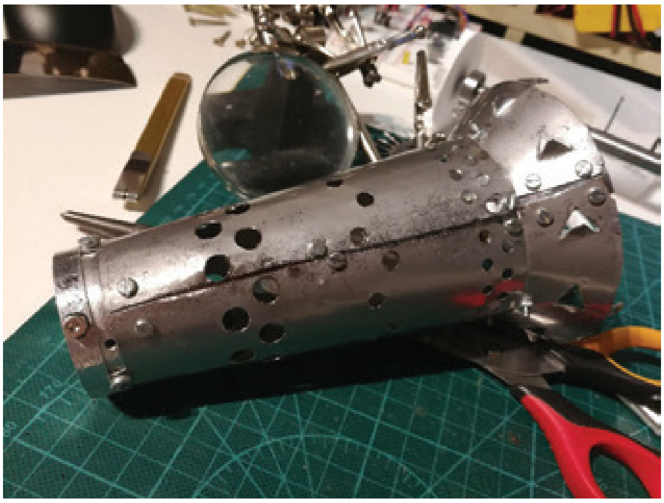


Figure 10.6: Previously cut sheets are bent, screwed together and attached to the structural component (see the left end)

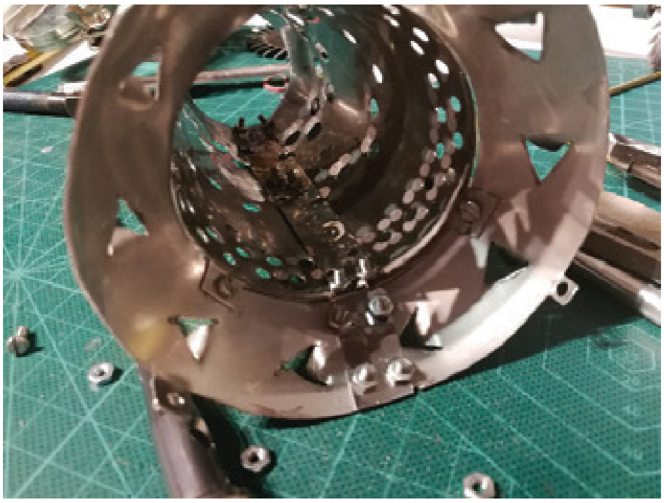


Figure 10.7: A additional strip of metal is used to cover up the gap between the sheet and is used to mount screws to.

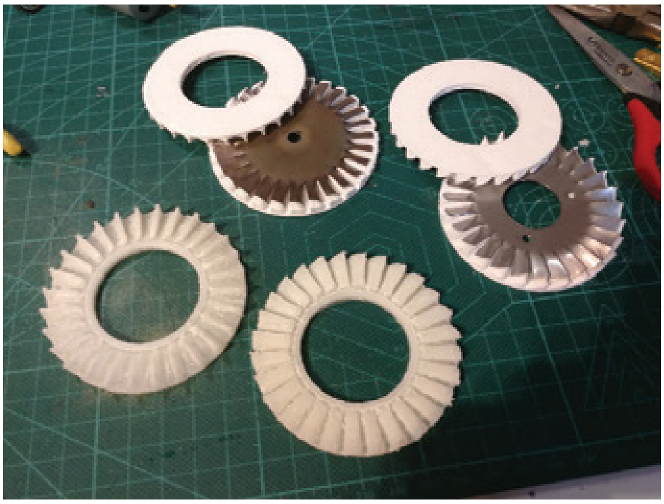


Figure 10.8: Cut out turbine wheels are placed in 3D printed presses to ensure constant geometry.

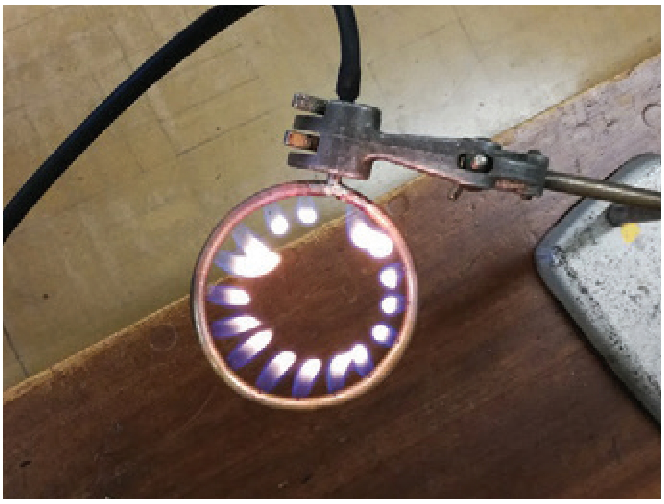


Figure 10.9: Copper combustion ring is now silver soldered (instead of lead soldered) to withstand the high temperatures and is tested to ensure that the flame holes are similar in size to each other.

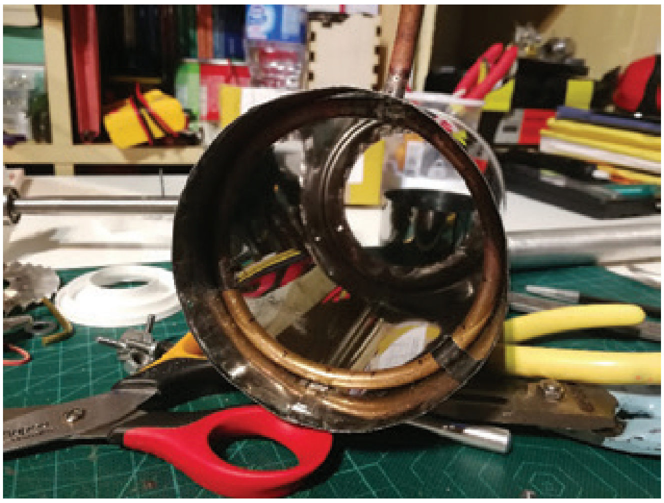


Figure 10.10: Copper mounted inside the main body by some strips of stainless.

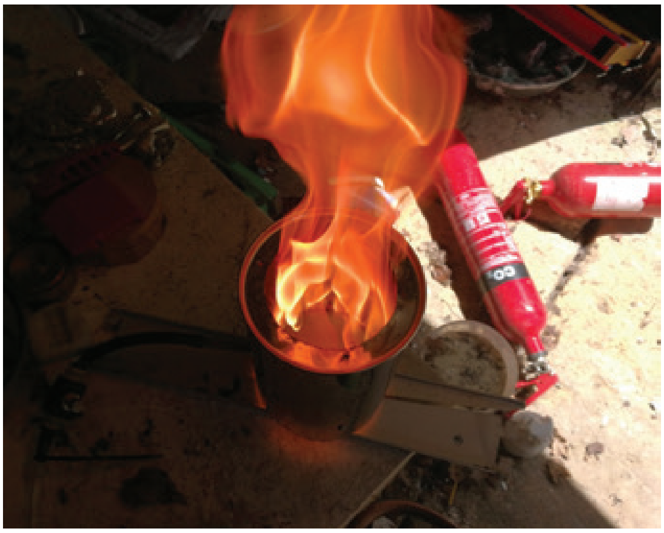


Figure 10.11: Combustion ring tested again while mounted inside body.

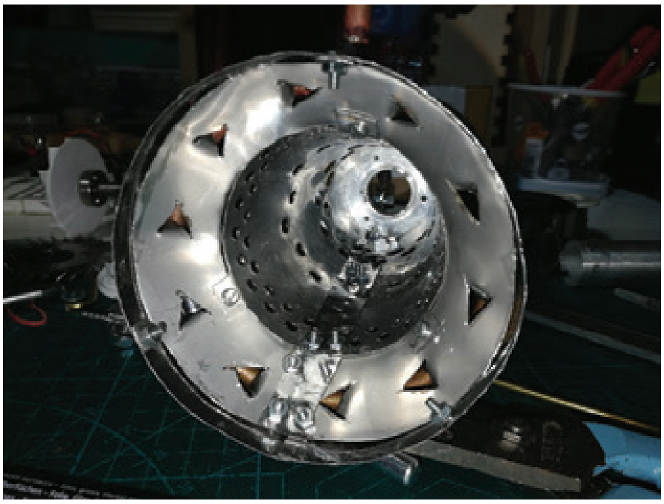


Figure 10.12: Combustion lining is also mounted with three screws around the side. (see combustion tube behind).

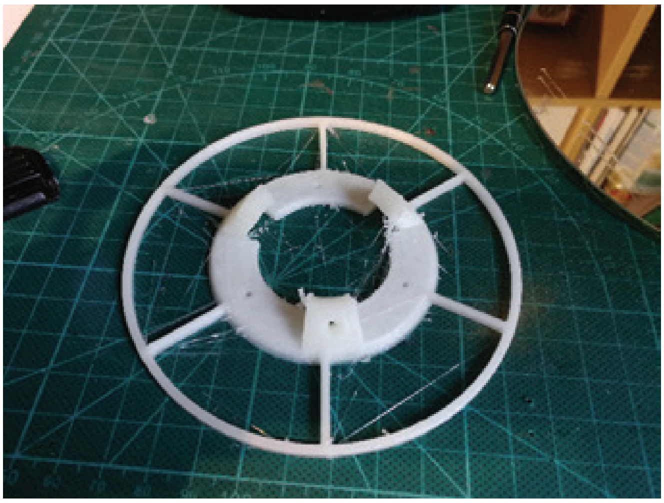


Figure 10.13: 3D printing of various parts. This one locates the front assembly and holds the radiation shield that reflects some of the heat away from the parts.

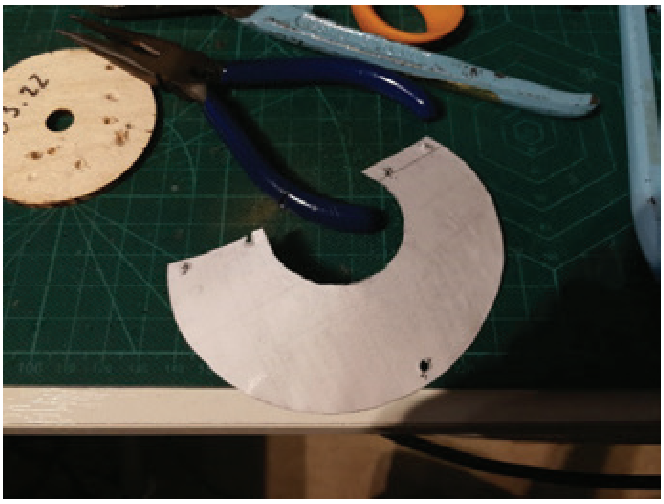


Figure 10.14: The radiation shield cut from a piece of an aluminium can.

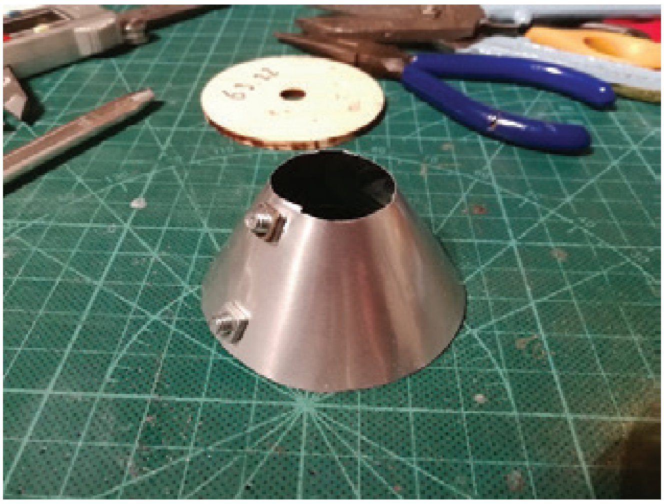


Figure 10.15: Shield bent and bolted together.

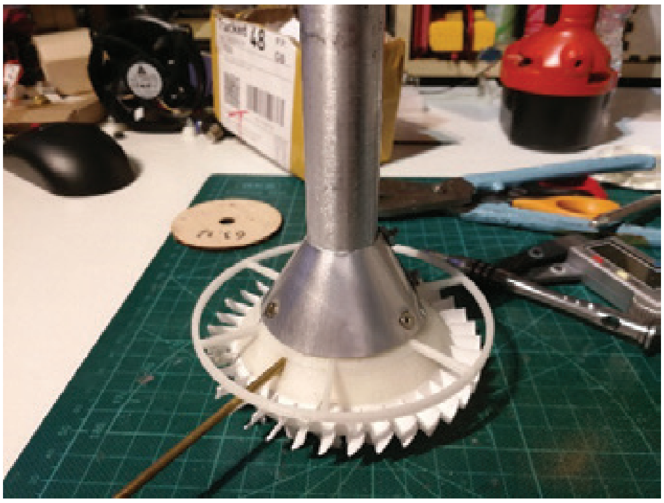


Figure 10.16: Structural tube and shield assembled with the 3d printed parts.

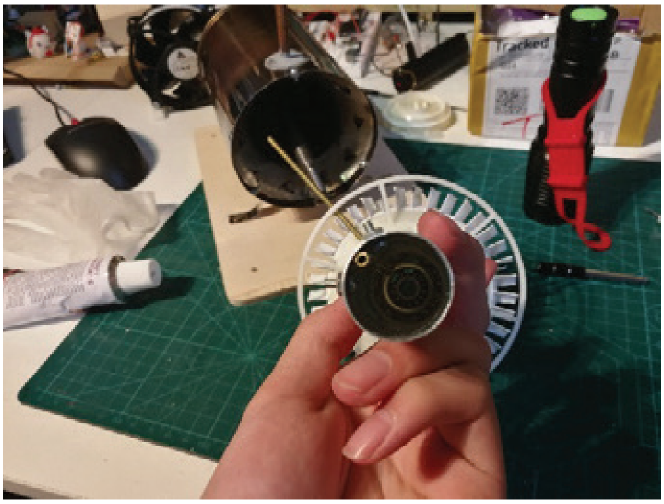


Figure 10.17: Back of the structural tube. See the brass cooling tube.

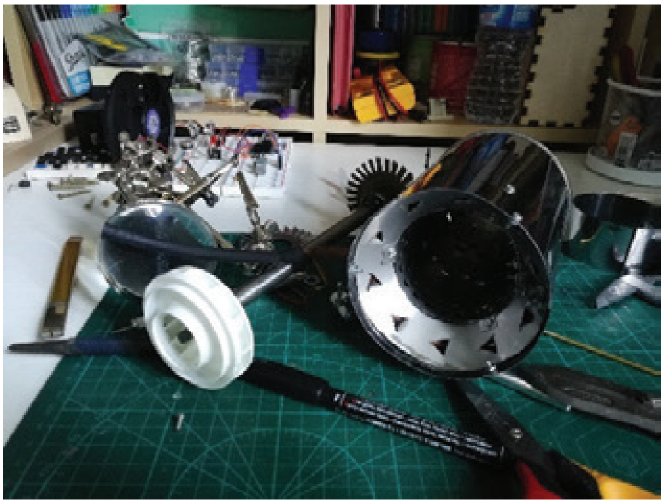


Figure 10.18: 3D printed compressor secured to the main shaft and reinforced with epoxy resin.

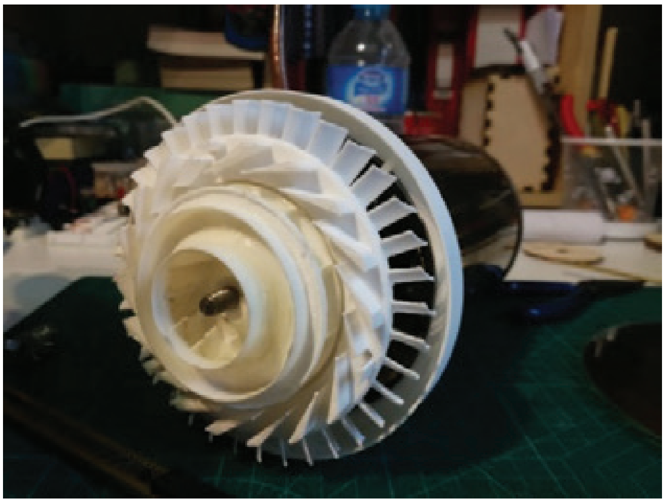


Figure 10.19: Assembled with the cover taken off.

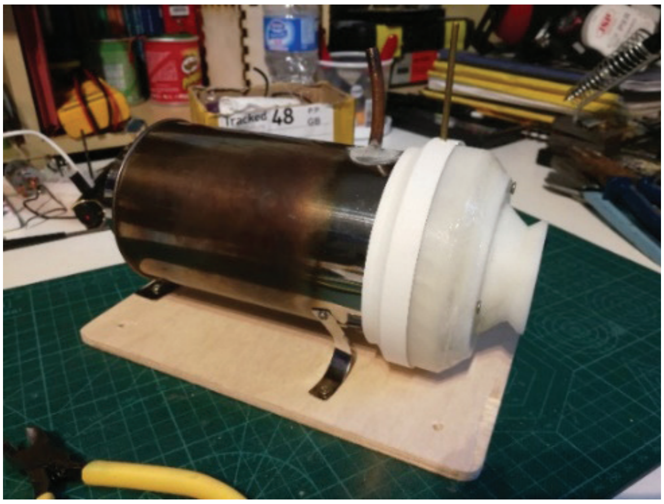


Figure 10.20: Full assembled and bolted down to a plywood base. See the colours of the outer body that resulted from the flame test in figure10.11.

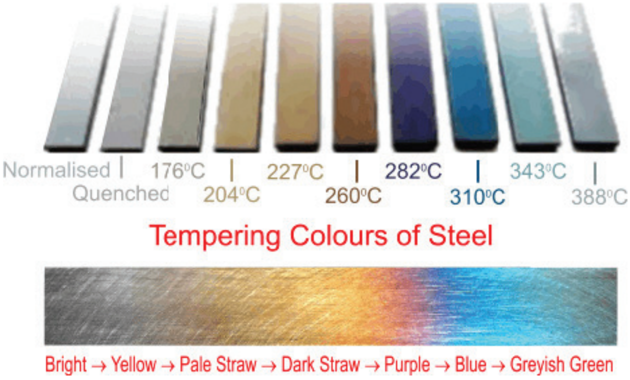


Figure 10.21: Image showing the temperature that steel turns into after being heated. Taken from source 10.1. This shows that the hottest part of the outer body only gets to around 260oC and that area around the 3d printed parts to not get excessively hot.

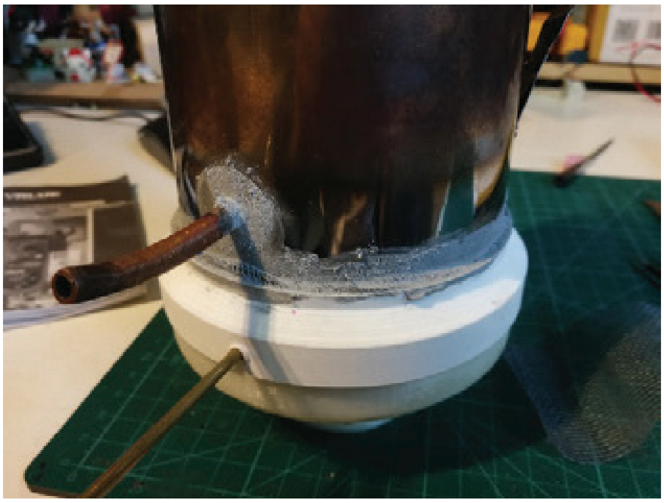


Figure 10.22: Silicone sealant is used to seal the distance between the parts.

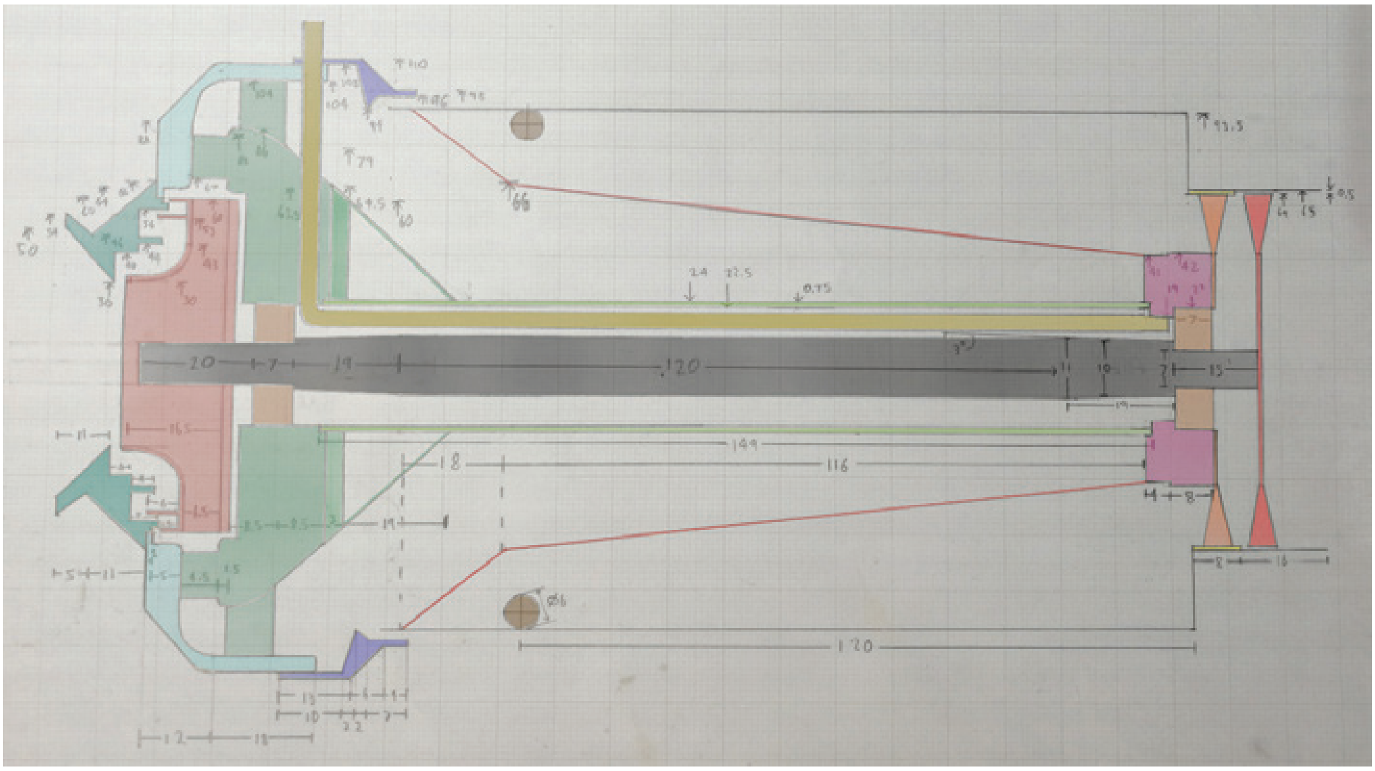


Figure 10.23: Drawn diagram that shows all the measurements for the engine.

TEST FIRE 1 OF THE MINI GAS TURBINE ENGINE:

TEST FIRE 1

As seen here, the 3d printed parts got very hot and started to deform. This was since my start-up procedure was wrong. Initially, I had started let the gas flow, lit the fire and then spun it up. Due to this mistake, it would have resulted in the entire engine filling with propane and thus when ignited would have caused a fire at the 3D printed parts before the air drawn in could direct the fire to the correct place. This caused the front cover to deform a bit however it was possible to reheat it and bend it back into shape.

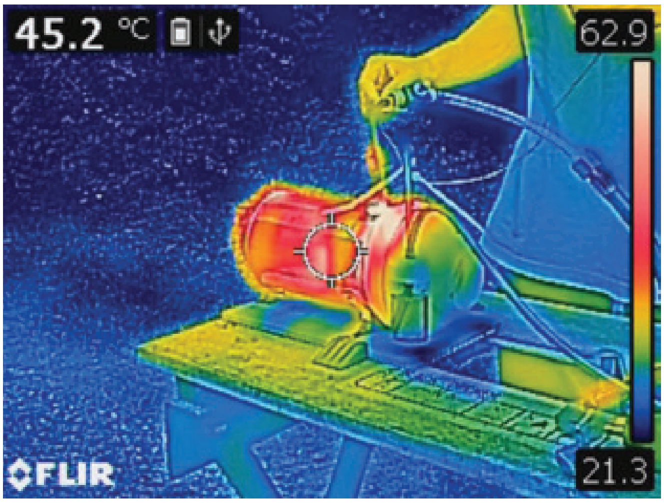


Figure 1.1: Thermal image of the engine after the first failed test fire. Notice how there is a bright white spot at the top of the engine at the printed parts. This resulted some damage.

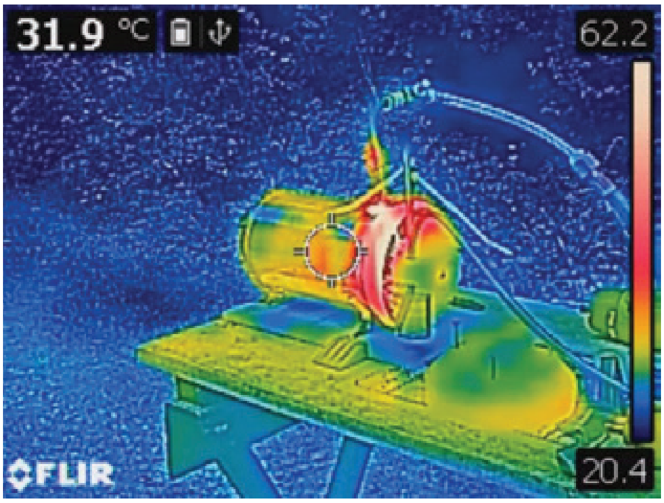


Figure 1.2: Thermal image with the cover taken off. The printed parts seem to be the hottest par

Unfortunately, the locating ring made of thin parts had deformed too much and I could not reheat it. Luckily, this part was not necessary for the operation of the engine and was only used in the process of balancing the main shaft.

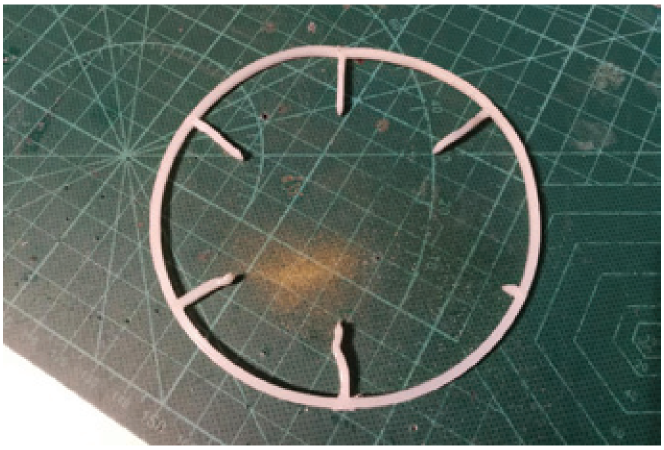


Figure 1.3: The locating ring that has deformed due to the weight of the system resting on this can to be cut from the system to continue testing.

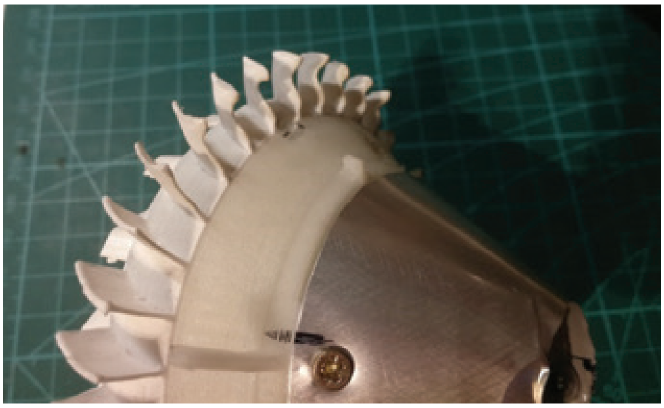


Figure 1.4: The printed guide vanes have deformed slightly/been burn by the flames.

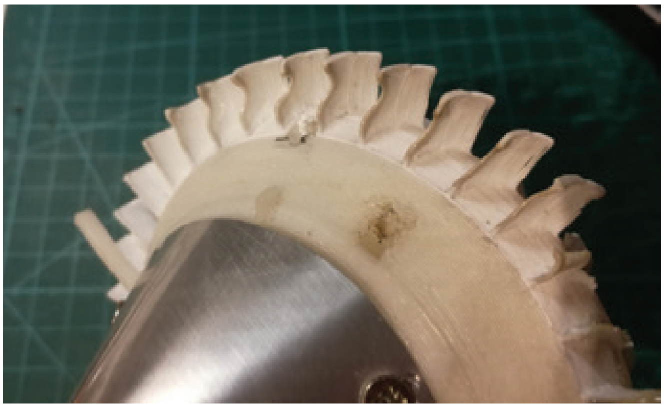


Figure 1.5: Other view of the guide vanes that slightly melted.

After removing the locating ring, further tests were performed. However, it was quickly discovered that each time, the flame would gradually decrease in size as the engine span up. This was almost definitely due to the compressor causing the pressure differential between the inside of the engine and the supply from the regulator to decrease.

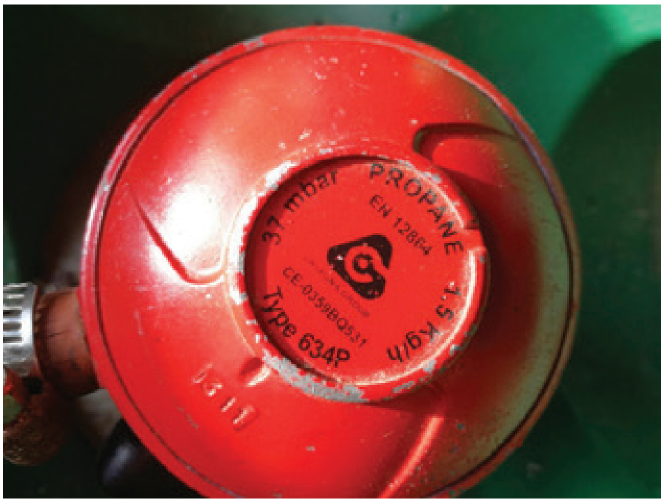


Figure 1.6: Image of the regulator used. Output at 1.037bar and a max flow rate of 1.5kg/h.

Looking back at the formulae, we can rearrange it to find the speed at which the pressure inside of the engine should equal the pressure coming from the regulator.

$$n_{max} = \frac{\sqrt{\frac{C_p \cdot T_1 (\lambda^{(\frac{\gamma-1}{\gamma})} - 1)}{\eta_V}}}{d_{middle} \cdot \pi}$$

$$n_{max} = 396.9s^{-1} = 23.8Krpm$$

As evident, it shows that at these speeds, the fuel flow would be greatly decreased where the fuel supply should be increased as the engine increases in rotational speed.

If we assume the max flow is always at 1.5kg/h regardless of pressure, it still would not be enough for the engine which requires a max fuel flow rate of 4.5kg/h, around three times what can be supplied.

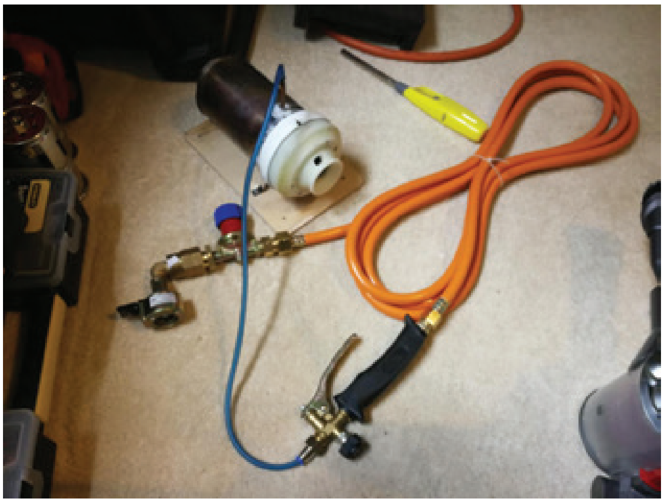


Figure 1.7: New regulator system. Rated up to 4bar and fuel flow of 8kg/h.

In addition, as seen in the video, it was very hard to get the shaft to spin at a constant speed either using the dremel or the leaf blower which only had on and off control. Thus, I added a PWM module to the leaf blower so I could perform fine adjustments to the windspeed.

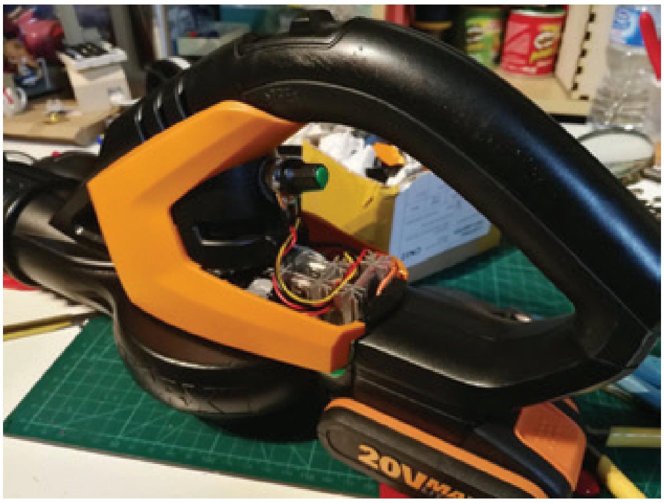


Figure 1.8: Image of the new leaf blower with a spare PWM module installed in-line with the original styling.

The tachometer was also modified so it could be toggled on or off using a slide switch instead of a momentary push button which had to be held down. Another modification was that its battery capacity was increased to 15.5Wh compared to 4.5Wh as stated in source XX with the 9V battery.

Figure 1.9: Modified tachometer mounted on a tripod.

TEST FIRE 2 OF THE MINI GAS TURBINE ENGINE:

TEST FIRE 2

After replacing the fuel supply, I ran into problems with the main shaft just not spinning fast enough. This led to me checking the bearings again after reducing all areas of friction notably the rear grub screw used to prevent the entire shaft assembly sliding backwards. After removing the bearings, I found that they were very dirty and left small metal flakes after cleaning, indicating that the ball bearings must have ground into the races causing additional friction.

We can find the speed limit for bearings using the following equations taken from source 1. These figures are assuming that the bearings are in static oil lubrication with the limit is mainly based on thermal limits. This is because the higher the speed, the higher the frictional losses and thus temperature. Figures obtained my measuring the maximum speed at which bearings can be continuously operated without failing from seizure or excessive heat causing seizing.

$$N = \frac{S_{LF}}{D_m}$$

N = *speed limit/rpm*

S_{LF} = *Speed limit factor read off table xx*

Bearing Type	Speed Limit Factor		
	Narrow	Wide	2 Row
Radial Bearings			
Ball, Deep Groove	500,000	-	400,000
Ball, Angular Contact	450,000	-	400,000
Cylindrical, 2 Piece Brass Cage	550,000	500,000	475,000
2 Piece Steel Cage	450,000	435,000	380,000
Stamped Steel Cage	330,000	300,000	-
1 Piece Brass Cage	600,000	420,000	-
Full Complement	170,000	120,000	140,000
Eng Ring Cage	80,000	60,000	60,000
Tampered Roller, Pin Type Cage	400,000	350,000	300,000
Brass, Land Riding Cage	450,000	420,000	400,000
Spherical, Brass Finger Cage	220,000	200,000	-

Table 1: Speed limiting factors for various bearings. Ball bearings used are narrow deep groove bearings. Taken from source 1.

$$N = \frac{S_{LF}}{\left(\frac{D_{outer} + D_{inner}}{2}\right)}$$

$$N = \frac{2 \cdot S_{LF}}{D_{outer} + D_{inner}}$$

N = *34500rpm*

As seen, the limiting speed is consistent with what the datasheet for the bearings state (33300rpm) as seen in source 2. Unfortunately, this is a lot below what is required however, by changing lubrication methods higher speeds can be obtained.

This involves a factor called the correction factor which comes into play if we use forced circulation oil lubrication, jet lubrication or oil mist lubrication. This allows the bearing to spin at much higher speeds as the lubricants are used to cool it down.

Bearing Type	Correction Factor
Cylindrical Roller Brgs. (single row)	2
Needle Roller Brgs. (except broad width)	2
Tapered Roller Brgs.	2
Spherical Roller Brgs.	1.5
Deep Groove Roller Brgs.	2.5
Angular Contact Roller Brgs. (except matched bearings)	1.5

Table 2: Correction factors for various bearings. Taken from source 3.

This means that the maximum speed that can be achieved is raised by a factor of 2.5 as the engine uses deep groove bearings. Thus, the maximum speed is 86200rpm which is enough.

SELF-SUSTAIN SPEED

In order to see if I could find the self-sustaining speed of the engine, I turned to source 4 for guidance.

The friction from the bearings can be estimated using various formulae with the moments from friction being split apart into:

1. Rolling friction moment
2. Sliding friction moment
3. Frictional moment of the bearing seals
4. Frictional moment of drag losses, churning, splashing ect.

Described mathematically on page two of source 4.

$$M_T = M_{rr} + M_{sl} + M_{seal} + M_{drag}$$

Firstly, we can eliminate M_{seal} as no seals will be used to minimise friction and to allot jet oil cooling.

As many of these depend a lot on the rotational speed the bearing is operating at, I will not use actual numbers until the very end.

ROLLING FRICTION MOMENT

Now we can start to calculate $M_{rr'}$ the rolling friction.

$$M_{rr} = \Phi_{ish}\, \Phi_{rs}\, G_{rr}\, (\nu \cdot n)^{0.6}$$

Φ_{ish} = *inlet shear heating reduction factor*

Φ_{rs} = *kinematic replenishment/starvation reduction factor*

G_{rr} = *rolling resistance function*

ν = *viscocity of oil/mm^2 s^{-1}*

Starting off with $\Phi_{ish'}$ it describes the reverse flow that the lubricant experiences as a ball bearing rolls past it. This reverse flow shears the lubricant, generating heat.

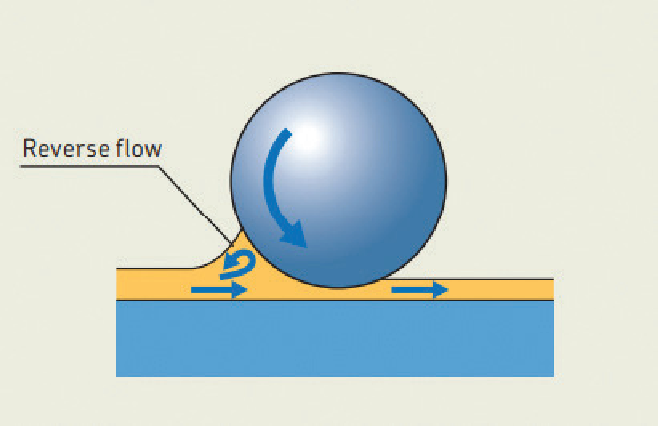


Figure 1: Taken from source 4, it shows the reverse flow the ball bearing makes.

This can be estimated using the following equation.

$$\Phi_{ish} = \frac{1}{1 + 1.84 \times 10^{-9} \cdot (n \cdot d_m)^{1.28}\ \nu^{0.64})}$$

n = *roational speed/rpm*

d_m =*mean diameter of bearing/mm*

For $\Phi_{rs'}$ it describes the additional friction caused by excess or insufficient lubricant being applied to the bearings, changing the thickness of the lubricant, affecting friction. Estimated using the following equation.

$$\Phi_{rs} = \frac{1}{e^{K_{rs} \cdot \nu \cdot n \cdot (d+D) \cdot \sqrt{\frac{K_z}{2\,(D-d)}}}}$$

K_{rs} = *replenishment/starvation constant:*

$= 3 \times 10^{-8}$ *for low level oil bath and oil jet*

$= 3 \times 10^{-8}$ *for grease and oil jet*

n = *roational speed/rpm*

d = *inner diameter of bearing/mm*

D = *outer diameter of bearing/mm*

K_z = *bearing geometric constant read off table* 3

Geometric constraints K_Z and K_L		
Bearing Type	Geometric constraints	
	K_Z	K_L
Deep groove ball bearings - single and double row	 3,1	 -
Angular contact ball bearings - single - double row - four-point contact	 4,4 3,1 3.,1	 - - -
Self-aligning ball bearings	4,8	-
Cylindrical roller bearings - with a cage - full complement	 5,1 6,2	 0,65 0,7
Tapered roller bearings	6	0,7
Spherical roller bearings	5,5	0,8
CARB toroidal roller bearings - with a cage - full complement	 5,3 6	 0,8 0,75
Thrust ball bearings	3,8	-
Cylindrical roller thrust bearings	4,4	0,43
Spherical roller thrust bearings	5,6	0,58[1)]

Table 3: Taken from page 14 of source 4, it describes the K_z values for various types of bearings. Bearings used are deep groove ball bearings.

Next, to calculate $G_{rr'}$ we can use this equation of there is no axial load taken from page 6 of source 4.

$$G_{rr_{no\ axial\ load}} = R_1 \cdot d_m^{\ 1.96} \cdot F_r^{\ 0.54}$$

R_1 = *geometric constant read off table* 4

F_r = *Radial load/N*

Bearing Type	Geometric constraints for rolling frictional moments	
	R_1	R_2
2, 3	4,4 x 10^{-7}	1,7
42, 43	5,4 x 10^{-7}	0.96
60, 630 62, 622 63, 623	4,1 x 10^{-7} 3,9 x 10^{-7} 3,7 x 10^{-7}	1,7 1,7 1,7
64 160, 161 617, 618, 628, 637, 638	3,6 x 10^{-7} 4,3 x 10^{-7} 4,7 x 10^{-7}	1,7 1,7 1,7
619, 639	4,3 x 10^{-7}	1,7

Table 4: Taken from page 8 of source 4, it describes the rolling constants for various bearings used. The 627 bearing is used.

However, if axial load exists, we need to perform a correction to account for it which changes the equation.

$$G_{rr} = R_1 \cdot d_m^{\ 1.96} \cdot \left(\frac{F_r + R_2}{sin \propto_f}\ F_a\right)^{0.54}$$

F_r = *radial load/N*

F_a = *axial load/N*

$$\propto_f = 24.6 \cdot \left(\frac{F_a}{C_0}\right)^{0.24}\ [°]$$

C_0 = *basic static load rating/N*

BEARING SEAL MOMENT

Despite not needing to consider seal friction, we can still model it using the following equations taken from page 11 of source 4.

$M_{seal} = K_{S1} \cdot d_s^{\ \beta} + K_{S2}$

K_{S1} = *constant*

K_{S2} = *constant*

d_s = *seal couterface diameter/mm*

β = *exponent read off table* 5

Seal Type	β	K_{S1}	K_{S2}	d_s
RSL	2.25	0.0018	0	D
RSH	2.25	0.028	2	D

Table 5: replicated from page 11 of source 4, it shows the different constants for different seal types.

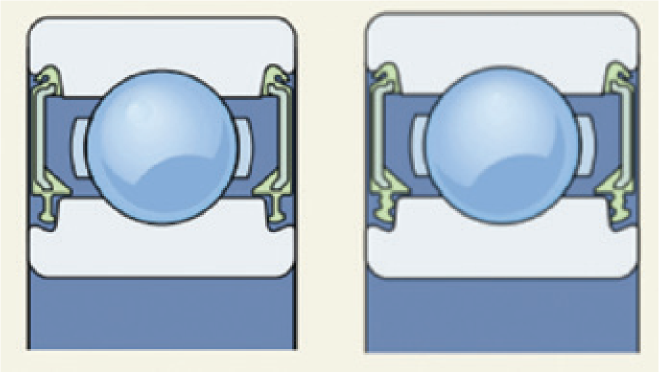


Figure 2: Showing the RSL seal on the left and the RSH seal on the right. Notice the way the seal attaches to the inner race differently. RSL seals are suited to high speeds however are not optimised for reducing water or particle ingress.

RSH seals create more friction due to their different attaching methods, offering more protection against particles at the cost of higher friction. Images taken and edited from page 4 of source 5.

SLIDING FRICTIONAL MOMENT

Next the sliding frictional moment can be estimated by this equation as seen in page 5 of source 4.

$M_{sl} = G_{sl} \cdot \mu_{sl}$

G_{sl} = *sliding friction function*

μ_{sl} = *sliding friction coefficient*

The calculation for G_{sl} is very similar to the one for G_{rr}, with them both having a correction needed for scenarios with axial load. The equation for G_{sl} with axial load is thus described.

$$G_{sl} = S_1 \cdot d_m^{-0.145} \cdot \left(\frac{F_r^5 + S_2 \cdot d_m^{1.5}}{\sin \propto_f} \; F_a^4 \right)^{1/3}$$

S_1 and S_2 are both geometric constants read off table 6

Bearing Type	Sliding frictional moments	
	S_1	S_2
2, 3	2,00 x 10^{-3}	100
42, 43	3,00 x 10^{-3}	40
60, 630 62, 622 63, 623	3,73 x 10^{-3} 3,23 x 10^{-3} 2,84 x 10^{-3}	14,6 36,5 92,8
64 160, 161 617, 618, 628, 637, 638	2,43 x 10^{-3} 4,63 x 10^{-3} 6,50 x 10^{-3}	198 4,25 0,78
619, 639	4,75 x 10^{-3}	3,6

Table 6: Taken from page 8 of source 4, it describes the sliding constants for various bearings used. The 627 bearing is used.

We also need to calculate the sliding friction coefficient with the equation in page 5 of source 4.

$\mu_{sl} = \Phi_{bl} \, \mu_{bl} + (1 - \Phi_{bl}) \, \mu_{EHL}$

μ_{bl} = *constant of movement*

= 0.12 *when the shaft is already spinning*

= 0.15 *for non spinning shaft.*

Used for starting torque calculation

μ_{EHL} = *sliding friction coefficient read off table* 7

$$\Phi_{bl} = \frac{1}{e^{2.6 \times 10^{-8} \cdot (n \cdot \nu)^{1.4} \cdot d_m}}$$

Lubricant used	μ_{EHL}
Cylindrical roller bearing	0.02
Tapered roller bearing	0.002
Mineral oils	0.05
Synthetic oils	0.04
Transmission fluids	0.1

Table 7: Replicated from page 5 of source 4. Describes the various coefficients for special bearings and certain oils. For our purposes, synthetic oils are used.

DRAG LOSSES

This section includes various additional drags such the resistance the oil poses on the rolling ball bearings, external oil agitation and oil viscosity which are all described in the following equations from page 12 of source 4.

$$M_{drag} = 0.4 V_m \, k_{ball} \, d_m^5 \, n^2$$

$$+ \, 1.093 \times 10^{-7} \, n^2 \, d_m^3 \left(\frac{n d_m^2 f_t}{\nu} \right)^{-1.379} R_s$$

V_m = *drag loss factor read off graph* 1

ν = *lubricant viscocity/mm² s⁻¹*

$$k_{ball} = \frac{i_{rw} \, K_z \, (d+D)}{D - d} \times 10^{-12}$$

i_{rw} = *number of ball rows*

K_z = *geometric constant read off table* 8

Bearing Type	Geometric constraints	
	K_Z	K_L
Deep groove ball bearings - single and double row	3,1	-
Angular contact ball bearings - single - double row - four-point contact	4,4 3,1 3.,1	- - -

Table 8: taken from page 14 of source 4, it describes the constant that is used for the deep groove ball bearings.

$$f_t = \begin{bmatrix} \sin(0.5t) & \text{when } 0 \leq t \leq \pi \\ 1 & \text{when } \pi < t < 2\pi \end{bmatrix}$$

$$t = 2\cos^{-1}\left(\frac{0.6d_m - H}{0.6d_m}\right)$$

H = *height of ball bearing/mm*

$$R_s = 0.36_{dm}^{\ 2}\,(t - \sin t)\,f_A$$

To calculate $V_{m'}$ the drag loss factor we can model it using the oil bath model, with the bearing being semi submerged in oil.

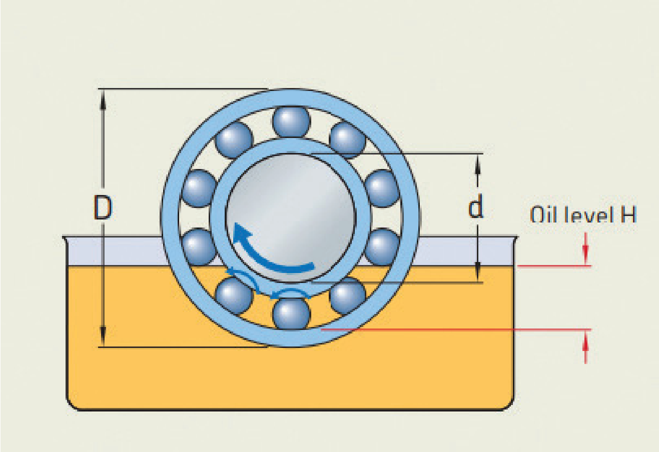
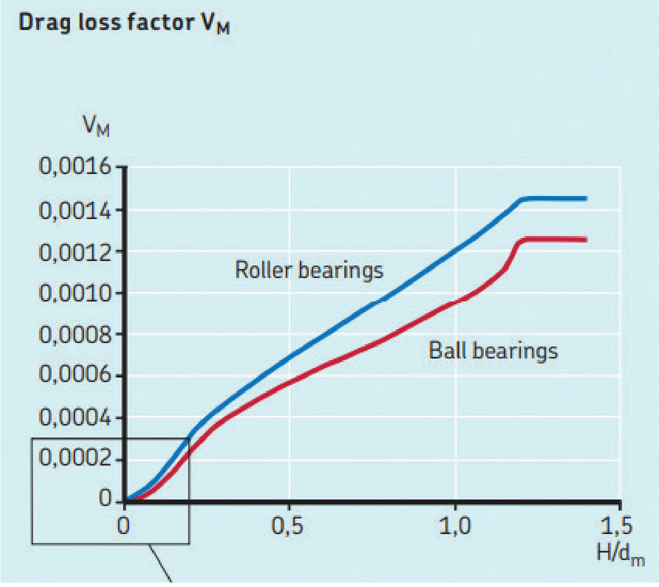
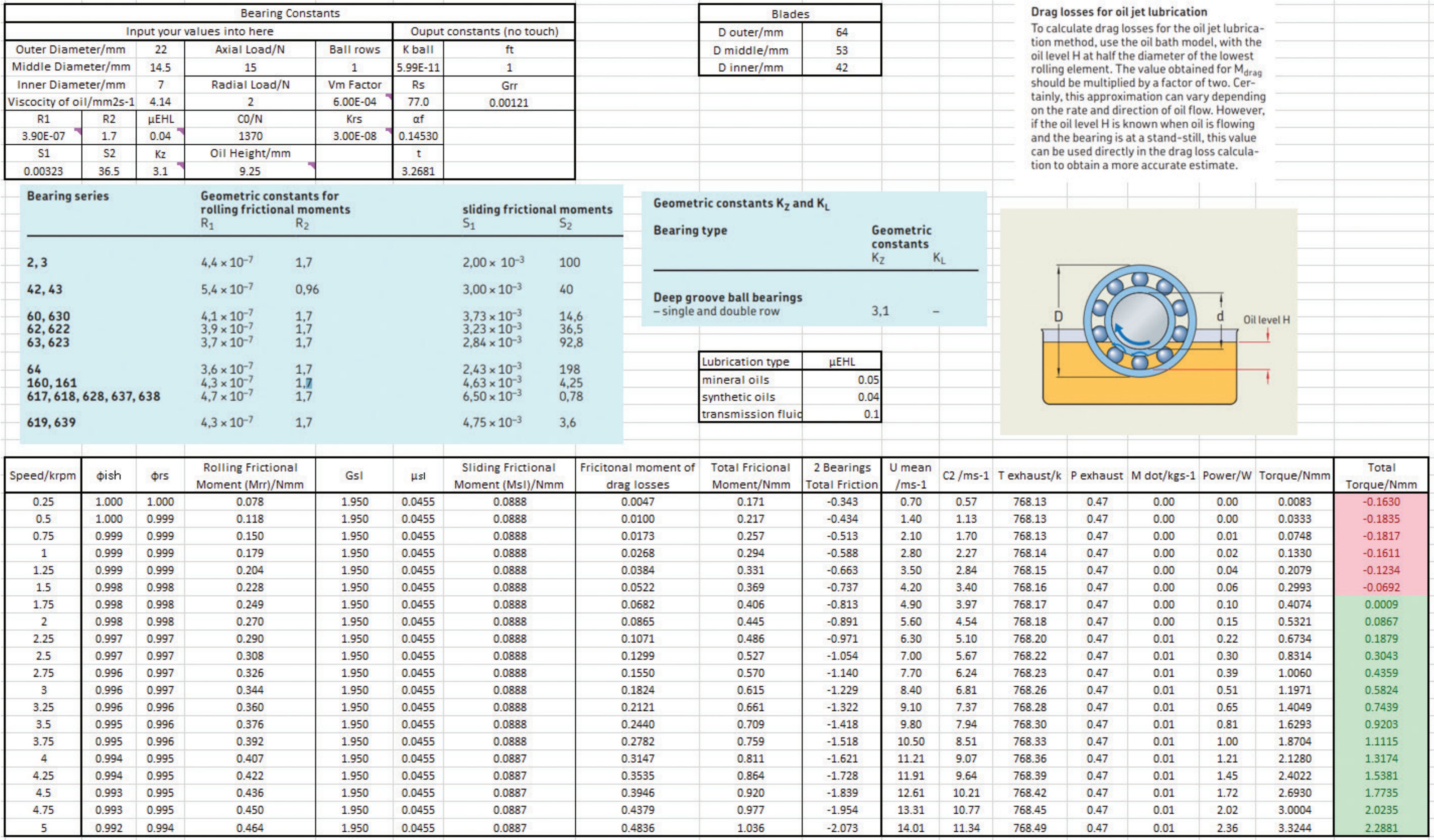


Figure 2: Taken from page 14 of source 4, it shows a bearing semi submerged in an oil bath.

However, because oil jet lubrication will be used, there has to be a correction factor. As source 4 advises, the oil depth can be calculated at half the diameter of the lowest rolling element. This then allows us to calculate the H/d_m ratio which then lets us find the V_m factor from the following graph.



Graph 1: Taken from page14 of source 4, showing the relationship between V_m and H/d_m .



Bearing Constants						
Input your values into here					Ouput constants (no touch)	
Outer Diameter/mm		22	Axial Load/N	Ball rows	K ball	ft
Middle Diameter/mm		14.5	15	1	5.99E-11	1
Inner Diameter/mm		7	Radial Load/N	Vm Factor	Rs	Grr
Viscocity of oil/mm2s-1		4.14	2	6.00E-04	77.0	0.00121
R1	R2	µEHL	C0/N	Krs	αf	
3.90E-07	1.7	0.04	1370	3.00E-08	0.14530	
S1	S2	Kz	Oil Height/mm		t	
0.00323	36.5	3.1	9.25		3.2681	

Blades	
D outer/mm	64
D middle/mm	53
D inner/mm	42

Drag losses for oil jet lubrication
To calculate drag losses for the oil jet lubrication method, use the oil bath model, with the oil level H at half the diameter of the lowest rolling element. The value obtained for M_{drag} should be multiplied by a factor of two. Certainly, this approximation can vary depending on the rate and direction of oil flow. However, if the oil level H is known when oil is flowing and the bearing is at a stand-still, this value can be used directly in the drag loss calculation to obtain a more accurate estimate.

Bearing series	Geometric constants for rolling frictional moments R_1	R_2	sliding frictional moments S_1	S_2
2, 3	$4,4 \times 10^{-7}$	1,7	$2,00 \times 10^{-3}$	100
42, 43	$5,4 \times 10^{-7}$	0,96	$3,00 \times 10^{-3}$	40
60, 630	$4,1 \times 10^{-7}$	1,7	$3,73 \times 10^{-3}$	14,6
62, 622	$3,9 \times 10^{-7}$	1,7	$3,23 \times 10^{-3}$	36,5
63, 623	$3,7 \times 10^{-7}$	1,7	$2,84 \times 10^{-3}$	92,8
64	$3,6 \times 10^{-7}$	1,7	$2,43 \times 10^{-3}$	198
160, 161	$4,3 \times 10^{-7}$	1,7	$4,63 \times 10^{-3}$	4,25
617, 618, 628, 637, 638	$4,7 \times 10^{-7}$	1,7	$6,50 \times 10^{-3}$	0,78
619, 639	$4,3 \times 10^{-7}$	1,7	$4,75 \times 10^{-3}$	3,6

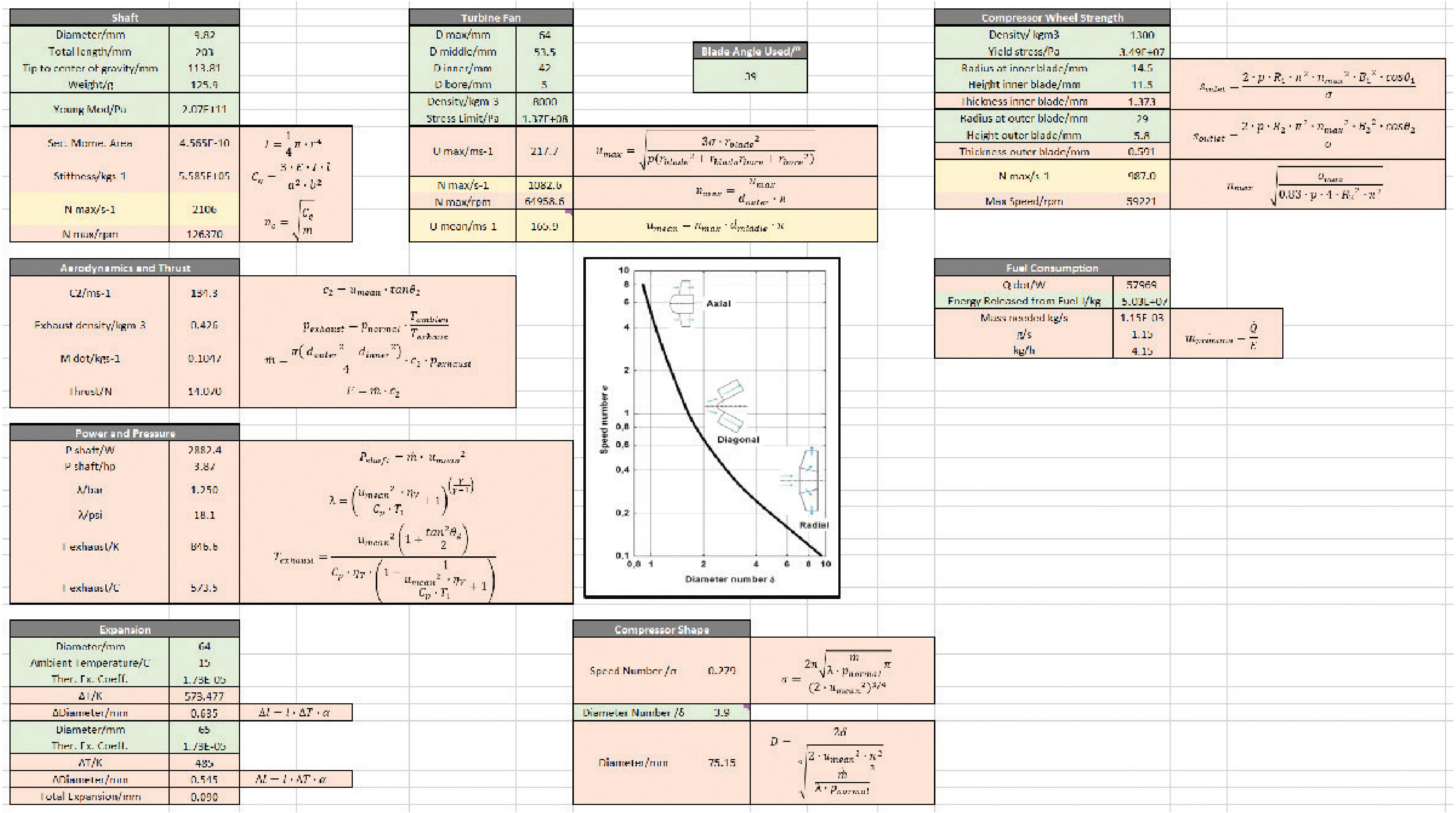
Geometric constants K_Z and K_L

Bearing type	Geometric constants K_Z	K_L
Deep groove ball bearings – single and double row	3,1	–

Lubrication type	µEHL
mineral oils	0.05
synthetic oils	0.04
transmission fluic	0.1

Speed/krpm	φish	φrs	Rolling Frictional Moment (Mrr)/Nmm	Gsl	µsl	Sliding Frictional Moment (Msl)/Nmm	Fricitonal moment of drag losses	Total Fricional Moment/Nmm	2 Bearings Total Friction	U mean /ms-1	C2 /ms-1	T exhaust/k	P exhaust	M dot/kgs-1	Power/W	Torque/Nmm	Total Torque/Nmm
0.25	1.000	1.000	0.078	1.950	0.0455	0.0888	0.0047	0.171	-0.343	0.70	0.57	768.13	0.47	0.00	0.00	0.0083	-0.1630
0.5	1.000	0.999	0.118	1.950	0.0455	0.0888	0.0100	0.217	-0.434	1.40	1.13	768.13	0.47	0.00	0.00	0.0333	-0.1835
0.75	0.999	0.999	0.150	1.950	0.0455	0.0888	0.0173	0.257	-0.513	2.10	1.70	768.13	0.47	0.00	0.01	0.0748	-0.1817
1	0.999	0.999	0.179	1.950	0.0455	0.0888	0.0268	0.294	-0.588	2.80	2.27	768.14	0.47	0.00	0.02	0.1330	-0.1611
1.25	0.999	0.999	0.204	1.950	0.0455	0.0888	0.0384	0.331	-0.663	3.50	2.84	768.15	0.47	0.00	0.04	0.2079	-0.1234
1.5	0.998	0.998	0.228	1.950	0.0455	0.0888	0.0522	0.369	-0.737	4.20	3.40	768.16	0.47	0.00	0.06	0.2993	-0.0692
1.75	0.998	0.998	0.249	1.950	0.0455	0.0888	0.0682	0.406	-0.813	4.90	3.97	768.17	0.47	0.00	0.10	0.4074	0.0009
2	0.998	0.998	0.270	1.950	0.0455	0.0888	0.0865	0.445	-0.891	5.60	4.54	768.18	0.47	0.00	0.15	0.5321	0.0867
2.25	0.997	0.997	0.290	1.950	0.0455	0.0888	0.1071	0.486	-0.971	6.30	5.10	768.20	0.47	0.01	0.22	0.6734	0.1879
2.5	0.997	0.997	0.308	1.950	0.0455	0.0888	0.1299	0.527	-1.054	7.00	5.67	768.22	0.47	0.01	0.30	0.8314	0.3043
2.75	0.996	0.997	0.326	1.950	0.0455	0.0888	0.1550	0.570	-1.140	7.70	6.24	768.23	0.47	0.01	0.39	1.0060	0.4359
3	0.996	0.997	0.344	1.950	0.0455	0.0888	0.1824	0.615	-1.229	8.40	6.81	768.26	0.47	0.01	0.51	1.1971	0.5824
3.25	0.996	0.996	0.360	1.950	0.0455	0.0888	0.2121	0.661	-1.322	9.10	7.37	768.28	0.47	0.01	0.65	1.4049	0.7439
3.5	0.995	0.996	0.376	1.950	0.0455	0.0888	0.2440	0.709	-1.418	9.80	7.94	768.30	0.47	0.01	0.81	1.6293	0.9203
3.75	0.995	0.996	0.392	1.950	0.0455	0.0888	0.2782	0.759	-1.518	10.50	8.51	768.33	0.47	0.01	1.00	1.8704	1.1115
4	0.994	0.995	0.407	1.950	0.0455	0.0887	0.3147	0.811	-1.621	11.21	9.07	768.36	0.47	0.01	1.21	2.1280	1.3174
4.25	0.994	0.995	0.422	1.950	0.0455	0.0887	0.3535	0.864	-1.728	11.91	9.64	768.39	0.47	0.01	1.45	2.4022	1.5381
4.5	0.993	0.995	0.436	1.950	0.0455	0.0887	0.3946	0.920	-1.839	12.61	10.21	768.42	0.47	0.01	1.72	2.6930	1.7735
4.75	0.993	0.995	0.450	1.950	0.0455	0.0887	0.4379	0.977	-1.954	13.31	10.77	768.45	0.47	0.01	2.02	3.0004	2.0235
5	0.992	0.994	0.464	1.950	0.0455	0.0887	0.4836	1.036	-2.073	14.01	11.34	768.49	0.47	0.01	2.36	3.3244	2.2881

Figure 3: A calculator that I made which calculates the frictional moments generated at certain speeds. This can then be used to figure out the minimum self-sustain speed needed which is predicted at a point just above 1500rpm. This is the very lowest theoretical limit as it does not account for various additional frictional losses such as described later on and does not take into account aerodynamic frictions



Shaft		
Diameter/mm	9.82	
Total length/mm	203	
Tip to center of gravity/mm	113.81	
Weight/g	125.9	
Young Mod/Pa	2.07E+11	
Sec. Mome. Area	4.565E-10	$I = \frac{1}{4}\pi \cdot r^4$
Stiffness/kgs-1	5.585E+05	$C_e = \frac{3 \cdot E \cdot I \cdot l}{a^2 \cdot b^2}$
N max/s-1	2106	$n_c = \sqrt{\frac{C_e}{m}}$
N max/rpm	126370	

Turbine Fan		
D max/mm	64	
D middle/mm	53.5	
D inner/mm	42	
D bore/mm	5	
Density/kgm-3	8000	
Stress Limit/Pa	1.37E+08	
U max/ms-1	217.7	$u_{max} = \sqrt{\frac{3\sigma \cdot r_{blade}^2}{p(r_{blade}^2 + r_{blade}r_{bore} + r_{bore}^2)}}$
N max/s-1	1082.6	$n_{max} = \frac{u_{max}}{d_{outer} \cdot \pi}$
N max/rpm	64958.6	
U mean/ms-1	165.9	$u_{mean} = n_{max} \cdot d_{middle} \cdot \pi$

Blade Angle Used/°
39

Compressor Wheel Strength		
Density/ kgm3	1300	
Yield stress/Pa	3.49E+07	
Radius at inner blade/mm	14.5	$s_{inlet} = \frac{2 \cdot p \cdot R_1 \cdot \pi^2 \cdot n_{max}^2 \cdot B_1^2 \cdot \cos\theta_1}{\sigma}$
Height inner blade/mm	11.5	
Thickness inner blade/mm	1.373	
Radius at outer blade/mm	29	$s_{outlet} = \frac{2 \cdot p \cdot R_2 \cdot \pi^2 \cdot n_{max}^2 \cdot B_2^2 \cdot \cos\theta_2}{\sigma}$
Height outer blade/mm	5.8	
Thickness outer blade/mm	0.591	
N max/s-1	987.0	$n_{max} = \sqrt{\frac{\sigma_{max}}{0.83 \cdot p \cdot 4 \cdot R_2^2 \cdot \pi^2}}$
Max Speed/rpm	59221	

Aerodynamics and Thrust		
C2/ms-1	131.3	$c_2 = u_{mean} \cdot \tan\theta_2$
Exhaust density/kgm-3	0.426	$p_{exhaust} = p_{normal} \cdot \frac{T_{ambient}}{T_{exhaust}}$
M dot/kgs-1	0.1047	$\dot{m} = \frac{\pi(d_{outer}^2 - d_{inner}^2)}{4} \cdot c_2 \cdot p_{exhaust}$
Thrust/N	14.070	$F = \dot{m} \cdot c_2$

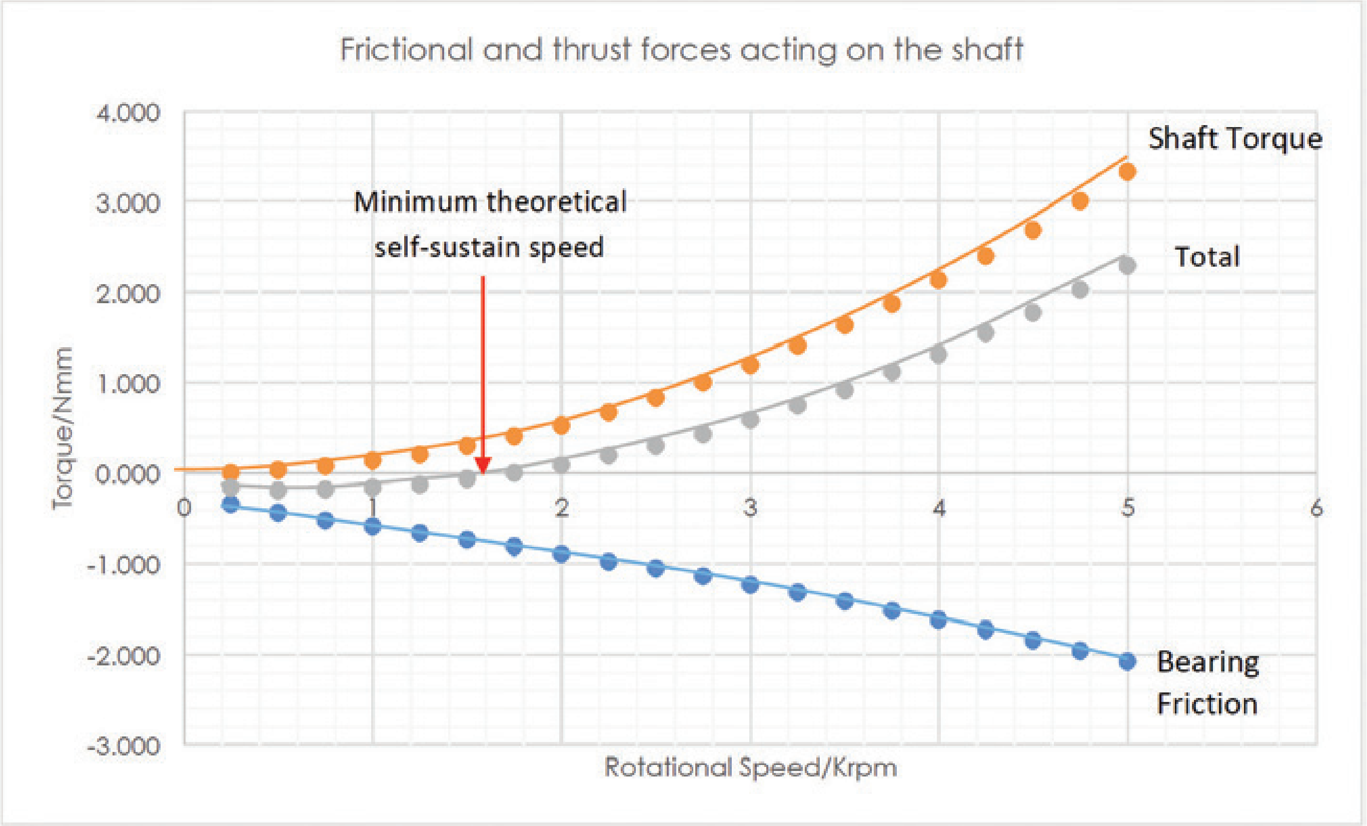
Fuel Consumption		
Q dot/W	57969	
Energy Released from Fuel J/kg	5.03E+07	
Mass needed kg/s	1.15E-03	$\dot{m}_{propane} = \frac{\dot{Q}}{E}$
g/s	1.15	
kg/h	4.15	

Power and Pressure		
P shaft/W	2882.4	$P_{shaft} = \dot{m} \cdot u_{mean}^2$
P shaft/hp	3.87	
λ/bar	1.250	$\lambda = \left(\frac{u_{mean}^2 \cdot \eta_V}{C_p \cdot T_1} + 1\right)^{\left(\frac{\gamma}{\gamma - 1}\right)}$
λ/psi	18.1	
T exhaust/K	846.6	$T_{exhaust} = \frac{u_{mean}^2\left(1 + \frac{\tan^2\theta_2}{2}\right)}{C_p \cdot \eta_T \cdot \left(1 - \frac{1}{\frac{u_{mean}^2 \cdot \eta_V}{C_p \cdot T_1} + 1}\right)}$
T exhaust/C	573.5	

Expansion		
Diameter/mm	64	
Ambient Temperature/C	15	
Ther. Ex. Coeff.	1.73E-05	
ΔT/K	573.477	
ΔDiameter/mm	0.635	$\Delta l = l \cdot \Delta T \cdot \alpha$
Diameter/mm	65	
Ther. Ex. Coeff.	1.73E-05	
ΔT/K	485	
ΔDiameter/mm	0.545	$\Delta l = l \cdot \Delta T \cdot \alpha$
Total Expansion/mm	0.090	

Compressor Shape		
Speed Number /σ	0.279	$\sigma = \frac{2\pi\sqrt{\frac{\dot{m}}{\lambda \cdot p_{normal}}\pi}}{(2 \cdot u_{mean}^2)^{3/4}}$
Diameter Number /δ	3.9	
Diameter/mm	75.15	$D = \frac{2\delta}{\sqrt[4]{\frac{2 \cdot u_{mean}^2 \cdot \pi^2}{\frac{\dot{m}}{\lambda \cdot p_{normal}}}}}$

Figure 4: The calculator I built with all the relevant formulae. This automatically calculates for me all the parameters if I need to make a small adjustment or for designing a new engine with different parameters.



Graph 2: Plotting the various moments on the shaft at certain speeds. Blue shows the friction from bearings. Orange shows the torque from the turbine section. Grey shows the total force on the shaft. The point at which the total forces becomes positive indicates the minimum theoretical self-sustain speed at around 1600rpm.

Additional sources of friction include effects of clearance, misalignment and material of the bearings. Hybrid bearings made of silicon nitride decreases the contact area between the ball and the raceways significantly reducing rolling and sliding friction. Additionally, the comparatively lower density of ceramic rolling elements against steel reduces centrifugal forces, also reducing friction at high speeds.

Hybrid bearings also come with the added benefit of being self-lubricating, requiring no oil as well as being optimised for high speeds and high temperatures. This is the main reason why they are extensively in heavy industry despite costing around ten times as much as regular bearings while offering marginal gains compared to a properly lubricated bearing.

SUMMARY

Despite not being able to get it to self-sustain, I believe now the main issue is the clearance between the turbine and the outer-casing. Hopefully I will be able to improve my machining ability and achieve self-sustain sometime in the future.

Governing Rojava: the democratic confederalist model in practice

ILA Arts & Humanities Winner 2019

Alfie Cherry

PREFACE

This paper seeks to examine the ideology of democratic confederalism, as theorised by Abdullah Öcalan, both in theory and in practice. The case study of democratic confederalism in practice is Rojava, an autonomous region of Syria. Supported by an explanation of the history and geography of the region, the paper examines each tenet of democratic confederalism: structure of governance; economics; feminism; ecology; multiculturalism and secularism. Each of these ideological ideals is then compared to the reality of Rojava, both as a whole and within each autonomous sub-region. Finally, the paper concludes as to whether democratic confederalism has been successfully implemented in Rojava; whether the problems Rojava faces currently are as a result of democratic confederalism; how Rojava can improve in practice and what the future holds for the stability of Rojava.

CHAPTER 1: INTRODUCTION TO ROJAVA

'Rojava is a beacon of light for the rest of the world' [1]
Lloyd Russell-Moyle, British MP

1.1: HISTORY OF ROJAVA

Amidst the bloodshed of the Syrian Civil War, a society unlike any other in the Middle East has emerged. In the predominantly ethnic Kurdish regions of North and East Syria, a multicultural, feminist, anti-capitalist, and democratic de-centralised federation has arisen; its borders inherently fluid. Its name is Rojava — translated from the Kurdish Kurmanji as 'where the sun sets'.

Formally titled the 'Autonomous Administration of North and East Syria', Rojava's first experience of autonomy was in 2012. After Syrian government forces withdrew from three predominantly Kurdish areas in the North and East of the country (Afrin, Jazira, and Euphrates) during the spring of 2012, the security and administration of the areas were left almost entirely to the local Kurdish militias. [2] Just months after the initial retreat of Syrian government forces, the two leading Kurdish political parties in Syria — the Democratic Union Party (PYD) and the Kurdish National Council (KNC) formed to create the Kurdish Supreme Committee (KSC) on July 12th, 2012. [3] The KSC governed Rojava until November 2013, during which time the committee created the People's Protection Units (YPG) and Women's Protection Units (YPJ). [4] The YPG entered the Syrian Civil War on the 19th July, 2012 — a mere week after its foundation — with the capture of the Northern city of Kobanî. [5] They were met with no resistance by the remaining few Syrian government forces, who entirely retreated from the city to focus on the ongoing Civil War in other regions of Syria. [6] The KSC disbanded in November 2013, after the KNC accused the PYD of monopolising power within the coalition. [7] Since the breaking down of the KSC, Rojava has been governed by the 'Movement for a Democratic Society' coalition (TEV-DEM), the primary member of which is the PYD. During this period, the most substantial political development has been the formal declaration of the autonomy of Rojava on 29th January, 2014 by TEV-DEM. [8] This declaration was denounced by both the Syrian government and the Syrian rebel coalition. [9]

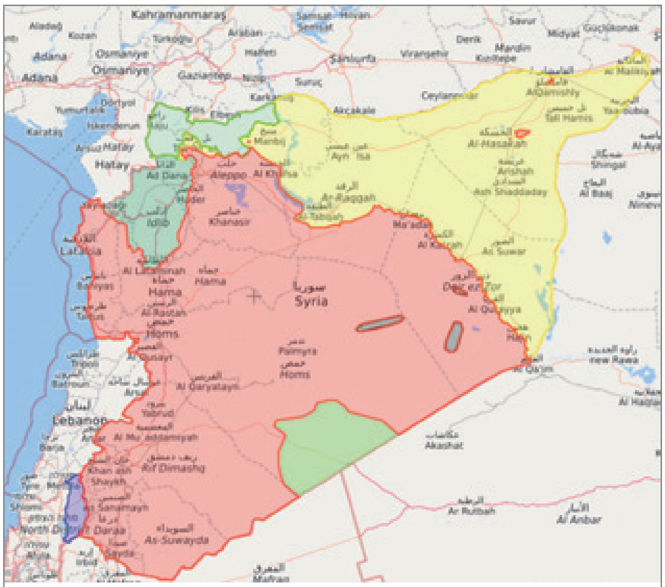
Major ground offensives from Islamic State (IS), alongside the entrance of Turkey into the Syrian Civil War in 2016 were and continue to be major threats to the stability of Rojava. Small-arms clashes between the YPG and the Syrian government continue, particularly in the Eastern region of Deir az-Zor, but this has become increasingly rare as the two sides have allied against Turkey. [10] On 10th October 2015, the YPG and YPJ joined a coalition alongside local militias to form the Syrian Democratic Forces (SDF). [11] The SDF proved to be the most effective ground force against ISIS, pushing ISIS back from more territory and entrenched strongholds such as Raqqa and Baghouz than the Syrian government. [12]

As of March 2019, Rojava makes up 28.9% of all territory in Syria. [13] Back in July 2012, Rojava was less than 5%. [14] The history of Rojava is one of struggle in the face of immense adversity, triumphing against the odds to establish a society described as an 'anarchist utopia' by Western journalists and politicians.

1.2: GEOGRAPHY OF ROJAVA

Rojava is situated in the North and East of Syria, comprising of the entirety of the Al- Hasakah governate, almost half of the Deir ez-Zor governate, the majority of the Raqqa governate, and two separate sections of the Aleppo governate. It borders both Turkey and Iraqi Kurdistan, and currently consists of approximately 32,600 square miles. [15] However, to discuss the geography of Rojava at any single fixed point is to take a limited understanding of the region. Rojava's borders are fluid due to the ongoing civil war in Syria. Villages, towns, and a small number of cities are still contested, with Rojava and the Syrian government officially sharing ownership of some cities (including Rojava's own capital city — Qamishli). Small-scale skirmishes between the YPG and Syrian government forces continue, particularly in the oil-rich region of Deir ez-Zor, where control of oil fields and refineries means a significant boost of regional income. [16]

The North-West of Rojava, in the Aleppo governate, is the site of the



Key:
Red — *Syrian government territory*
Yellow — *Rojavan territory*
Light green — *Free Syrian Army territory (including Turkish-backed)*
Dark green — *Hayat Tahrir al-Sham territory*
Purple — *Israeli-occupied Golan Heights*
Grey — *Islamic State territory*

The borders of Rojava as of June, 2019.

fiercest fighting the YPG is currently involved in. [17] Turkey and its Syrian rebel proxy, the Turkish-backed Free Syrian Army, are currently engaged in Operation Olive Branch. This military operation, which began on 20th January 2018, has caused the majority of the Afrin district of the Aleppo governate to be taken over by Turkey. [18] Rojava's territory in the area consists of roughly 92 square miles, surrounded by Turkish and TFSA territory, with Syrian government territory to the South. [19]

1.2.1: STRUCTURE OF THE REGIONAL SYSTEM

Rojava is divided into self-governing regions. Consequently, not only does it act autonomously from the Syrian government, but each region acts autonomously from the rest of Rojava. The official regions of Syria, as designated by the Syrian government, are referred to throughout as 'governates'. The official regions of Rojava will henceforth be referred to simply as 'regions'. This decentralised system begin in January of 2014, when the regions of Afrin (in the Afrin Governate), Jazira (in the al-Hasakah Governate, and Euphrates (in the Afrin, al-Hasakah, and Aleppo Governates) were established. [20] There are now seven regions of Rojava in total, with Raqqa (in the Raqqa governate), Tabqa (also in the Raqqa governate), Manbij (in the Aleppo governate), and Deir ez-Zor (in the Deir ez-Zor governate) all joining the existing three in September, 2018. [21] Each region has a legislative assembly, a president, and a set of ministers. [22] They are able to pass laws, decide upon funding projects and enforce security within their region through the YPG and YPJ. [23] Each region has their own set of regiments, which the government and military leaders can direct.

1.2.2: RELATION TO GREATER KURDISTAN

As established, Rojava is a majority ethnic Kurdish region of Syria. Kurds are the second largest ethnic group in Syria after Syrian Arabs, making up approximately 10% of the population. [24] In Iraq, Kurds are also the second largest ethnic group, making up approximately 15%. [25] Highly concentrated in the North-West of the country, Iraqi Kurds gained autonomy in 1992, when the Iraqi government agreed to allow the official formation of Iraqi Kurdistan [26]. The concept of Greater Kurdistan is based on irredentist claims, and stretches into the North-East reaches of Turkey, West and North-West Iran, and even Southern sections of Armenia. Within Greater Kurdistan, Rojava is considered to be part of Northern Kurdistan.

However, whereas Iraqi Kurdistan has its Kurdish identity as the very basis for its existence, Rojava does not. The struggle for autonomy amongst the inhabitants of Rojava has not been based solely upon Kurdish nationalism, but instead, support for the federalisation of Syria as a whole. Whilst there are elements of Kurdish nationalism within the Rojava movement (such as links to Kurdish nationalist parties in Turkey), it is a multiethnic and multicultural society. In official Rojavan governmental statements and papers, little mention is given to the concept of Kurdistan, and the name was even changed from the 'Democratic Federation of Rojava' to the 'Autonomous Administration of North and East Syria', specifically to convey an inherently multicultural society; so as not to alienate the non-Kurdish citizens of Rojava. [27]

1.3: MODERN HISTORY OF KURDS IN SYRIA

Whilst the history of the Kurds in Syria dates to before the 11th century, it is only necessary to examine the presence of Kurds in Syria from the advent of independent Syria (in 1946) in order to understand their current plight. The United Nations has described the treatment towards Kurds by successive Syrian government as 'ethnic discrimination and national persecution', through the imposition of 'ethnically-based… regulations and exclusionary measures.' [28] For example, under the Ba'athist government in August 1962, approximately 120,000 Syrian Kurds were stripped of their Syrian citizenship, leaving them stateless. [29] This equated to 20% of the Kurdish population of Syria at the time, and barred them gaining employment, being educated in schools and universities, and engaging in the political process. [30] The Ba'ath party's policy of Arabisation continued in the 1970s, when, in 1973, 750 square kilometers of agricultural land was taken over by Syrian government security forces. [31] This farmland, situated in the al-Hasakah Governate, belonged to tens of thousands of Kurds, and the Syrian government gave it to ethnic Arab families, brought in from other governates. [32] In more recent years, 6,000 square kilometers of rural land was seized from tens of thousands of Kurds, forcibly evicting them, and handing over the land to ethnic Arab families. [33]

This denial of Kurdish identity, alongside a broader destruction of multicultural society by the Syrian government, directly inspired Syrian Kurds to create a society, with multi-ethnic and multi-cultural communities. Furthermore, the Kurds saw the danger in having such a large, centralised state, which lent itself to oppressive strategies to further its national hegemony, through the denial of equal rights and democracy to many of its citizens. This historical oppression inspired much of Rojava's policies.

CHAPTER 2: POLITICS OF ROJAVA

'A society can never be free without women's liberation' [34]
Abdullah Öcalan, Leader of Kurdistan Workers' Party

2.1: DEMOCRATIC CONFEDERALISM

Rojava is guided by the ideology of democratic confederalism at the level of both institutionalised politics and social interaction. Democratic confederalism is based upon a set of guiding principles, spanning government structure, economy, social policy, and ecology. Primarily, it focuses upon power relations within society, and seeks the removal of hierarchy. It achieves the (theoretical) removal of hierarchy through a system of 'popularly elected administrative councils, allowing local communities to exercise autonomous control over their assets, while linking to other communities via a network of confederal councils'. [35] Therefore, it values direct democracy above all else.

It posits itself as the blueprint for a democratic nation, in opposition to the nation-state — the model of national governance that has dominated the world since the Enlightenment. [36] However, the ideology goes far deeper than a system of governmental organisation. Democratic confederalism creates the framework for a post-capitalist economy, with ecological protection, feminism, multi-culturalism, and secularism all championed as part of the overarching opposition to hierarchy. In a more broad perspective, in can be classified as a form of libertarian socialism. This chapter seeks to explain the ideology of democratic confederalism in abstraction to the practice of it in Rojava, focusing purely on the theoretical aspects. An analysis of the ideology in practice in Rojava will be the subject of Chapter 3.

2.1.1: ABDULLAH ÖCALAN

The original philosopher of democratic confederalism is Abdullah Öcalan, a Kurdish-Turkish political theorist, militant, and political leader. Öcalan is widely viewed as the ideological figurehead of Rojava, with his writings the basis for its constitution. Öcalan was originally a revolutionary Marxist-Leninist, who founded the Kurdistan Workers' Party (PKK) in November, 1978. [37] The PKK fought for a combination of revolutionary Marxist- Leninist socialism alongside Kurdish nationalism, aiming to form a Communist nation of Kurdistan, as they asserted that Kurds had been oppressed by the hands of the Turkish state for hundreds of years. However, after he was imprisoned by Turkey in 1999, Öcalan moved away from the revolutionary Marxist-Leninism platform, and formulated the ideology of democratic confederalism. He was inspired by the American eco-anarchist philosopher Murray Bookchin, who is considered to be his main influence. [38]

In March 2005, just seven years before Rojava's first experience of autonomy, Öcalan released the 'Declaration of Democratic Confederalism in Kurdistan', which advocated for a peaceful solution to the PKK-Turkey insurgency, alongside the establishment of a borderless confederation of Kurdish regions in Turkey, Syria, Iraq, and Iran. [39] Throughout his publications and speeches, Öcalan has affirmed the principles of democratic confederalism, and shown himself to be a figurehead for Rojava. His face regularly features on flags flown by governmental, military, and civil figures alike across Rojava. [40]

2.1.2: STRUCTURE OF GOVERNANCE

The structure of governance under Democratic Confederalism is a form of de-centralised direct democracy. It can be described succinctly as a 'system of popularly elected administrative councils…linking to other communities through a network on confederal councils.' [41] This form of governance allows local communities to have greater autonomy as to how their day-to-day lives are run, without fear of a centralised, hierarchical state imposing its own agenda onto them. Accordingly, all citizens are able to take part in the administrative councils (which administrate villages, towns, and cities), which can then feed into wider governance in specific regions or areas (through confederal councils; a confederation of the administrative councils). Citizens elect ministers to implement their passed legislative proposals raised in the administrative councils, and these same ministers then represent the views of the citizens in the confederal councils.

Whilst the use of ministers may present a challenge to the ideal of direct democracy, Öcalan responds to this by stating that these ministers 'only serve the co-ordination and implementation of the will of the communities that send their delegates to the general assemblies. For a limited space of time, they (the ministers) are both the mouthpiece and executive institutions'. [42] Therefore, the ministers are merely a matter of short-term practicality, as entire communities are not able to attend confederal councils, which could be held hundreds of miles away from their village or town. Upon returning to their community, the ministers hold no power over the will of the community. By implementing direct democracy at nearly the most localised level possible, Öcalan asserts that individuals are able to gain self-determination, and allows 'the power of decision to rest with local grassroots institutions.' [43] This decentralisation is key in the removal of overarching hierarchy, which both Öcalan and Bookchin believe to be the root cause of oppression of minority groups throughout human history (be that class, gender, or race).

2.1.3: ECONOMICS

The economic theory of Democratic Confederalism lies broadly within the libertarian socialist paradigm. The ideology is guided by opposition to social hierarchy, and opposes capitalism on the basis that it 'grounds itself on maximum profit' and 'a means to control and enslave society.' [44] Instead, Democratic Confederalism supports the existence of a 'communal economy, grounded on a theory-of-value based on societal need.' [45] This communal economy would be based around a worker-owned means of production, against both state capitalism and state socialism. Furthermore, the profit motive would be removed; therefore whilst similar to market socialism in the allowance of co-operative and autonomous businesses, it differs in that the profit motive would be removed — not through government legislation, but by societal consensus (presumably brought about by teaching future generations) towards the profit motive as being inherently exploitative and conducive to societal hierarchy. According to Öcalan, capitalism, driven by its endpoint desire of capital accumulation, 'domesticates society…and alienates the community from its natural foundations.' [46] This means that the form of nation-state capitalism that has dominated the world for the majority of the 19th, 20th, and 21st centuries is, in the view of Öcalan, unnatural for small-scale, organic communities, and thus brings about unjust hierarchy through the centralisation of capital.

Öcalan's further development of the economic position of Democratic Confederalism rests largely with the practices of certain industries. For example, the system of financial markets would be allowed to remain under a democratic confederalist society, so long as it 'serves economic productivity'. [47] Therefore, he does not dogmatically reject certain economic structures borne of capitalism, so long as they move away from the motive of accumulating capital. However, the practice 'making money from money' (presumably referring to stock market trading, interest on money lending, and investment banking) is considered 'the most effortless form of exploitation' by Öcalan, and as such, cannot remain under Democratic Confederalism. [48]

2.1.4: FEMINISM

One of the central tenets of Democratic Confederalism is the championing of equal rights between the sexes; a revolutionary concept for much of the Middle East. Öcalan describes the prevalence of sexism in modern nation-states as having 'consolidated the traditional framework of hierarchies' — the hierarchy of the sexes builds the framework for hierarchies in class, religion, sexuality, and even species. [49] Therefore, in Öcalan's view, for any society to remove hierarchy as a whole and implement a society based on equality and freedom, feminism is essential. Theoretically, this should equate to equal representation on both the administrative councils and general councils, as well as in other branches of government, business structures, public speaking opportunities, and the military.

The specific form of feminism advocated by Democratic Confederalism is 'jineology'; this translates from Kurmanji Kurdish as 'the science of women'. Jineology was pioneered by Öcalan, and is embodied in the doctrine that 'the level of women's freedom and equality determines the freedom and equality of all sections of society'. [50]

Therefore, jineology is the feminist belief that the emancipation of women is necessary in the emancipation of all peoples in society against an oppressive hierarchy. For Öcalan, the hierarchical basis for traditional patriarchal social systems (as is prevalent across many Middle Eastern countries) stems from an overarching justification of hierarchy. Once one form of hierarchy is taken down, the same process can be used to take down all hierarchy. Jineology follows the trend of the other tenets of Democratic Confederalism, in that it posits specific social change in order to bring about wider, more general social change — that is, the total removal of hierarchy.

2.1.5: ECOLOGY

The influence of Murray Bookchin's work on Öcalan is evident in his formulation of social ecology. Bookchin pioneered the concept of social ecology — the belief that humanity's assertion of superiority and domination over nature (be that animals killed for food; trees cut down for paper, and entire ecosystems destroyed through urbanisation) stems directly from social hierarchies of human-over-human (be that class, race, gender, or sexuality). Therefore, mass ecological devastation is a direct product of hierarchical society, and in order to cease environmental destruction, it is necessary to remove social hierarchy. Following from Bookchin, Öcalan asserts that 'we may gain a better understanding of the essence of collective life' from following social ecology, and protecting ecological systems in the face of mass pollution and climate change. [51] Consequently, Öcalan believes that by respecting nature, humanity is led towards respecting collective life.

In practice, this means that nature must be cared for, and not commodified as it has been for hundreds of years. According to Öcalan, capitalism, with its inherent propensity to create artificial needs for consumers, has caused overconsumption that has further drained nature of its resources and stability; it must therefore be replaced by a new system that does not lead to commodification and overconsumption. [52] The upmost goal of Democratic Confederalism's social ecology is to undo the anthropogenic effects on nature, and this can be achieved through re-wilding, mass planting of trees, moves toward veganism, de-urbanisation (which stems directly from localised geopolitical decentralisation), and the minimisation of pollution.

2.1.6: MULTICULTURALISM AND SECULARISM

Democratic Confederalism was formulated by Öcalan largely in response to the treatment of the Kurdish people in Turkey. Just as the Syrian government had followed a policy of Arabisation against the Kurds for much of the 20th and 21st centuries, so too had the Turkish government followed a policy of Turkification. Kurds were killed and displaced en masse by the Turkish state; by the mid-1990s, approximately 378,000 Kurdish civilians had been displaced due to the policies of the Turkish military during the ongoing Kurdish- Turkish conflict. [53] Until 1991, the Kurdish language was banned in both public and private settings — with anyone breaking this ban being imprisoned. [54]

Öcalan views this attempted destruction of Kurdish culture as a wider, inherent symptom of the centralised nation-state model, which seeks to create a single national culture, identity, and religious consensus. As he has seen the violence and misery this hegemonic claim brought onto the Kurds, he wishes to formulate a societal model that allows for the rich diversity of culture found in Middle Eastern society. Furthermore, he abhors theocracy, believing it to be an assertion of false hegemony, and purely a means to control any societal dissent. He exemplifies this in Iran, which he describes as 'multiethnic and multi-religious and blessed with a rich culture', which has been partly eroded by the 'hegemonic claim of theocracy'. [55] The decentralisation of Democratic Confederalism allows 'all cultural identities' to express themselves in the local meetings, and ensures that one cannot dominate another, regardless of the ethnic or religious makeup of the entire federation. [56] By decentralising government to a local level, the natural diversity of culture is able to be maintained; cultural, ethnic, and religious conflict will be kept to a minimum, as all peoples are allowed to be heard in their community.

2.2: CONSTITUTION OF ROJAVA

When Rojava first declared autonomy on January 29th, 2014, it adopted a provisional constitution — officially titled 'Charter of the Social Contract'. The constitution was later reformed in June, 2016 and the four cantons that had formed since January, 2014 affirmed their support for it. It is evident that the constitution has been deeply inspired by the principles of Democratic Confederalism, with support for a de-centralised federal system (in both Rojava and the entirety of Syria), alongside the enshrinement of women's and minority rights into law. The Constitution offers a bridge between the purely theoretical concepts of Democratic Confederalism, Öcalan's texts, and the purely practical, actualised structures of Democratic Confederalism found in Rojava.

In Section One, the general principles of the constitution are laid out. First, it promises that 'all languages are equal in all…social, governmental, educational and cultural matters', thus Rojava is an inherently multicultural society. Furthermore, it promises that 'all administrative bodies and councils (will be) formed from the results of elections' — this necessitates decentralised administration, in order to allow for direct elections. So long as the cantons do not contravene the articles of the constitution, they may 'freely elect their administrative and representative bodies and may pursue their rights'. Crucially, Rojava is established as a border-fluid federation — Article 7 states that 'all cities, towns and villages in Syria which accede to this charter may form cantons in autonomous regions'. Regarding the role of women in society, it is further stated that Rojava will 'ensure the freedom of women' and that 'women represent themselves equally with men in all areas of life'. From the onset, the stark differences between the ideology of Rojava and the ideology of its neighbouring countries are clear to see. The enshrinement of gender equality, direct democracy, and multiculturalism are in deep contrast to the patriarchal, centralised, and culturally hegemonic states of Turkey, Israel, Iraq, and Iran.

In Section Two, the universal rights of all citizens of Rojava are explicated. This includes the right to life (thus abolishing the death penalty); the right to a fair and free trial, the right for local communities to pursue self-determination (particularly culturally); the right to equal and universal participation in political life (both to vote and to run for office); the right of private property and the right to freedom of conscience and belief (both politically and religiously).

In Section Three, the administrative structure of Rojava is explained in great detail. The constitution establishes the creation of 13 different governmental structures at all levels of social organisation — from the smallest of communes to the federation-wide Democratic Assembly. Throughout the constitution, there is a considerable focus on the importance of communal self-determination, against the centralised nation-state model. However, in theory, the constitution replaces the problems stemming from a centralised model with the problems stemming from such a radically decentralised model. With over a dozen different layers of government (not including the individual bodies that make up these different layers, such as departments of health or defence), it would not be unfair to label it bureaucratic and overly complex. Whether or not this theoretical bureaucracy actually does create problems is the subject of the next section.

One of the most important provisions in Section Three is the safeguarding of minority representation at all levels of government. Firstly, Article 47 ensures at least 40% of any administrative body must be made up of either sex. Secondly, in order to ensure ethnic minority rights, Article 46 requires all administrative bodies with a presidential head to also have a co-presidential head from a non-Kurdish ethnic group. Whilst affirmative action is widely practiced, particularly in Western nations, it is arguable that this undermines the direct democracy principles of Democratic Confederalism. However, it is necessary to balance between principles when they seemingly conflict — if women and ethnic minorities are institutionally underrepresented, then is it more important to ensure that their voices are being heard proportionally (and thus safeguard minority rights), rather than practice pure direct democracy with no affirmative protections (and thus practice Öcalan's key principle)?

2.3: SUCCESS OF GOVERNANCE IN PRACTICE

In assessing the success of the governance of Rojava, it is necessary to consider two questions: first, is the ideology of Democratic Confederalism being upheld and practiced? Second, is the governance of Rojava successful and effective in purely practical, realpolitik terms?

An in-depth analysis of all levels of governance in each canton would not be realistically feasible in this report — not only because of the enormously time-consuming undertaking this would be, but also because detailed analysis of Rojava at the very localised levels is sparse, with much of the analysis being journalistic, anecdotal accounts. Nevertheless, an analysis on a wider scale of governance (at federation-wide, canton-wide, and districtwide levels) is possible.

2.3.1: STRUCTURE OF GOVERNMENT AND CONFEDERATION

The structure of government as laid out in Rojava's constitution has been actualised — all 13 levels of government, alongside the canton-wide and confederation-wide executive bodies (referred to as 'authorities'; comparable to the US' Executive Departments) have been created. To take a case study of the city of Qamishli (the capital of Rojava), it consists of six neighbourhoods — each one with 18 communes, and each commune with approximately 300 households. [57] The ability for the authorities of Rojava to develop such a de-centralised system in the midst of a brutal civil war and still effectively govern society is nothing short of remarkable. There is little criticism of the system not holding true to its ideals — the federation-wide council (referred to as the Democratic Autonomous Administration) has not been seen to overbear the governance of even the smallest communal administrative bodies. Furthermore, the effective provision of government services under the decentralised system has been noted even by critics. [58]

However, there is one major criticism of the practice of the confederated government. This comes primarily from a Chatham House report in 2016, which found that 'the Rojava leadership's relationship with opposition groups remains fraught' due to the PYD 'alienating its political opposition…and forcing them to either abide by its rules or leave'. [59] The report argues that the PYD, despite its supposedly democratic intentions, has systematically undermined the legitimacy of any rival parties — particularly the KNC. [60] This is fundamentally at odds with the ideology of Democratic Confederalism, and violates the Constitution of Rojava.

However, Chatham House's findings are at odds with other major studies of Rojava. For example, Michael Knapp, in his book 'Revolution in Rojava', states that 'no other distinction is made between the PYD and the other parties', and that the only inter-party conflict is the result of the KNC deliberately undermining the PYD. [61] The KNC has even resorted to supporting embargoes against Rojava, and supplying intelligence to the CIA and Syrian government, all whilst attempting to undermine the PYD's authority through accusations of being a puppet of the Assad government. [62] These two reports cannot both be wholly correct — instead, it is necessary to take a balanced view, drawing on the evidence of each. Following Chatham's report, it may be the case that the PYD had suppressed its political opponents, and the KNC in particular. However, following Knapp's report, it is likely that this suppression has only been due to repeated efforts of the KNC to undermine not just the PYD, but Rojava itself.

2.3.2: SOCIAL POLICIES

The social policies of Rojava encompass all government action related to gender equality, minority representation, and ecology. The most successful of these policies has been gender equality and the furthering of women's rights in the region. The constitutional quota of 40% of administrative roles being filled by women has been greatly surpassed. In the Afrin canton, approximately 65% of roles in civil society, political, and military institutions are taken by women. [63] Across Rojava, 55% of administrative roles at the district level (comprising all neighbourhoods of a section of a canton) are women. [64] The creation of women-only councils and courts has further allowed for gender equality to prevail in Rojava. This has largely come about through the creation of Kongreya Star (the Star Union) — an all-encompassing women's movement, which has, since 2005, educated women in Rojava on the importance of engaging with civil society, politics, and autonomy. [65]

Ilhan Ahmed, head of the Syrian Democratic Council (the political wing of the YPG), has stated that in the Afrin canton, 'men's influence… is very weak'. [66] Knapp claims that this is because the traditional, patriarchal clans 'play no special role' in Afrin's society. [67] However, the same cannot be said for other cantons, such as the Euphrates canton. The influence of patriarchal clan society is considerably stronger in Euphrates, and Kongreya Star has found it more difficult to encourage women to join their movement. [68] Nevertheless, progress has been made — in October 2015, a set of laws were implemented to further women's rights, including the banning of child marriage. [69] Furthermore, all councils in Euphrates abide by the 40% gender quota in the Rojava constitution, and a women's council has been created. [70] The practice of decentralisation in Rojava has made uniform progression much more difficult, but no canton is without women's rights, women's councils, and the 40% gender quota.

Ethnic minority representation and relations in Rojava have also been mostly successful. Kurds, Syriacs, and Arabs (of all religious identities) have developed strong levels of trust, and the localised administrations have strived to overcome ethnic prejudices and conflicts that have plagued the Middle East for centuries. [71] The representation of ethnic minorities in local administration has, however, been fraught with accusations of ethnocentrism towards the Kurds by the governing PYD. These accusations reached a boiling point in May 2019, when protests broke out amongst the Arab populations of the Deir ez-Zor canton. [72] Arab residents accused the PYD of anti-Arab discrimination, particularly in regards to participation in the administrative bodies and communal assemblies. [73]

Furthermore, there was anger against the alleged conscription of Arab youths in the SDF — a practice used towards Kurdish youth, but allegedly to a lesser degree. [74] These accusations have been rallied against the PYD and SDF for a number of years, with one of the earliest institutional accusations being made in June 2014, when the nongovernmental organisation Human Rights Watch (HRW) raised concerns over the PYD's inclusion of Arab citizens in its administrative bodies. [75] However, these claims of anti-Arab racism have been met with skepticism. Many of the Arab populations in Rojava were initially part of the anti-Assad rebellion — under the moniker the 'Free Syrian Army'. [76] Beginning in April 2015, the YPG (and later the SDF) began working alongside Syrian government forces against both ISIS and (to a lesser extent) Free Syrian Army rebel groups. [77] Anti-Assad Arab populations in Rojava saw this as a betrayal, as the PYD had used anti-Assad propaganda to entice Arab support since 2013. [78] Therefore, it has been speculated that the charges of anti-Arab racism at the hands of the PYD are mostly propaganda tactics, seeking to undermine PYD legitimacy, and rally the Arab population to an anti-PYD and anti-Assad position. Nevertheless, there is evidence from the HRW reports that anti-Arab racism is existent within sections of the PYD; due to the decentralised structure of Rojava's governance, however, it cannot be said that this accusation rings true of the entirety of the PYD nor the SDF.

Ecology has been one of the less prevalent, but nevertheless existent practices of the Rojava administration. Decades of environmental mistreatment at the hands of the Syrian government left much of Rojava in a state of ecological devastation; oil extraction (particularly in the Deir ez-Zor region) and agricultural monoculture (centred around wheat farming), alongside the mass dumping of industrial waste in rivers all contributed to this destruction. [79] Currently, the Turkish government has been constructing dams along the Euphrates river — causing vast sections to dry up, and further harming Rojava's environment. [80] This, alongside the ongoing civil war, have caused environmental protection in Rojava to be severely hindered. Nevertheless, multiple ecological initiatives in all cantons of Rojava exist, and are working to solve the crisis. The most successful and wide-reaching of these initiatives is the 'Make Rojava Green Again' campaign, launched by the Internationalist Commune of Rojava (ICR). The campaign seeks to 'address… cultivation of food,…reforesting…alternative form of electricity, limiting fossil fuel usage, preserving the water supply, and…developing waste management solutions'. [81] In the Jazira canton, the ICR worked alongside the canton's Ecology Committee to develop reforesting in the Hayaka forest. This forest was declared a protected natural reserve by the Jazira administration in 2014, with construction, fishing, and hunting banned. [82] Furthermore, the ICR has committed to planting over 50,000 trees in the Hayaka region, having already planted 2,000 in a tree nursery in their commune in the summer of 2018. [83] The ICR has also planted 50,000 fruit tree shoots in the spring of 2018 alone, seeking to increase biodiversity in Rojava. [84] Further plans are underway for alternative energy solutions; increase in public transport to reduce fossil fuel emissions; improved recycling methods and moves away from non-biodegradable materials, such as plastic. [85] Whilst most of these are only blueprints at the moment, the efforts of the ICR in tandem with the cantons' administrative bodies is laudable. The ecological plans that have been implemented, including reforestation, designating natural reserves, increasing biodiversity, and protecting water sources from overuse, have been successful. To be able to make strides in ecological protection during a brutal civil war is previously unheard of, and these plans are long-term. Rojava's administrations and NGOs do not have ample funding, as the majority is needed for maintaining security through SDF forces. Therefore, these ecological plans cannot be implemented overnight — they will be the work of years, and in some cases decades of civil and governmental action.

2.3.3: ECONOMY AND GOVERNMENT POLICY

The economy of Rojava is principally centred around worker-owned co-operatives — against the two predominant forms of business across the world: privately-owned and state-owned. Whilst the constitution of Rojava does protect private property, it is somewhat uncommon — only 20% of farmland is privately owned, with a prohibition of any new private landowners. [86] These co-operatives businesses are tied to their local communes, with the leadership of the co-operatives elected by the commune. This ensures direct accountability between workers and the administrative teams within the cooperative businesses. Due to decentralisation, there are few, if any, 'chain' businesses with Rojava-wide market access — instead, the administrations of each communes are able to communicate with higher-level authorities to co-ordinate shipments of products (be it food, oil, or clothing) to communes lacking in these goods. [87] In 2016, TEV-DEM, the ruling coalition of the entirety of Rojava, stated that the division of profits in a business are as follows: 20% to the commune (a form of taxation), 30% to the co-operative (used for investment), and 50% to the shareholders (the workers, with an equal number of shares). [88] On top of their share of the yearly profits, workers also receive a monthly salary for their labour.

The two biggest industries in Rojava are oil and agriculture. The Jazira canton is considered to be the economic heartland of Rojava, with major production of cotton, wheat, and oil. [89] The Deir ez-Zor canton is similarly abundant with oil reserves, further providing income to Rojava, whilst the Afrin canton is centred around the production of olive oil and soap. [90] As of January 2019, it was estimated that Rojava produces 125,000 barrels of oil per day, with a production capacity of 400,000 barrels per day. [91] This estimate came from alleged correspondence between the SDC and an Israeli-American oil executive, Motti Kahan, in which the SDC promised Kahan rights to oil extraction and refining in the region. [92] However, the SDC responded to the leaked correspondence claiming it to be a 'fabricated on the part of Turkey…to pollute the image of the SDF'. [93] Kahan has similarly denied the correspondence, claiming it to be fraudulent.

The criticism of foreign involvement in the extraction of oil was one of the main issues that sparked the aforementioned Arab protests in Deir ez-Zor in March, 2019. This criticism was compounded by reports of collaboration between Rojava and the Syrian government in oil refinery, which critics argued was a betrayal of democratic confederalism, as it meant Rojava was assisting the Ba'ath regime — an ideology considerably opposed to that of Öcalan's. [94] Furthermore, there has been ecological criticism of Rojava's oil industry, as it is the largest contributor to global warming and air pollution in the region. However, whilst the criticisms of ecological hypocrisy and regime-cooperation may ring true, it is arguably a necessary evil to continue the economy of Rojava. Although figures on the exact income from oil have not been publicly disclosed by TEV-DEM, it was stated by Abdul-Karim Malak, the minister for oil and gas, that these two natural resources are the main source of revenue for Rojava. [95] If this is the case, then Rojava simply cannot afford to abandon these industries for the sake of ideological purity — they are an unfortunate realpolitik necessity for the continuation of Rojava, and TEV-DEM authorities (in tandem with the ICR) are working towards long-term goals of alternative energy sources and cleaner practices of oil extraction.

2.3.4: INTERNATIONAL RELATIONS

Rojava's international relations — both positive and negative — are one of the most complicated in the Middle East. The region's umbrella organisation of armed groups, the SDF, has been given military support from the Combined Joint Task Force (CJTF) — comprising of nine countries, including France, the UK, the USA, and Saudi Arabia. This has primarily been through air support, in which ISIS targets were engaged on the ground by the SDF, and bombed from the skies by the CJTF. However, beginning in April, 2015, the YPG (later the SDF) became increasingly co-operative with the Syrian government and Russian forces. This includes joint military operations (both on the ground and through Russian air support), alongside sharing major checkpoints on Syria-Rojava borders. This co-operation increased drastically after President Trump announced the withdrawal of US operations in Syria, causing the SDF to request the movement of Syrian government forces to the city of Manbij in the Aleppo governate. [96] However, the relationship between Rojava and the Syria-Russia coalition has fluctuated, with small-scale skirmishes between the groups continuing into 2018. Despite this fluctuation, it can be said that a moderately stable agreement between Rojava and the Syria-Russia coalition has been reached.

Rojava's primary enemy on the international stage is Turkey. This is primarily due to the YPG's links to the PKK — Turkey's deadliest terrorist organisation. Though the links are not formally recognised by any nation other than Turkey, there is evidence of co-operation between the two groups, with the PKK entering the Syrian Civil War to support YPG forces against ISIS in July, 2014. [97] Both groups also share the ideology of Democratic Confederalism, and both see Öcalan as their ideological figurehead. Furthermore, Turkey has been accused by YPG figures of seeking to destroy Rojava, less it further inspires Kurdish separatism in Turkey. [98] Turkey has even been accused by Rojava, alongside numerous Western media outlets and think tanks, of supporting ISIS against both Rojava and the Syrian government. [99]

This began during the Siege of Kobane in September 2014 (lasting under March 2015), where YPG fighters were attacked by ISIS sniper fire and vehicle-born improvised explosive devices (VBIEDs) from Turkish territory. [100] Turkish military action against Rojava culminated in Operation Euphrates Shield, beginning in August 2016 (lasting until March 2017). The result of the operation was Turkish-occupation (primarily under the guise of the Turkish-proxy Free Syrian Army) of approximately 2000 square kilometres of territory between Afrin and Manbij, crippling supply lines between YPG groups. [101] This crippling of supply lines allowed Turkish forces to capture the entirety of the city of Afrin from YPG control, and the majority of the Afrin canton in an operation lasting from January-March, 2018. [102] It is evident that Turkey's military action against Rojava will continue so long as there are perceived links between the YPG and PKK.

Rojava has been criticised by a small number of far-left commentators and media outlets for its alliance with the US and UK, describing it as antithetical to the principles of antiimperialism within Democratic Confederalism. [103] Whilst it is true that US foreign policy in the Middle East from the Gulf War onward could be described as imperialistic, their involvement with the YPG is far from the Western military actions in Iraq and Libya. Crucially, the US has not backed the YPG in a fight against the Assad government, and Western support for anti-Assad rebels has dwindled as the war has progressed. [104] Instead, the West's focus has been entirely on the destruction of ISIS in Syria, and not on any kind of regime change. Therefore, the criticism of Rojava as allying with imperialism is misguided.

CHAPTER 3: CONCLUSION

'Rojava is the world's most progressive democracy' [105]
Rahila Gupta, CNN Reporter

3.1: IS ROJAVA SUCCESSFUL?

Against all odds, the people of Rojava have created a society that should be the envy of much of the Middle East. With the rights of ethnic and religious minorities, alongside gender equality and direct democracy, constitutionally enshrined, Rojava is the most progressive region in the Middle East. The confederal model, despite its complexity and lack of centralised network, has proven itself to be effective in establishing direct democracy and providing government services to its citizens. Ethnic tensions are remarkably peaceful in comparison to the rest of the Middle East, and although some tensions do arise (such as the 2019 Deir ez-Zor Arab protests), they are few and far between. The progression of women's rights in the region is perhaps the most remarkable of all, with centuries-old patriarchal traditions overturned in a matter of years across Rojava, promoting gender equality and revolutionary feminism against the misogyny of much of the Middle East.

Rojava has proven itself to the rest of the world as a serious geopolitical entity through its relentless fight against ISIS. The YPG has been described as the most effective ground force against the terror group, rivalling the considerably more advanced Syrian government forces. This, in turn, has caused military support from both Western and Eastern nations — from the United States to Russia to Iraq. Although hindered by the onslaught of Turkish backed forces in 2016-2018 in the North-West of Rojava, the YPG has nevertheless been able to hold its own against a military force over five times its size. [106] The Democratic Confederalist model as practiced in Rojava is demonstrably successful. From the mass progression in minority rights to the extremely effective direct democracy, Democratic Confederalism has prevailed in the region. Whilst the ongoing civil war has regrettably necessitated some realpolitik measures (such as ecological damage through oil extraction and partnership with the Syrian Ba'ath government), Rojava is a remarkable and commendable example of anti-capitalism and directly democratic society in action. The region is proof that, with the right societal mindset, an alternative to capitalism, centralisation, nationalism, and indirect democracy is possible.

3.1.1: HOW CAN ROJAVA IMPROVE?

As previously detailed, there are legitimate concerns to the conduct of PYD and SDF authorities in the region. The forced conscription of youths, including underage Kurds and Arabs, is a severely disheartening practice by the SDF. Furthermore, the allegations of anti-Arab racism at the hands of PYD administrations is of concern, and further steps towards inter-ethnic peace are necessary in order to maintain multicultural society. The aforementioned oil extraction is a serious threat to Rojava's ecosystem and ecological ideals. Whilst it may be a necessary evil during the civil war, it is encouraging to know that TEV-DEM and the ICR are working towards moving away from reliance on fossil fuel extraction for economic income.

Finally, and perhaps controversially, it is necessary for Rojava to distance itself from the PKK. Turkey is currently the only serious threat to the security and integrity of Rojava, and will continue to be so as long as the YPG maintain military and infrastructural links with the PKK in Turkey and Iraq. Whilst this move may prove to be difficult, and sadly may not have any effect of Turkey's geopolitical actions against the region, it is hopeful that a peace process between Rojava and Turkey can take place in the near future.

3.2: WHAT DOES THE FUTURE HOLD FOR ROJAVA?

The landscape of Rojava has changed considerably over the course of the civil war — both literally and figuratively. Alliances have evolved over this time, with Rojava in the rare position of support from both the US and Russia. However, as the Syrian Civil War moves away from the threat of ISIS and towards Turkish-backed rebels in the North-West of Syria, Rojava faces a new challenge. Assad and Russia (both in a somewhat unstable alliance with Rojava) are engaged in fierce fighting against the Turkish-backed Free Syrian Army. It should be a clear-cut alliance — the YPG, Assad, and Russia working together to defeat the Turkish proxy rebels. However, it is highly unlikely that it will be this straightforward. Turkey is a member of NATO, and as it is engaged in a proxy war against Russia, could receive the backing of Western militaries in the future — this would mean the total withdrawal of the already dwindling support for Rojava by Western allies.

Conversely, Turkey has recently bought into the Russian S-400 missile defence system — at the behest of other NATO members, who have forcibly removed Turkey from the F-35 fighter jet program, fearing Russian access to the planes' vulnerabilities. [107] This could signal a shift of Russian support from Assad to Turkey — leaving the fate of the Kurds at the hands of Western allies, who may be unwilling to back Assad and Rojava, even against Russia. The complex geopolitical entanglements of all sides in the Syrian Civil War mean an accurate prediction of the conflict in even a year's time is not wholly possible.

Even if the civil war does end without Rojava's borders changing considerably, its fate under Assad is still unclear. The Syrian government, despite co-operating with Rojava, has vowed to retake ‘every inch of Syria’. [108] Whilst Russia has stated on numerous occasions that it is open to the idea of the federalisation of Syria, Assad has made his stance against it abundantly clear. [109] Nevertheless, it is foreseeable for Rojava to remain an autonomous region of Syria, as it is unlikely that Assad would wish to reignite a second civil war against Rojava having just come out of the first.

The fate of Rojava is unclear, and it depends entirely upon the next wave of the Syrian Civil War. The geopolitical alliances of the Western and Eastern nations will make or break Rojava — it is up to the international community to rally their governments against its destruction.

Alliance to Empire: the transition of the Delian League into the Athenian Empire in the period 478-454 B.C.

Alexander Norris

INTRODUCTION

The Delian League was formed in 478 BC as an alliance of Greek city-states in the immediate aftermath of the Persian Wars, primarily for the purpose of mutual military support and to launch campaigns against their common enemy, the Persians – however, by 454 (when the League's treasury with all its contents was moved from the eponymous island of Delos to Athens, after which all meetings of the League's assembly were held there too) it had de facto ceased to be a confederation of free states, but rather had become a federation under Athenian control: what would later be known as the 'Athenian Empire'. How this came about remains relevant to this day; in fact, understanding the process of how an alliance can become an empire by extrapolating from this ancient example can be particularly useful in comparing it with modern political alliances such as the United Nations, European Union and, most appropriately of all, NATO. What this essay will demonstrate is that the transition of the Delian League from alliance to empire was essentially through a centralisation of the League's functioning on Athens, caused in part by growing Athenian imperial ambitions and in part by the misjudgement and inaction of the League's other members.

FORMATION OF THE LEAGUE

The League itself was formed as a response to the Persian aggression that had resulted in their invasion of Greece in the Persian Wars, both the First Persian War (where they were defeated upon arrival at the Battle of Marathon in 490) and the Second Persian War (where they were defeated in 479 by land at Plataea and by sea at Mycale, after engagements at Thermopylae, Artemisium and Salamis).

REASONS FOR THE LEAGUE'S FORMATION

Ostensibly, the Athenians' reason for the League's formation was (as Thucydides, the main source for this period, writes) 'to take revenge for their losses by devastating the Persian King's territory.'[1] However, the word Thucydides uses for 'reason' ('πρόσχημα') can more accurately be translated as 'pretext' or 'excuse'; in fact, it literally means 'a screen in front of [something]' and so Thucydides would seem to be implying that in his view this was not the real reason why the League came into being, as Hunter R. Rawlings makes clear.[2]

Whether this view of Thucydides' is reliable or not is doubtful, as he would have been writing most likely about fifty years after the events, and so his judgement could have been clouded by hindsight (given the large degree of imperialism present in Athens at the time of the Peloponnesian War) – he implies that the Athenians had imperialistic ambitions from the very formation of the League, being concerned less with fighting the Persians than with furthering their own interests[3] and yet Thucydides would have been able to see the Athenians' later, and much more blatant, imperialism, and consequently may have assumed it always to have been present. He could nonetheless see it from the other perspective, as when he reports the Athenians as claiming, 'Fear was our first motive; afterwards honour, and then interest stepped in'[4] which hints that they only began trying to gain power for themselves once they realised they had the opportunity of doing it, having set up the League originally for quite a different purpose.

Whatever the Athenians' initial intention, it is revealing that such a relatively short time after the League's establishment historians as methodical as Thucydides were unable entirely to settle the matter. In this regard, it would appear to make scant difference which of the two is correct for the purpose of this essay, since regardless of when the Athenians' desire to subjugate their erstwhile allies emerged, the fact that it did eventually emerge, as well as the subsequent consequences of that emergence, are clear.

NATURE OF THE LEAGUE

Thus, despite the apparent purpose of the League to oppose the Persians militarily, it would seem from the Athenian campaigns that Thucydides lists directly after detailing the circumstances of the League's origin that they are more aimed at conquering other Greek city-states than at taking revenge on the Persians; for example, the expeditions against Skyros, Carystus and Naxos.[5] This contrasts strongly with the 'pretext' of the League (as Rawlings has again pointed out)[6] and makes it appear that the Athenians used all their new-found military might to add to their dominions within Greece – nonetheless, it's possible (as with the analysis of their primary intentions) that Thucydides is adding his own spin on the events in an attempt to illustrate the Athenians' disloyalty with hindsight, and it may be that he chooses to emphasise these events to demonstrate how shocking their actions were as abuses of their authority, as A. French has argued.[7]

The original nature of the League, moreover, would appear to all intents and purposes to have been temporary, inasmuch as it relied on hostilities between the Greeks and the Persians, and so after the alleged Peace of Callias in 449 the League ought to have lost its raison d'être; nevertheless, it continued, and this continuation shows it had come to be viewed by that point as a permanent bloc.

This permanence is evident in some of its ceremonies too, such as throwing heavy weights into the sea to symbolise states' allegiance, which are described as 'iron bars' by the Old Oligarch[8] and 'lumps of rock' by Plutarch.[9] This implies a lasting commitment to the League, because such weights would permanently remain in the ocean, and specifically (as the Old Oligarch mentions) 'to have the same friends and enemies [as the Ionians]'[10] which indicates that the League was viewed as an indefinitely enduring and binding union.

ORIGINAL MEMBERSHIP OF THE LEAGUE

In terms of the League's original membership, it is very difficult to determine which states were a part of it from the very beginning, given the lack of contemporary records that can be accessed; one likely view of the early League is that the 'core' states were many, if not all, of the 12 Ionian states, which is implied in the Old Oligarch's references to 'the Ionians' and by their strong links with Athens through the Persian Wars. Also, the Athenians may not have had a large amount of control over members of the League to begin with, which would explain why such prominent states as Samos, Chios and Lesbos were so eager to join.[11]

Some historians, such as Hammond, have proposed that the Athenians had multiple separate alliances with their various allies[12], but this is unlikely because of examples of cases where the allies of Athens work together to pursue a common goal, implying an alliance between them too (e.g. Plutarch 23.4). Others, such as Meiggs, have advocated for a very wide-reaching alliance right from the beginning encompassing states as remote as Cyprus[13] but again, this seems unrealistic considering the very few members mentioned by Plutarch and Thucydides regarding the League's formation.

Overall, while the original membership of the League is unclear, it would seem likely that it included several Ionian states in addition to the specifically referenced states of Chios, Lesbos and Samos, but unlikely that many more were a part of it to begin with.

STRUCTURAL DEVELOPMENT OF THE LEAGUE

The structure of the League and the roles its member states played developed over time, becoming more centralised on Athens as the Athenians grew in power, and thus having an impact on the way payment was made and collected to be put into the League's central treasury at Delos.

STATES' ROLE IN THE LEAGUE

The role individual city-states played in the early League is ambiguous, insofar as that Thucydides' accounts of how the voting systems were organised is ambiguous – thus, he describes the allies' voting as 'κοινῶν ξυνόδων', 'πολυψηφίαν' and 'ἰσοψήφους' ('in a common assembly', '[with] an excess of votes' and 'equal in vote')[14] from which it appears that they had theoretically the same power as the Athenians, although this could mean that each city-state including Athens had one vote in a common assembly, or alternatively it could signify – as Meiggs has suggested[15] – a bicameral assembly whereby an assembly of the allies had collectively the same power as the Athenian assembly.

On one hand, the bicameral arrangement had contemporary precedent, with Sparta using a similar system for its allies, whereas a 'one state, one vote' policy could have been humiliating for the Athenians to deny themselves the privilege (as de facto leaders of the League) of maintaining their own assembly's independence; on the other hand, though, such an arrangement would have been in turn resented by other strong states which felt Athens was growing too powerful, and the Athenians could have been persuaded to settle for an equivalent say in the League to each of their allies by the knowledge that they would be able to influence the voting of the smaller states, and so would be able to wield far more actual power in that environment than in one where there was a clear Athenian and non-Athenian distinction.

In fact, it seems that this may have been the case from the growth of Athenian power over the League, together with the resentment that some allies felt to being swamped by prevalent pro-Athenian sentiment in the Assembly, and so a system where each individual state had an equal vote seems most likely.

ATHENIAN HEGEMONY

Athens was clearly the pre-eminent member state of the League from its formation, but Athenian hegemony of the League is less clear so early on – the reason Athens was chosen over Sparta, which was renowned for its military capabilities and so would seem to be the natural leader of anti-Persian military action (as indeed it was in the Persian Wars) was because of the arrogant and aggressive behaviour of Pausanias at Byzantium, as Thucydides makes clear.[16] In theory, the Hellenic League that had opposed the Persians up to now continued under Spartan leadership until 461[17] and so it is curious that states should have felt so repelled by Pausanias' behaviour that they started another alliance.

Of course, after the great land power of Sparta, the naval power of Athens was the logical leader, which is why they formed the focal point of the League (as can be seen by the oaths sworn to them and their allies rather than to a list of members[18]). By the time of the Athenian Empire, though, there were undertones of political sovereignty which were not evident at the beginning of the League, with Pericles in 449 proposing a 'Congress Decree' that would have asserted Athenian superiority in no uncertain terms, and the use of imperialistic language in inscriptions such as the description: 'The cities which the Athenians rule.'[19] How this developed was in large part due to a development of the economic functioning of the League.

NATURE AND METHODS OF PAYMENT

The premise on which the League was founded was that member states would pay a certain amount as a tribute into a common treasury (which was located on the island of Delos, hence the League's name), which the Athenians would use to continue military action against the Persians. This payment was known as 'φόρος' and it was collected either in ships or in money, depending on the state. [20]

This 'φόρος' would be collected from each state by Athenian officials called 'ἑλληνοταμίαι' and given to the Assembly to deploy wherever was deemed best. The actual amount states would have paid originally is disputed, since while Thucydides claims the first collection was in total worth 460 talents[21] this seems unrealistically high, although Meiggs has pointed out that it could have been lowered in later years for that very reason.[22] Furthermore, given the large amount of money that had accumulated in Delos by 454 – 8,000 talents according to Diodorus Siculus[23] – it doesn't seem improbable for the Athenians to have requested such a large amount.

As it turned out, a large amount of that went to rebuilding Athens after it had been ravaged by the Persians in 480, especially the Acropolis, rather than on military funding, but there was little the other members of the League were able to do about it. This extraction of tribute from League members was certainly important in ensuring the Athenians had a solid financial foundation on which they could build their empire, but it should be remembered that the tribute was agreed to by those paying it, as opposed to other forms of Athenian control which Athens imposed arbitrarily on its allies; these are far more significant in turning the League from being an alliance into being an empire.

ATHENIAN INTERFERENCE IN ALLIED AFFAIRS

The interference of the Athenians in the member states of the League developed gradually, since at the beginning Meiggs has shown that 'the autonomy of members was taken for granted and there was no question of their leader interfering in political cases.'[24] However, by the time the Old Oligarch was writing he made the point that the Athenians 'force the allies to sail to Athens for judicial proceedings'[25] even going so far as to refer to them as 'οἱ σύμμαχοι δοῦλοι τοῦ δήμου τῶν Ἀθηναίων' ('the ally slaves of the people of Athens').[26] Therefore, it is fundamental for understanding how the League became an empire to grasp how the Athenians were able to interfere in the affairs of other states.

JUDICIAL AUTONOMY OF STATES

The judicial interference of which the Old Oligarch speaks, where the Athenians would impose their judicial system on their allies, is well documented in instances such as the later Chalcis Decree of 446, where autonomy was granted to states 'except in cases involving exile, death, and loss of rights' and also required the inhabitants to take an oath 'not to revolt against the Athenian demos' and 'to be obedient to the Athenian demos.' Another instance of this was the Phaselis Decree of c.469, where in much the same way cases normally heard in Phaselis were transferred to Athens – an apparently clear example of the violation of an ally's judicial autonomy.

Alternatively, though, it could be viewed as merely a form of judicial standardisation, which involved a certain degree of centralisation, and de Ste. Croix haas argued that this arrangement actually favoured Phaselis because the Athenian polemarch's court was known for giving preference to foreigners.[27] The Athenians prided themselves on their legal impartiality (as Thucydides himself, hardly the most pro-Athenian commentator, claimed: 'In Athens the laws [are] impartial'[28]) and the legal standardisation of the League could be argued to have benefited all its members.

Nonetheless, regardless of the Athenians' intentions in imposing their laws or the potential benefits of such an imposition, it remains clear that by doing so they aided, consciously or unconsciously, the influence of the Athenians and ultimately the formation of the Athenian Empire. More significant than this, though, was the Athenians' interference into the political functioning of other states.

POLITICAL AUTONOMY OF STATES

Although member states of the League comprised a large variety of seemingly autonomous political systems – ranging from the Athenian democracy to states such as Mytilene and Chios which had managed to hold on to oligarchy until the time of their revolts, in 428 and 412 respectively – there are also examples of the Athenians inflicting democracy on the nominally independent states within their influence. For example, the Decree of Erythrae of 453 imposed on Erythrae a democratic constitution establishing similar political institutions to Athens (including the use of lot, as well as a military garrison led by a 'φροὺραρχος', or governor, who also assisted the 'ἐπὶσκοποι', or overseers, in their capacity of ensuring Athenian demands were met).

There are further instances of political interference in the cases of states being forced to be members of the League, such as happened to Naxos after its inhabitants revolted against Athens after the battle of Eurymedon; after laying siege to the island, the Athenians were able to prevent them from leaving the League. Other examples of this include the states Melos and Calystus which were compelled to join the League because they were in Athenian eyes potentially hostile, and consequently it was politically expedient for them to have an alliance with Athens.

Another way Athens maintained a political presence in other states' affairs was by establishing officials – such as 'ἐπὶσκοποι' (overseers), ' φροὺραρχοι' (officers in charge of garrisons) and 'φὺλακες' (guards) – who would ensure that the states complied with Athens' wishes. Similarly, multiple colonies were set up, of which the most significant ('κληρουχὶαι' or cleruchies) would be inhabited by Athenian citizens who kept their citizenship with all the rights that it entailed, as opposed to other colonies where they would be citizens of those colonies and so forfeit their Athenian citizenship. Such cleruchies were set up on islands such as Scylos and Melos, and because their inhabitants were Athenian citizens they remained directly under Athenian control, maintaining an Athenian presence wherever they were.

The most significant example of Athenian interference in states' political autonomy was through the Delian League itself, whereby the Athenians necessitated the transition of states' status within the League from ship-paying to money-paying (such as Thasos in 465). The importance of this lies in the fact that Athens had much more control over the use of money than of ships, and hence much more influence over the money-paying states than over the ship-paying ones. According to de Ste. Croix, moreover, this was stressed by Thucydides, who consistently 'conceived the condition of the tributary allies, whom he describes as "ὑποτελεῖς φόρου, φόρῳ ὑπήκοοι"[29] as one of "δουλεία" but except on one occasion he is willing to call the naval allies "αὐτόνομοι" and "ἐλεὺθεροι".'[30] De Ste. Croix admits that there was no official distinction between the levels of autonomy a state enjoyed since each case was dealt with on its own basis[31], but nonetheless the way Athens was able to twist states' arms into obligations of their membership in the League that they had never originally accepted through the recently acquired financial and military influence that supremacy in the League gave them was fundamental in the League's transition from alliance to empire.

CONCLUSION

In conclusion, therefore, it was this centralisation of the League's resources that enabled the Athenians to have so much say in the running of the League which also enabled them to turn it into an empire; this was partly a result of Athenian ambitions, but also of Athenian fear. As Thucydides has the Athenians say to the Spartans, 'You turned against us and begun to arouse our suspicion: at this point it was clearly no longer safe for us to risk letting our Empire go, especially as any allies that left us would go over to you'[32], and later adds about the Empire that, from Pericles' point of view, 'it may have been wrong to take it; it is certainly dangerous to let it go.'[33]

A third factor crucially comes into play here, and that is the League's members' inaction in providing checks and balances to Athenian power but preferring to mind their own business and allow the Athenians to gain more and more power in the League. This too is highlighted by Thucydides, when he states of the Empire that 'for this position it was the allies themselves who were to blame'[34] implying that had they acted in opposition to the Athenians they may have been able to thwart their imperialistic aspirations. Nonetheless, they did not, and so the opportunity was made available to the Athenians to centralise power in the League at Athens, a centralisation that ultimately resulted in the transferral of the League's treasury to Athens in 454, after which the Congress of Allies ceased to meet and the Athenian assembly alone decided what should be done with League funds.

It is this centralisation that was the key reason why the Delian League became the Athenian Empire, and this is why the question continues to have relevance today; in order to prevent modern alliances (such as the UN and NATO) from becoming empires, it is imperative to keep any centralisation to the minimum necessary to organise the alliance. As the Athenians' case shows, with this centralisation comes the inevitable imperialism, and accordingly it can be seen that the transition of the Delian League from alliance to empire came about through the concentration of the League's authority in Athens, which in turn was facilitated by the inertia of the League's members and their failure to curb Athenian power substantially enough to safeguard their own long-term political independence.

VEGF-D function and mechanism in coronary angiogenesis

Kai Rohde

ABBREVIATIONS

- □ VEGF = Vascular Endothelial Growth Factor
- □ VEGFR = Vascular Endothelial Growth Factor Receptors
- □ PDGF = Platelet Derived Growth Factor
- □ C-FIGF =C-Fos-Induced growth factor
- □ AdVEGF-D = Adenovirus-Mediated VEGF-D
- □ CAD = Coronary Artery Disease
- □ RFA = Refractory Angina
- □ ELISA = Enzyme-Linked Immunosorbent Assay
- □ CTO = Chronic Total Coronary Occlusion
- □ CAC = Coronary Artery Calcium
- □ EF = Ejection Fraction

Within the text, {-} refers to appendices while [-] refers to references.

ABSTRACT

The aim of this study is to build up a preliminary understanding of the function and importance of the vascular endothelial growth factor VEGF-D in natural angiogenesis before a clinical trial on it takes place in the coming years. This study is part of a larger phase II study evaluating the "safety and efficacy of catheter mediated endocardial adenovirus-mediated vascular endothelial growth factor-D (AdVEGF-D) regenerative gene transfer in patients with refractory angina to whom revascularisation cannot be performed" [1], being done by Kuopio University Hospital and eight other institutes; including my hosts Queen Mary University of London [1]. Currently not much is known about the process of angiogenesis and the function and mechanism of the vascular endothelial growth factors, even less on VEGF-D. Therefore, this study will help build an understanding of the effects of VEGF-D to assess its importance as a treatment to many coronary artery diseases.

ACKNOWLEDGEMENTS

I would like to extend my deepest gratitude to Professor Anthony Mathur for allowing me to partake in this project and supervising me through this incredible experience. In addition, Dr Deshan Weeraman, who allowed me to base my work on his thesis and show me a fascinating in-depth look at how research is conducted within medicine. I also would like to thank James Whitefold for conducting the difficult procedures and helping to explain all the details of the experiment. Finally, I would like to thank all the staff at both Bart's cardiac section and Queen Mary of London University for their hospitality.

KEY WORDS

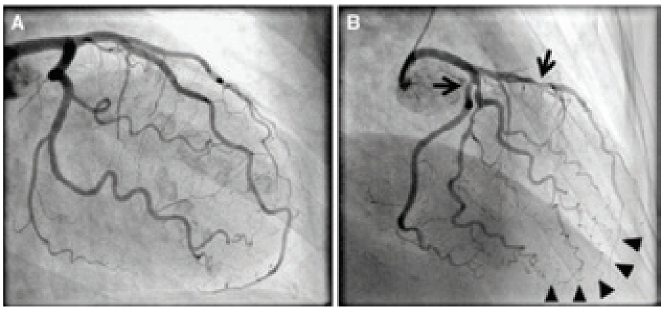
VEGF, VEGFR, VEGF-D, VEGFR-2, VEGF-DΔNΔC, CAD, CAC, angiogenesis, atherosclerosis, calcium, coronary artery

CORONARY ARTERY DISEASE (CAD)

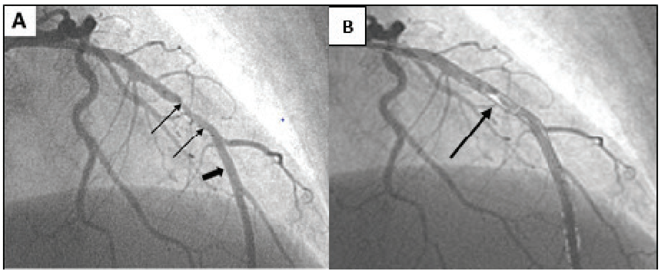
The heart is the main organ involved in the circulation of blood around the body beating around or more than 3 billion times in a lifetime. This requires a huge amount of energy to achieve and so the heart must be constantly fed oxygen in order to supply the muscles with energy. This is achieved through a dense network of arteries and capillaries which cover the heart known as the coronary arteries and are fed by the residual back flow from the Aorta during diastole. However, over time these arteries and capillaries often become coated with plaque and other molecules known as atherosclerosis, reducing the flow of blood or even ceasing it. Limiting the blood supply results in a reduction of oxygen, transported by RBC's bound to haemoglobin, to the heart muscles which can lead to weaker pumps and cell death through oxygen starvation (hypoxic). The dead heart muscle cells form an ischaemic tissue or scared tissue which isn't able to contract, weakening the heart and leading to angina [4]. These hypoxic conditions stimulate the release of HIF's (hypoxic inducing factors) which in turn stimulate the release of VEGF's with the function of promoting new blood vessel formation around the blockages (angiogenesis) and restoring blood supply to the ischemic area [17].

CAD develops when the arteriole wall is damaged allowing small molecules (often cholesterol) to get into the endothelium. A type of WBC (white blood cell), called a macrophage, is then signalled to the affected area and begins to break down the molecules. Often indigestible, molecules such as cholesterol, "bloat" the macrophages turning them into Foam cells (often referred to as "plaque"). These build up a small mass underneath the vessel's wall all the while releasing more signalling chemicals to promote more WBC's to the area. This cycle continues, the build of cell debris, macrophages and foam cells becoming an atheroma, which eventually pushes into the lumen of the vessel and begins impeding blood flow [20].

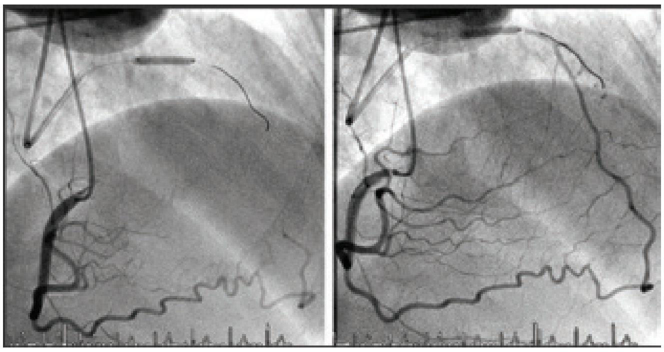
As plaque is constituted out of fatty molecules among other substances. It is no surprise that "fat" conditions, such as Diabetes and a high cholesterol level (from conditions such as hyperthyroidism), can act as serious risk factors and speed up the process of atherosclerosis [17] [18]. Other risk factors include smoking and high blood pressure, both of which can cause damage to the arteriole walls, induces the process of atherosclerosis. Lastly, age and genetics can play a large role in CAD and CHD (Coronary Heart Disease). A family history in CAD, CHD or other risk factors (such as a family history in diabetes or hypertension) can also lead to increased atheroma development for similar reasons as explained. Age on the other hand acts as a risk factor



[6] Angioplasty showing coronary arteries of two patients. A: healthy patient with wide dilated arteries. B: patient with CAD, narrowing of arteries shown by the arrows and has developed angina pectoris.



[9] Angioplasty of a single patient. A: thin arrows representing narrowing of arteries and bold arrow pointing towards a healthy artery. B: same patient, however, white areas (as directed with arrow) indicate areas of atherosclerosis – directly relates to image A's first arrow from left



An angiogram of a patient before and after angiogenesis has occurred. Where blood (which is infused with a dye) previously couldn't get through, the blood now can. Hence, the addition of new blood vessels not seen beforehand.

due to the natural wear and tear that the arteries sustain through one's lifetime. Almost impossible to escape, age is a key factor in determining the risk of CAD in a patient, and although lifestyle changes can reduce it, they can't inhibit it completely. Lifestyle changes include, healthy diets with less fatty foods and regular exercise [19].

If left untreated, CAD and CHD can lead to many problems and severe conditions. With the reduction of blood flow to an area of heart's myocardium, some heart muscle cells will die of oxygen starvation. This can result in scarring tissue of ischemia which reduces the heart ability to pump. This can cause angina, severe pain in the upper chest region, sometimes spreading down the arms or up the neck, which is relieved with rest. CAD can also have knock on effects including hypertrophy (enlargement of muscle) in any muscular area of the heart in order to compensate for the reduced cardiac output. Similarly, this can lead to increased blood pressure as well. It is also possible for a clot or thrombosis to form on the atheroma, which if broken off, can cause a MI; myocardial infarction (or otherwise known as a heart attack).

Current methods of treatment for CAD include CABG (coronary artery bypass graft where a section of vascular culture is removed from one area of the body and attached to the affected area of the heart in the hopes of stimulating blood vessel growth), PCI (percutaneous coronary intervention which includes angioplasty and stenting) as well as numerous drugs including: anticoagulants, beta blockers and anti-platelet drugs like Clopidogrel. Limitations within each treatment include scarring and infection risks linked with surgery as well as the risks of long-term blood thinner medication, which will see the patient bruise and bleed easily.

CAC (coronary artery calcium) is one of the major factors of CAD and CHD. During an atherosclerotic plaque's formation, it undergoes numerous stages of rupturing and regeneration followed by calcification. As calcium is not found in arteries normally it can be used as an indicator of CAD [21]. The actual process of measuring a patients CAC levels involves the use of an ultrafast CT scan of the chest (the process is required to be superfast in order to capture pictures of the beating heart) [22]. Any calcium will show up as white dots or streaks on the (usually) grey canvas [23]. The results are recorded numerically, with scores ranging from zero to infinity. Scores are assigned based off the Agatston score, which uses the "weighted density and area of the calcification identified" [24] to determine the severity of the artery's calcification. These scores are presented in ranges [24]. A score of zero represents no calcium in the arteries, whereas, a score over 100 is troubling and would likely be followed up with a meeting with the patient's doctor [22]. Despite results under 100 being less worrisome, the test records a range and so even a low measurement can pose a risk of a developing atheroma and CAD. As such, a "45-year-old" with a CAC score of "25" is still a "major concern" [22].

Coronary artery calcium score	Calcification grade = risk of imminent coronary event
0	None
0 - 10	Minimum
11 - 100	Mild
101 - 400	Moderate
401 - 1000	Severe
> 1000	Very severe

[24] A table showing the scores within their corresponding ranges and the risk of an adverse CAD related events. Here 100 is considered mild, however, in a younger patient this may be more problematic and require lifestyle changes as opposed to in an elderly.

VASCULAR ENDOTHELIAL GROWTH FACTORS (VEGF'S)

VEGF's are a group of glycoproteins and a subclass of PDGF [2] fitted with a characteristic cysteine knot. They are present during both angiogenesis (the de novo formation of blood vessels from existing blood vessels) and vasculogenesis (the de novo formation of blood vessel from nothing). VEGF's are able to bind to certain receptors most notably VEGFR's, a group of receptor tyrosine kinases (RTK's) of which there are three main ones: VEGFR's -1, -2, and -3 [8]. Some isotopes of VEGF's bind to co-receptors such as neuropilin 1&2 (NRP-1 and -2)

VASCULAR ENDOTHELIAL GROWTH FACTOR RECEPTORS (VEGFR'S)

VEGFR's are found predominantly on endothelial surfaces, although it is important to note that, VEGFR-1 and VEGFR-2 is more localised to vascular endothelial surfaces as opposed to VEGFR-3 which is found predominantly on lymphatic endothelial surfaces and is a known contributor to lymphomagenesis [10] (the growth and development of lymph vessels [11]). During angiogenesis VEGFR-2 is the more significant receptor, controlling endothelial permeability, proliferation and cell migration [7]. VEGFR-1, on the other hand, increases morphology of blood vessels as well as acts as a decoy to monitor the action VEGFR-2 receives and regulate angiogenesis, binding to both VEGF-A and –B [7].

VASCULAR ENDOTHELIAL GROWTH FACTOR D (VEGF-D)

VEGF-D belongs to the family of proangiogenic factors; VEGF. VEGF-D is a protein coded for by a 38, 817 base pair long length of DNA found on the X chromosome part 22.2 [12]. VEGF-D is comprised of a central VEGF homology domain (VHD) and two additional propeptides, a C- (carboxyl) and N- (amino) terminal polypeptide extension. This protein can then undergo "maturation", whereby the C- and N- are cleaved by specific proteases [13]. The resulting molecule VEGF-DΔNΔC has a higher affinity for the receptor VEGFR-2 [14]. VEGF-DΔNΔC, like VEGF-D, is able to promote both angiogenesis and lymphomagenesis [14]. VEGF-D, or C-FIGF, is a ligand that closely resembles another VEGF, VEGF-C, in its shape, structure and function [2], both ligands are comprised of a VHD flanked by a C- and N- propeptide [15]. VEGF-D was called C-Fos- (an oncogene) -induced growth factor, but was later changed to VEGF-D due to its resemblance to VEGF-C. Like VEGF-C, VEGF-D is able to bind to two VEGFR's; VEGFR-2 and VEGFR-3, which are both located on the endothelial surface [5], the VEGF-D ligand is non compatible with the VEGFR-1 shape and so can't dimerize with the receptor [7].

PRACTICAL PART OF PROJECT

AIM:

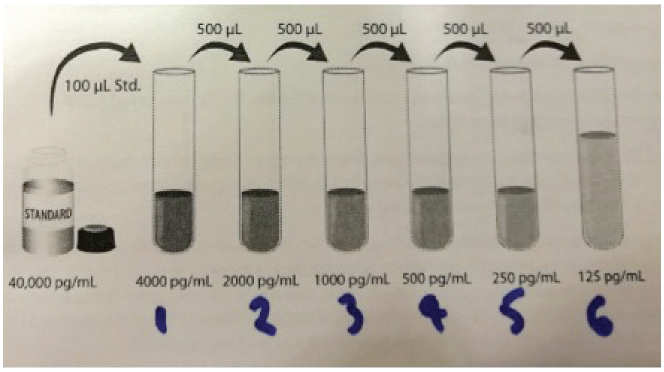
To discern the natural level of VEGF-D in patients with CAD or factors which increase the chance of CAD.

HYPOTHESIS:

VEGF-D concentration would be higher in patients with high CAC scores than that found in healthy patients. The reasoning behind it being that with CAD comes the hypoxic conditions caused by reduced blood flow through the coronary arteries. As I have found through research these hypoxic conditions stimulate the release of VEGF's in order to stimulate angiogenesis and vessel development in an attempt to restore blood flow to areas of oxygen deprivation.

METHOD BACKGROUND:

We opted for an ELISA test to find the levels of VEGF-D in the patients. The ELISA test, or enzyme linked immunosorbent array, used a microplate pre-coated with a monoclonal antibody specific to human VEGF-D [16]. Prior to this blood was extracted from 150 volunteers with the following selection condition of anginal chest pain requiring urgent investigations. The blood was taken from the brachial vein in the anti-cubital region (inside of the elbow) and centrifuged for 15 minutes at 1000 x g in order to achieve blood plasma (yellow) containing the VEGF-D [16]. This was then pipetted out of the (centrifuged) blood collection tubes and into small vials for storing, which were then refrigerated. A VEGF-D standard was used, starting with a 40,000 pg/mL standard, and samples of concentrations ranging from 4,000 pg/mL to 125 pg/mL were obtained via serial dilution [16]. The importance of these standards were to act as a contrast to the wells containing the volunteers blood plasma. The ELISA test is usually analysed based on colour, in this particular case yellow, with a stronger yellow colour resembling higher levels of VEGF-D. During steps in this process the samples were kept refrigerated at -80 degrees celcius at either St Bartholomews Hopsital or at Queen Mary of London University's John Vane building.



[16] An extract from the manual showing the process of serial dilution to create the 7 standards (the 7th standard, 0 pg/mL included no standard).

VOLUNTEER METHODS:

Every volunteer gathered for this experiment was asked about certain medical conditions. This was to enable us to build a picture of the physical and mental condition of each volunteer as to possibly locate links between factors and VEGF-D levels. The conditions recorded included: wether they were diabetic; wether they had hypertension; wether they had high cholestrol; if they had any family history with CAD; wether they smoked; wether they ever had a myocardial infarction; wether they ever had PCI (percutaneous coronary intervention), balloon angioplasty, stenting or CABG (coronary artery bypass graft); their CAC score; the severity of CAD (syntax); wether they had CTO or ischemia; their kidney function and finally, their ejection fraction (how much blood flows through the vessel, norm around 60%). In addition, each volunteer was tested for Rentrop, the level of angiogenesis that had occurred naturally.

ASSAY PROCEDURE:

Once all the reagents, standards and samples have been brought to room temperature, the ELISA test process can start. Primarily, apparatus and solutions were organised before starting the experiment, including trimming the microtiter plate to the desired numbers of wells. To begin the experiment, 100μL of Assay diluent RD1X was added to each of the wells. After this, 50μL of either a standard or a plasma sample were added to each well, where ant VEGF-D is able to bind to the monoclonal antibodies attached to the plate. An adhesive strip was attached above the microplate and the entire plate incubated for 2 hours at RT (room temperature). While waiting preparation for the next stage of the experiment can be completed. After the incubation period the wells were aspirated and washed. The wash was completed by using 400μL of Wash Buffer per well and any liquid post-wash was completely removed. This process was repeated 3 more times, totalling four washes. After the last wash was completed, the microtiter plate was aspirated a final time and inverted to blot any Wash Buffer from the wells. The next step was to pipette 200μL of VEGF-D conjugate into each well. The well was then covered again by an adhesive strip and left to incubate for a further 2 hours at RT. After the incubation period, the aspiration and wash procedure were repeated, with careful handling again as to ensure wells are completely dry afterwards. After this 200μL of substrate solution is pipetted into each well and incubated at RT for 30 minutes, however, no light was allowed to reach the microplate, so it was kept under a layer of foil. after the 30-minute incubation stage, 50μL of stop solution was added to each well and a colour change observed from blue to yellow. sometimes wells did not exhibit the correct colour change, however, after a gentle tap to the microtiter plate (to ensure complete mixing) all results exhibited the correct change. A microplate reader set at 450 nm was then used to determine the optical density of each well, the resultant value is concentration of VEGF-D from the blood sample [16]. Three results were acquired for each volunteer and an average was taken. [Methodology taken from procedure in kit manual as well as personal experience]. All aspects of the experiment were completed by James Whiteford from Queen's University of London due to both the complexity and cost of the kit and experiment which was undertaken.

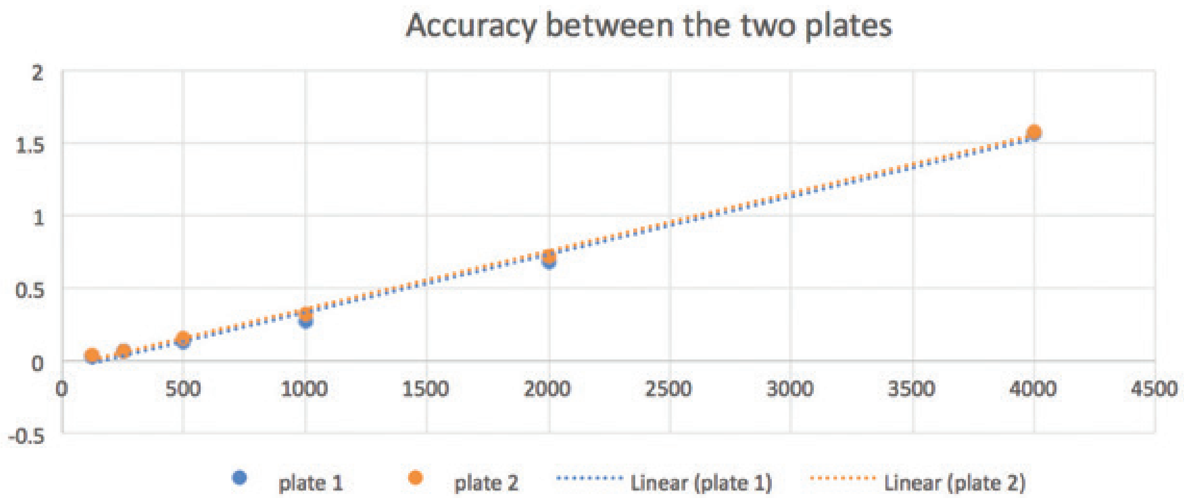
RESULTS

The full list of data can be found in the appendices {1} and photos {2}. I mainly used the patients' CAC score and their averaged zeroed VEGF-D concentration.

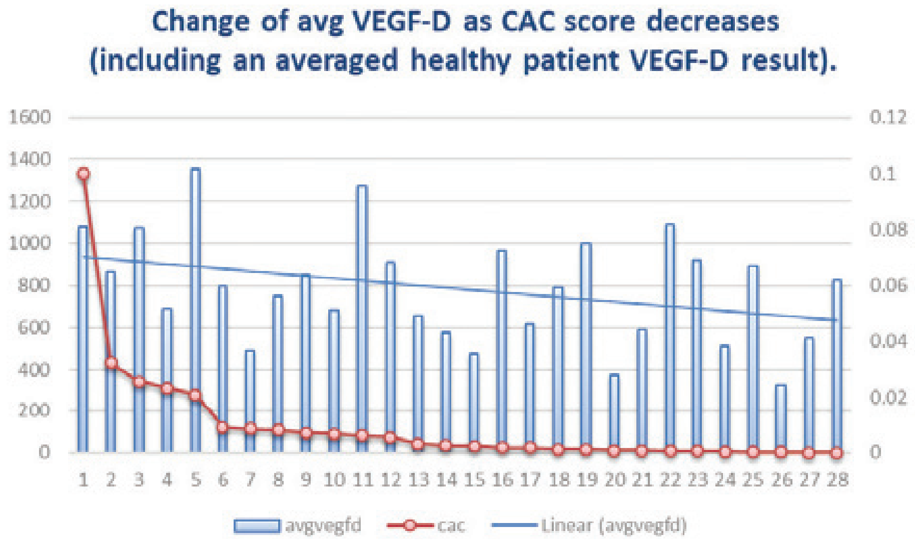
GRAPH ANALYSIS

Below are two graphs I was able to map out. The first is the accuracy between the two plates which I calculated {3} using the standard values. The result is a very close reading for both plates, which allows me to confidently compare results from the two plates.

The other graph I produced are my findings from the experiment. In this graph I have plotted descending CAC score against their respective average VEGF-D concentrations. The result is a clear yet not-significant trend downwards, so as the patients CAC score decreased their respective VEGF-D concentration was found to as well. This is really impressive as not only was my sample size small, but this is also a preliminary study into the subject, yet it still produced a notable trend. Two more graphs {4} {5} also show this trend, albeit in a different way. Graph {4} shows the correlation between a descending VEGF-D concentration and the respective CAC scores, producing yet again a noticeable trend downward for CAC score this time.



Graph shows difference between the two assays is negligible and so we are able to compare results between plates.



A graph I mapped out by arranging the CAC score in decreasing order and converting the corresponding VEGF-D concentrations into a combo chart on Excel. The graph had an R value of 0.3447 and a P value of 0.064573, which is close but greater than a 0.05 significance level. As such, these results are insignificant.

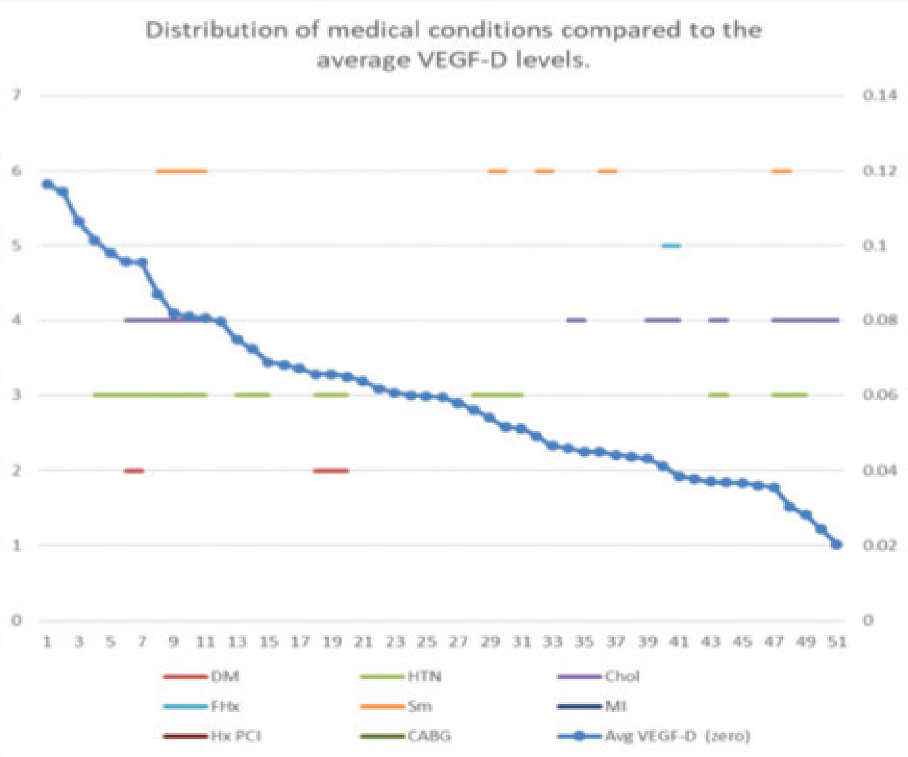
DISCUSSION

Although I lacked the entire 150 volunteer samples, I can still deduce conclusions and evaluations from these results, albeit less reliable. In addition, due to the lack of a measure for CAD I've used the patients CAC score as a measure of CAD.

In general, my results tended towards my hypothesis; that the VEGF-D concentration would increase as CAD, represented by CAC score, increased. Due to the clinical use of a CAC CT scan in identifying CAD, it is plausible for me to assume it as a marker for CAD. Due to the range of VEGF-D values I received for patients with a CAC score of 0, I decided to first remove their values and see the trend with only a decreasing CAC score {6}, I then calculated an average VEGF-D value (turning out at 0.06054 pg/mL).

Possible reasons for the fluctuations in my data could have come from different points in production and use of VEGF-D. it is possible that a healthy person could have genetic defaults which lead to them producing more/less of VEGF-D, distorting their result in one direction. Similarly, it is possible for patients with CAD to be producing VEGF-D but then using it up to produce the collateral vessels, falsely advertising the patients VEGF-D concentration as lower rather than higher.

VOLUNTEER ANALYSIS



This graph compared the distribution of medical conditions against VEGF-D concentrations. It can be seen here that both hypertension and diabetes are found predominantly in patients with higher concentrations of VEGF-D. although both smoking and high cholesterol levels are found in patients with a high VEGF-D concentration, it is also found in patients with low VEGF-D concentrations and so doesn't show a significant trend.

CONCLUSION

Based on the results I have produced, as CAC increases the concentration of VEGF-D also increases. Due to the relation between the CAC score and possible CAD and atheroscleroma, it is a plausible evaluation to say that VEGF-D concentrations increase during CAD. Through research a potential mechanism that I have come up with for this is that during CAD, the narrowing of the blood vessel lumen (due to atherosclerotic plaques forming – hence being picked up by the CAC CT scan) causes hypoxic conditions which could lead to the stimulation of VEGF-D release.

EVALUATION

Due to these findings it is possible that VEGF-D could be used as either a treatment or a determining marker for CAD or CHD in the future. However, one must be aware that these findings could be affected by other factors.

How is China's presence on the global stage and its internal policy affected by its geography?

Winner of the Trinity Group Geography Essay Prize2020

Freddie Lonie

INTRODUCTION TO CHINESE GEOGRAPHY

China's geography has been the mould for its presence on the global stage and shaping its internal affairs for many centuries. Chinese borders have been altered with the growth of other regional powers such as the Mongol Empire, the Indian Subcontinent and Russia. Despite this, the official borders of China have remained almost unchanged since the 1950s. This was when Chairmen Mao was consolidating his power in the People Republic of China. However, smaller border disputes have been an irritation for the Chinese government. With each shift in its geography, China unearths new problems on the global and domestic stage but also finds solutions for existing ones.

With a landmass of approximately 9.597 million km² and a population of 1.404 billion[1], China is an immense country with many different cultures, biomes and landscapes. This has provided China with a unique internal socio-politico-economic climate- fashioning the country's internal affairs for many dynasties and states. Each individual state has distinct geography which results in an equally distinct internal political climate. This ranges from the Shang Dynasty in 1766 BC to the current People's Republic of China.[2]

THE 3 GEOPOLITICAL IMPERATIVE

Geopolitical imperatives are the aims of Chinese international and internal policy. These are formed as a result of its Geography. China has three fundamental geopolitical imperatives. Firstly, to maintain internal unity in the Han Regions, secondly to maintain control of the buffer regions and thirdly to protect the Chinese borders (especially the coastline) from foreign encroachment.[3] Almost all Chinese foreign and internal policies should be viewed through the lens of these imperatives because they are key to understanding Chinese objectives home and abroad. For example, some perceive the intervention by China in the 1950s Korean War as colonial expansion and the reinforcement of communism in the region. However, if this military intervention is viewed through the lens of the Chinese geopolitical imperatives, it is clear that China was only trying to protect its borders from the western led United Nations forces in Korea- a border nation of China. Furthermore, North Korea was a semi-buffer state. Although fully autonomous, China had an extensive level of influence over the Communist government of North Korea. This demonstrates how the Chinese Geopolitical Imperative, to maintain control of the buffer regions, strongly impacts China's presence on its internal political affairs and indeed, on the global stage.

PROTECTION OF 'CHINA PROPER'

One of the key areas where China's geography shaped the internal climate of China and been a beneficiary to the state is through the protection of the China Proper region. This is considered the nucleus of modern China. Located on the east of the country, this area is where 95% of the population live, where most of the economic activity occurs and where there is the majority of the arable land is. Almost all major Chinese cities are located in this region. Figure 1 below illustrating the distribution of major cities in China, demonstrates that all but 6 of the major cities of China are located in the China Proper area.

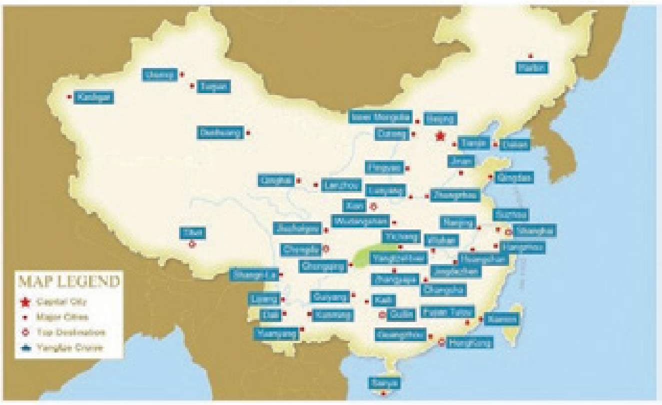


Figure 1: A map to show the Distribution of Major Cities in China[5]

The term 'China Proper' was coined by western writers in the 18th century to describe the Manchu Qing dynasty. It is used to express the distinction between the inner regions and the frontier (buffer) regions like Tibet and Xinjiang. The China Proper region is determined by the areas to the east of the Chinese Isohyet. An isohyet is an isoline separating areas by annual rainfall, east of this isohyet is where more than 15 inches of rain falls each year and west of which the annual rainfall is less than 15 inches.

The China Proper region is dominated by Han Chinese, who make up 92% of the population.[6] As a result of this population dominance, it is the Chinese government's priority to protect this heartland. The ruling elite of China often view the outer regions as buffer zones with their primary purpose to protect the economic hub of 'real' China. The resulting domestic policies reflects this view illustrating the extensive effect that geography has on the formation of internal policies of China.

Historically, the Han civilisation has faced considerable challenges from the nomadic tribes that surrounded the dynasty. In order to secure the Han core, China developed a policy to fight its neighbours in Tibet, Manchuria and Xinjiang. This was intended to establish a system where Han China had minimal, but enough, influence over the area in order to stop these buffer regions from attacking but where minimal military presence was needed. This control was achieved through a nominal tributary system, where the tributaries (the modern-day buffer regions) sent tokens of submission to the superior power of Han China. This gradually brought the neighbouring states under the influence of the Hans until the Han Chinese State eventually engulfed them. However, these regions are not entirely controlled by China, they are semi-autonomous. China has enough influence to stop independence campaigns. The geographical proximity of these threatening primitive civilisations dictated early Han Foreign Policy to control these buffer regions. These barrier territories are exemplified in the blue areas on the map below (fig. 2).

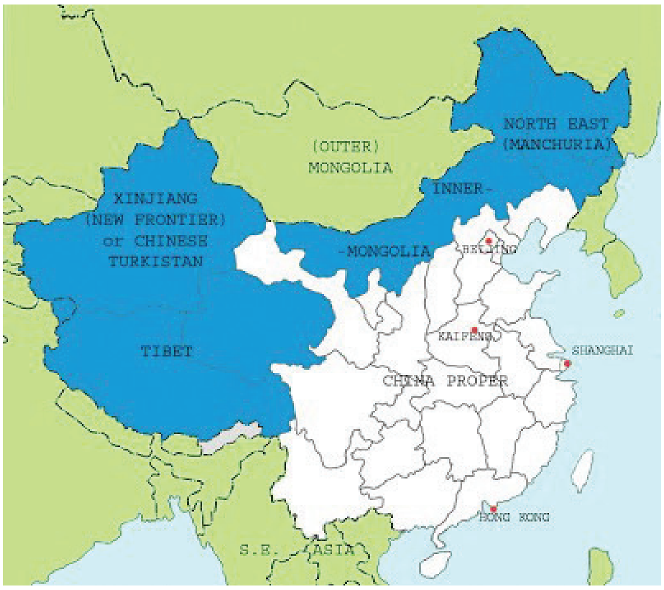


Figure 2: A map to show the Buffer regions of China[7]

The buffer region's geography provides security for the Hans because they extend the Chinese borders towards the natural borders of the deserts and savannah of Xingjian, the jungles of South China and the Himalayan mountains of Tibet. These physical barriers provide exceptional security for the country because it would be extremely difficult to lead a substantial military force through these natural geographic fortifications. The natural barrier that the physical geography provides is reflected in the internal policies of China. This is because the natural buffer serves as a replacement for artificial defence. Although China controls the regions of Tibet, Xingjian, Manchuria and Inner Mongolia, it has a point of friction on the South East border with Vietnam.

The buffer regions containing perilous territory are not only there to extend Chinese borders but also to make the supply lines for a potential invading force almost impossible to maintain. A comparison can be made with the German invasion of Russia in Operation Barbarossa in 1941-44. When the terrain became hostile in the winter months the German invading force (one of the leading militaries of its time) was unable to sustain the supply line. This devastated the whole invasion (in particular affecting the battle of Stalingrad) and turning the tide of World War 2 in favour of the allies. The terrain in the buffer regions of China is even more intimidating than the Russian winter and, it is not feasible to foresee a successful invasion of Han China from the west. This idea has prompted the creation of the Island of China Theory. This theory considers modern China as an island but not one surrounded by water, but rather impassable land. Figure 3 is an illustration of this, the impassable regions of China have been artificially replaced with a blue colouring to represent the isolation of modern China as it is surrounded by ocean to its east and indeed hostile landscape to its west.

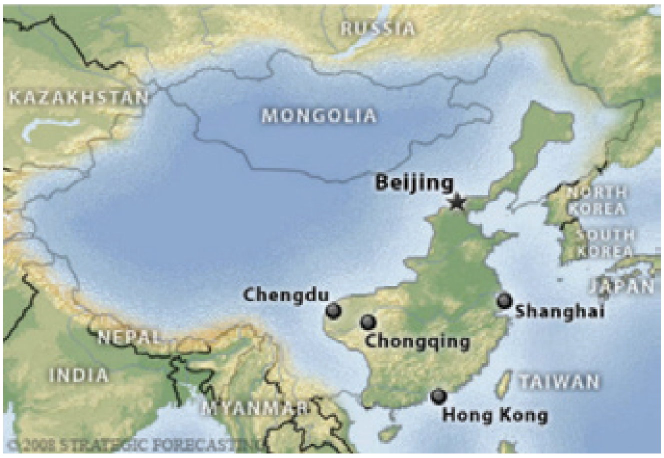


Figure 3: A map to show the Island of China theory illustrating the impassability of the buffer regions[8]

Many military strategists often state that the weakest border of China is to the sea. China's most significant war was after the Japanese invasion in 1937 which came from the sea. This led to the subsequent occupation of eastern China, including Manchuria. Despite the striking imbalance in military power and the many years of war, Japan was not able not to force the Chinese government to capitulate. This was because China, with its large population density and inhospitable geography was an impenetrable force. This exemplifies how the geography of China has shaped the country's internal policies. Furthermore, the indispensability of this impassable buffer terrain in its function to stop an invading army, has led to the Han government's iron fist rule over these regions. Moreover, this territory it gives China a greater power on the global stage because it is protected against invading forces. Unlike other nations they do not have to worry about the repercussion of controversial international actions because the resultant threat of invasion is less prevalent.

MAINTAINING CONTROL OF THE BUFFER REGIONS

The buffer regions, which are so integral to the defence of Han China, are often difficult to control due to the ethnic geographical differences between different groups such as the Muslim Uyghurs in Xinjiang and the governing Chinese Han. As a product the policies of the Han government are ones that assert hard-line control over these regions.

When the Han government are distracted or are in a weak position some of the buffer states start to drift out of the iron fist rule. This is exemplified, in the Sino-Japanese war. The war consumed all of Han China's military attention, and ergo some of the buffer regions (Tibet and Xingjian) started to move away from Han China rule. When Mao rose to power in 1948, he understood the indispensability of these buffer regions and began to work on policies to control them. Mao pursued policies to regain a firm control over the buffer regions and to remove the influence of the USSR over the regions of Mongolia and Manchuria. This successfully reinstated Chinese influence over the regions. Mao then mobilised the military to remove the warlord, Yang Zengxin, from Xinjiang enabling him to retake this region so indispensable for China. In 1950 Mao turned his attention to Tibet where he attained complete control in 1951. These polies can certainly be considered a result of the tensions caused the by the geographical ethnic makeup of China.

After the military action to re-assert Han dominance in the buffer regions, the Chinese government created a policy that gave these regions semi-autonomy. This allowed them to partly govern themselves but with China exerting significant levels of influence both economically and militarily. The buffer regions are resentful of the Han occupation and as a result independence movement in the area are prevalent. For instance, the East Turkestan Independence Movement. These self-proclaimed freedom fighters are deemed by the Han Government to be terrorist organisations, and thus are heavily persecuted especially in Xinjiang. In the deserts of Xinjiang, there have been 're-education' camps set up, reminiscent to concentration camps, for the detention of these ethnic minority Muslins Uyghurs. There has been consistent friction between the independence groups and the Han government for many years. The movements were formed as a result of varying ethnic groups in China's anthropological makeup therefore showing that China's geography has a major impact on internal policies, particularly ones that subdue these minorities.

China's natural geography has resulted in a vast drought problems due to their hot climate, their lack of water sources and their large population. With many areas of the country having unsustainable population density growth, it is projected that this drought problem will be amplified in the future. Beijing has had to conform its internal policy to try to achieve greater levels of water security as a result of their geographically disadvantageous position. As well as water issues, China is in a food crisis since it has approximately 20% of the world's population but only 7% of the arable land. If China's water supply were to decrease even slightly it would dramatically diminish the ratio of population to arable land and exacerbate an already problematic situation.

The region of Tibet is crucial to the Han government because it is where six of Asia's major rivers fount from. These include the Salween, the Yangtze and the Yellow River. These rivers ensure that Tibet is indispensable to China because if they lose control over Tibet, they lose control over a large proportion of China's water supply. Given that China's internal policies often revolve around water supplies, controlling Tibet is an important factor in its domestic considerations - losing Tibet and its groundwater would be cataclysmic for the whole of China.

Exasperating this issue is Tibet's proximity to India, China's main regional rival; if India can traverse the Himalayan mountains and occupy Tibet, then China's water supply would be at India's mercy. This could parch the majority of China's population and decimate its growing economy as no economy can work without water. Furthermore, One could hypothesise the disastrous effects for China if its course were to change. If India were able to control Tibet and all its rivers, China would be in serious trouble not just from drought but flood. An example of how the flooding of the Yangze is devastating is in 1954 where due to excessive rainfall there was a flood causing 30,000 deaths. A flood, as a result of the changing course in the river as a potential result of a Sino-India conflict could induce would devastate China. Chinese internal policy of suppressing the independence movement (and preventing Indian control) is as a direct result of this water insecurity.

The river Yangtze is the Nile of Southern China. As shown on the map below (fig 4) it spans the length of the country. It is crucial to sustaining life in Southern and Eastern China, not only for the all-important rice crop, the farmers working in the fertile land and those whose sole water supply is from the Yangtze but also for the thousands of people who rely on the aquafarming provided by this river. For example, crab farming in the Hongze Lake and fishing all along the river and its tributaries.[9]

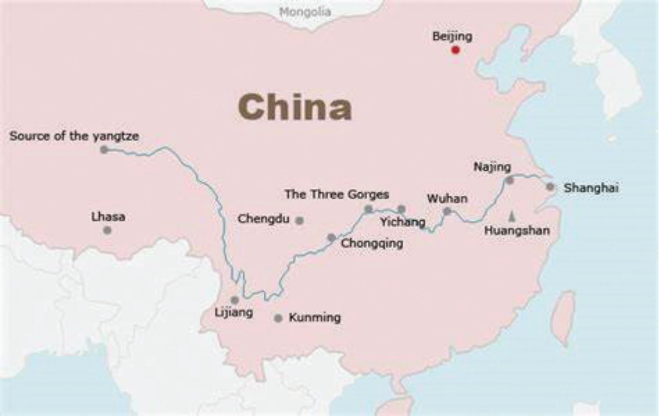


Figure 4: A map to show the route of the River Yangze sourcing in Tibet and ending in Shanghai[10]

Furthermore, the urban areas of southern and eastern China are also dependant on the Yangtze's water supply. Cities like Chongqing, Nanchang, Nanjing and Shanghai are utterly reliant on the river. The Yangtze River Delta Economic Zone accounts for 20% of China's GDP. This emphasises the importance of this river and by extension, the importance of policies that control it. This reliance can certainly account for the harsher political control of regions that the Yangze pass through, again illustrating the impact geography has on the internal policy of China.

GEOGRAPHICAL RESTRICTIONS ON TRADE BY LAND

Although the impassability of its borders prevents invading armies from entering China, it also prevents trade from exiting China. Traders are unable to easily traverse much of the border to sell Chinese goods. For example, as a result of the impassable Himalayan Mountain range on the south border of China, Chinese international influence is reduced in Southern Asia, Nepal, India, Bhutan and Pakistan. China's geographical wall does have some holes where trade can pass, one of the most accessible being the land bridge on its northern border with Kazakhstan where trade has passed through for centuries. This land bridge was the gate to the Eurasian world and resulted in the formation of the Silk Road which ran from Han China through Xinjiang and Kazakhstan on its way west. Yet, the border with Kazakhstan is nearly a thousand miles away from the Han provinces so although there is a gap in the natural border, the traders must cross a thousand miles of inhospitable terrain to get there. So, policies needed to be developed in order to combat this issue. An early example of this is the creation of the Silk Road which utilised the land bridge between Xingjian and Kazakhstan as shown in the map below, the red line on fig 5 representing the Silk Road going through this land bridge.

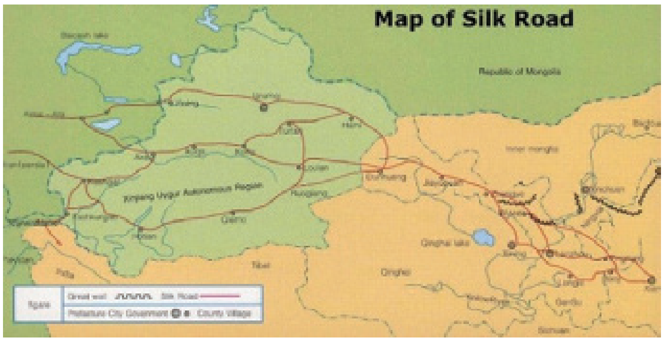


Figure 5: A map to show the route of the Silk Road passing through Xinjian and Crossing into Kazakhstan[11]

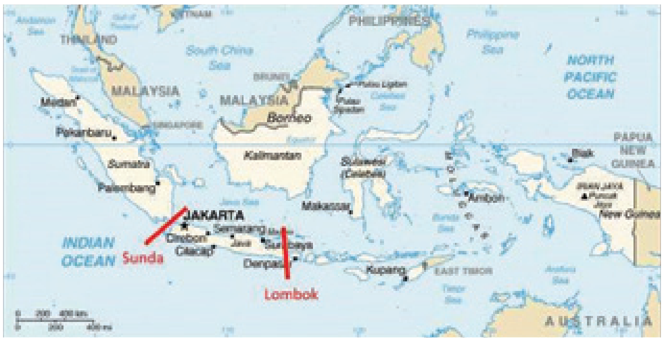


Figure 6: A map to show the straits between the South China Sea and the Indian Ocean. The Sunda, Lombok and Malacca Straits[12]

The development of the Silk Road was one of the first foreign policies that China implemented as a result of its geography; this road was created and improved because there was a need for a passable road for merchants to use. This has not only shaped Chinese foreign policy historically but also in the present. China has developed a $900 billion project, 'The Belt and Road Initiative' which is the largest ever infrastructure project. It is an ambitious development campaign that aims to boost trade and international co-operation with the heart of this co-operation being Beijing. The land part of this projects aims to build the Silk Road of the 21st Century with bridges being rapidly constructed in Pakistan, Bangladesh and many more. The reason for this project is far from an altruistic 'global development' policy. The initiative is more driven by Chinese foreign policy to increase trade along routes which China can access and control. As a result, China has a greater ability to trade by land therefore China's economic, and therefore political, influence will increase dramatically. This major foreign policy for China is all a result of the geographical restrictions forced upon China by its borders.

GEOGRAPHICAL RESTRICTIONS ON TRADE BY SEA

China's geography has not only imposed restrictions on China's land trade routes but also on its sea links. China's only access to the sea is its eastern border, which severely limits the proportion of this vast country which is open and available to sea trade. China, currently, does not have easy access to Eurasia via the sea; it only has limited access to the Pacific and India oceans.

China is an export-led economy, with 37.8%[19] of its GDP made up of exports, meaning that trade is crucial to the Chinese economy. Without trade, Chinese industry and its economy would fail, and its GDP growth would plummet. This, combined with the fact that approximately 80% of global trade is by sea, illustrates the imperative need to develop foreign policy to maintaining and increase access to these oceans. This foreign policy development is a result of the geographical position of China. A thriving Chinese economy is the only thing that holds the government in place. In China, there is a trade-off between political freedom and economic prosperity - the people of China will accept a lack of freedom including restricted freedom of speech for high levels of economic growth and low unemployment. If the Chinese economy starts to decay because of declining exports the people of China could revolt against the oppression which might lead to the ousting of the current ruling party.

China's geography, with its impassable borders and limited sea access, has a direct effect on its global trade security and in order to rectify this the Chinese government has employed quite drastic and controversial foreign policy measures such as the South China Sea territorial expansion which will be expanded on below.

To access the Pacific Ocean, Chinese ships must pass through the East China Sea, where there are many straits owned by other nations. For example, the Ryukyu islands off the south coast, owned by Japan creates a wall of straits. Japan could easily impede Chinese trade routes via these straits. Given their limitations to safeguarding these trade routes, the most accessible way for China to enter the Pacific Ocean would be via the Luzon Strait between Taiwan and the Philippian Island of Luzon. This strait is approximately 200 miles wide, so would be difficult but not impossible to patrol, so even this route to the Pacific Ocean is not entirely secure. Whilst not completely protected, it is the most accessible trade route available to China. That being said, the only significant markets China can access via the Pacific are the United States, Canada, Mexico and Brazil, which only makes up 25.2% of China's exports.[13] Taking this one step further, the Pacific Ocean only allows China to supply the west coast of these countries restricting access to the complete market. So possibly an even lower percentage of China's exports go via the Pacific Ocean, illustrating, the need to develop successful policies to access their more significant markets via the India Ocean.

The Belt and Road Initiative, a Chinese economic foreign policy, recognises that the markets accessible via the Indian ocean has a greater importance which led to the creation of the "Maritime Silk Road Initiative". The routes accessible to Chinese shipping to access the Indian ocean, and therefore most of China's significant export partners (Europe, Africa, the Middle East and the East Coast of America) are much more limited and blockadeable. Significantly, there are only a few viable routes for Chinese exports to take to access the western markets. This would involve going via the Malacca, Sunda or Lombok Straits. The preferred course for Chinese ships would be through the Malacca Strait, a conglomerate of islands where the strait at its narrowest is no more than 1.5 miles wide meaning that is is easily blockable. This would be a problem for China as it prevents it from trading with some of its bigger partners. Furthermore, the Malacca Strait is crucial to China with 25% of world trade going through this strait a high proportion of which is Chinese. For this reason, China has embarked on foreign policies that attempt to protect this route.

Should the Malacca Straits were to be blocked, China could use the Sunda Strait; however, at only 15m deep, it is inadequate as a shipping route. Alternatively, the Lombok Strait is 20km wide, and

250 m deep, and this is better proportioned to handle large ships.[14] However, if one was trying to access the Suez Canal which is the route taken by Chinese ships when accessing Europe, then going through the Lombok Strait would add approximately 3200km to the journey adding over 143 hours to the passage of a boat traveling at 12 knots which could cost up to $1,191,000 extra for each ship. Over 94,000 ships use this trade route annually and if we conjecture that over half of these are Chinese owned, then the cost to China would be around $55,977,000,000 (56 billion) which would be a massive loss to the Chinese economy. [15] This again illustrates how geography has created a problem which requires the application of successful international economic and political policies to solve.

THE ECONOMIC WEALTH UNDER THE SOUTH CHINA SEA

Trade is not the only geographical factor that has influenced Chinese foreign policy in the region. Natural resources have played a massive role in shaping Chinese policy in the South China Sea. There is a huge wealth of natural resources under the South China Sea; 11 billion barrels of oil waiting to be tapped for example. This supply would last the United States just over one and a half years if they were solely reliant on it. Moreover, it has a huge economic value. Using the current price of oil at $52.28 per barrel the economic value of this oil is five hundred and eighty-six billion US dollars. [16] China consumes the second highest amount of oil in the world, and currently are heavily reliant on foreign imports. This causes issues for the Han government. If China was able to secure a large source of oil it would gain greater economic security. On top of its economic value, it also accentuates Chinese desire to control the South China Sea – one that is reflected in its Foreign Policy.

There is also a massive reserve of natural gas under the South China Sea. There is approximately 190 trillion cubic feet, priced at over $760 billion. This makes the sea a hugely valuable economic asset to control. Consequently, Chinese presence on the global stage and its relationships with other nations in the regions are shaped around controlling it and the resources it endows.

One of China's policies to mitigate their geographical problems is to dominate the region by creating artificial islands. The 1982 United Nations Convention on the Law of the Sea states that 200 miles from any land mass under the sovereign control of a nation is the exclusive economic zone (EEZ) of that nation including the natural resources within. By creating artificial land mass, China claims a larger area of sovereign EEZ around the South China Sea and thus controlling the oil and gas reserves in those areas. China has been building military bases on reclaimed shoals and sandbars to reinforce its territorial claims in the South China Sea. This completely opposes US efforts to prevent Chinese dominance in the region.

China's foreign policy, claiming the natural islands in the South China Sea, for example, the Spratley and Parcel Islands, is detailed in the 'Position Paper of the government of the People's Republic of China on the matter of Jurisdiction in the South China Sea."[17] The Philippines submitted a notification and statement of claim to initiate arbitration proceedings under article 287 and Annex VII of the United Nations Convention. The Philippines stated that China had overstepped the convention when it claimed/invaded these islands. In response, China created a policy of disregarding these UN rulings and continues to occupy these natural islands. The UN proved ineffective in reinforcing the regulations over China. In fact, Chinese foreign policy was specifically designed to take advantage of the rich geography of the South China Sea with complete disregard to the UN rulings. This is a prime example of how China's relationship with the rest of the world has suffered as a result of its geographical desires.

Furthermore, it is not only the Chinese and the Philippians that are staking claims to these islands; nations such as Vietnam, Taiwan, Bruni and Malaysia are also demanding parts of these archipelagos. Vietnam states that these islands fall under Hanoi's control according to the UN International Convention on the Sea, however, the Chinese government has repeatedly blocked Hanoi's attempts to look for oil inside what both countries regard as their territory.

In an attempt to control the South China Sea, China is seeking to enforce an air-identification fly zone inside its 9-dash line, see map in fig 7 below. They insist that all planes identify themselves to the Chinese Aviation Authority and state that no foreign military aircraft should fly through the specified area. This air-identification zone is a vital part of Chinese foreign policy in the region of the South China Sea, and it is clear to see that the geography of the area has induced this initiative. However, many other countries, including those in the EU, Japan and the USA have dismissed this and are continuing to fly through the zone. The consequence of this is increased tensions between China and the rest of the world on the Global Stage. Tensions were highlighted in July 2017 when the Foreign Secretary Boris Johnson committed two aircraft carriers to carry out Freedom of Navigation 'training' exercises in the waters of the South China Sea. This is considered as an open message to Beijing that western countries will not tolerate this Chinese policy.

China is not only trying to control the land and air within this Nine-Dash-Line they are also trying to control the seas. Tim Marshall in Prisoners of Geography wrote: ''in October 2006, a US naval supercarrier group led by the 1,000 foot USS Kitty Hawk was confidently sailing

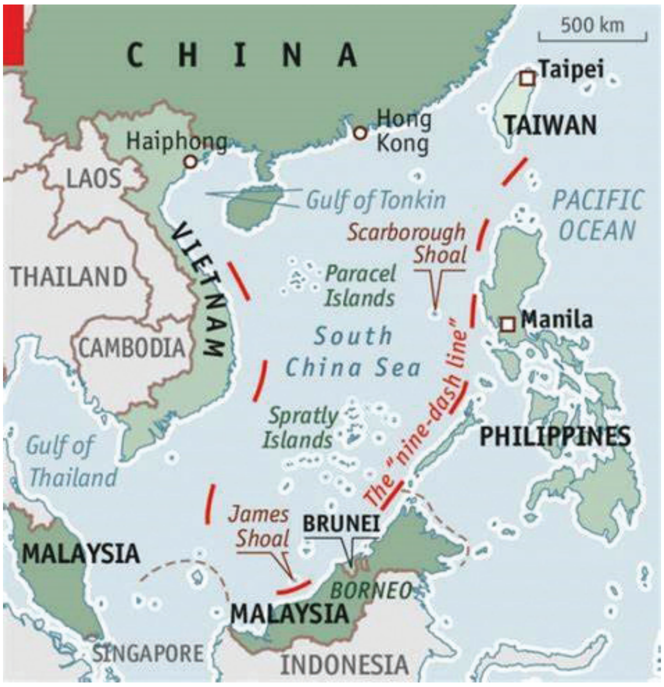


Figure 7: A map illustrating the 'Nine-Dash-Line' and therefore Chinese territorial claims.[18]

In the East China Sea between Japan and Taiwan, minding everyone's business when, without warning, a Chinese navy submarine surfaced in the middle of the group''. [18] He is validated in this view because this was an extremely provocative move by China who deliberately attempted to scare the western countries to force them out of their sphere of influence.

Due to the increased presence on western capitalist countries, especially the United States of America in the area surrounding the South China Sea (Thailand, Taiwan, the Philippines, Japan, South Korea and many others), China has becoming increasingly hostile towards foreign nations. Souring of an already very bitter relationship between China and the rest of the world. The geography of the South China Sea is certainly the most significant factor in the worsening of relations between China and the rest of the world on the Global Stage.

China's geography has been the mould for its foreign policy for many years and will continue to be so for the foreseeable future. An area where China will continue to expand its influence is Africa. China is investing heavily in the continent; building infrastructure and educating the population. There are many reasons for this, one of which is the attraction of the abundant natural resources of Africa - Diamonds, Gold, Iron Cobalt and many more. This need for African resources is perhaps a result of Chinese geography, it has a lack of natural resources in its country. China also has a foreign policy of seeking accessible naval ports around the world so it can secure its threatened shipping routes. China achieves this by giving developing African countries loans to enable them to build ports and bases. However, this loan comes with a catch, if a country fails to repay the loan, China will annex this base strengthening their position in Africa. Chinese infrastructure investment in Africa is of concern for western countries because many believe that that Chinese soft, economic power will develop into hard military power. China will use a natural disaster where Chinese workers or investments have been affected as an excuse to send in their military in the 'interest of the Chinese workers' whereas in reality it is a rouse to control the resources. The Sikyong of Tibet, Lobsang Sangay, stated during a recent speech in the Palace of Westminster that there are 'worrying similarities that can be drawn between the soft power investment in Africa and Tibet'. When China first took over control of his country, they invested heavily in the region; building roads, bridges and railways. The Tibetan people were delighted with this economic injection and resulting prosperity, they even created songs and poems praising the Communist Party of China. However, after this soft power investment in Tibet, along the roads built by Tibetan workers came the guns, tanks and military to crush any opposition. In Africa, the process has only just begun, the roads are starting to be built, and the question on everyone's lips is: what will be coming along these roads, will it be trade, or will it be guns?

CONCLUSION

In conclusion, China's geography has a profound impact on the internal policies of the Han Government. The oppression of minorities and independence movements in Xinjiang is due to its crucial, impassable, terrain protecting the Chinese Economic heartland of China Proper. Similarly, the oppression of independence movements in Tibet results from China's desire to protect its vital water table, crucial to sustain an already parched China. Geography has also been a significant factor in shaping China's presence on the global stage with issues such as the restrictive geography of the trade routes to access the Indian Ocean contrasting with the favourable geography of the natural resources in the South China Sea. Although these issues are conflicting, they have the same effect on Chinese foreign policy - to assert its dominance and control, both for trade security and natural resources. The geographical restrictions on China's trade has forced the Chinese government to pursue innovative policies to circumvent these barriers. Increasingly China has turned to the air as a mode of transportation. The sustainability of airborne transportation is questionable as it puts China at an economic disadvantage. Looking forward, in order to maintain their economic competitiveness, China will have little option but to direct their trade policy increasingly towards rail, road and sea.

Why do Greenland sharks live for 400 years?

James Miller

AN INSIGHT INTO THE COMPARATIVE BIOLOGY OF AGING MECHANISMS AND THE EVOLUTION BEHIND THEM

In December 2014 a group of scientists conducting research in the North Atlantic made a surprising discovery. Using radiocarbon dating, they were able to accurately estimate the ages of several individual Greenland Sharks, and they found that the longest-lived of those was 392 years old (1) (give or take 120 years)

This made Greenland sharks the longest-lived vertebrate species on record, far outdoing its nearest competitor, the bowhead whale, at a mere 211 years.

In this essay I will attempt to outline some of the various biological factors that seem to correlate to lifespan in different species, and examine the mechanisms that Greenland Sharks might use to prolong their lifespan so substantially above all other vertebrates. Seeing as very little research has been performed on this species (they are quite hard to study) the conclusions that I draw from this are mostly just thoughts based on my understanding of the biology of aging. This field, called gerontology, is particularly hazy, with very many theories and relatively little proven fact.

Before I begin to explain my findings, it would be worth specifying a formal scientific definition of aging; it's not as simple as one might expect and until we are entirely sure of what aging is, we can't define when species are evading it. The accepted definition for scientific purposes is *'a progressive deterioration of physiological function, an intrinsic age-related process of loss of viability and increase in vulnerability'* (2). Some descriptions also include a decline in reproductive capacity.

BODY MASS

One factor that correlates very distinctly with maximum lifespan across several taxa is body size. From the figure below, there is a general trend of tmax= $5.58M^{0.146}$ $r^2 = 0.340$, suggesting body mass accounts for 58% of the variation in longevity in the 1,701 species recorded. This trend is called the 'allometry of lifespan'.

Greenland Sharks are one of the largest species of shark, potentially growing up to 7.3m and 1,400 kg in weight (3), so I assumed that whatever factor is related to the body mass to create this trend is very likely to be contribute to the Greenland Shark's longevity.

What could be the biological mechanism that produces this relationship?

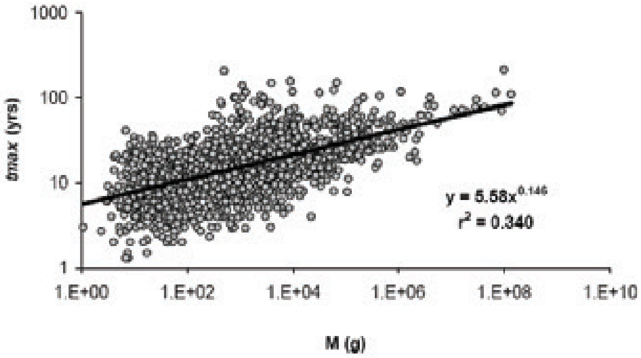


Figure 1 – Correlation between maximum lifespan (tmax) and typical adult body mass (M) using all species (n = 1,701) present in AnAge build 8. Plotted on a logarithmic scale. Comprises almost entirely of Chordata species, including endotherms and ectotherms. hiip://www.senescence.info/comparative_biology.html

METABOLIC RATE

Metabolic rate is, as a general rule, proportional to body mass - Max Kleiber's law concluded that basal metabolic rate (BMR) could be accurately predicted by raising the body mass to the power of ¾. Seeing as metabolic rate is a popular theoretical cause of aging, and Greenland Sharks are thought to have an extremely low metabolism (having only one heartbeat every ten seconds as a rough indication) I thought it would be worth investigating.

In the early part of the last century, Max Rubner discovered a correlation between metabolic rate of a species and its longevity. It seemed that the faster the metabolism of a species, in general, the shorter its lifespan.

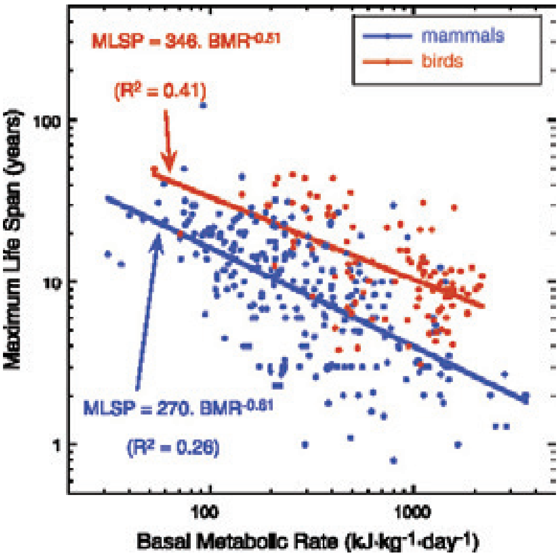


Figure 2 – A graph showing the relationship between the maximum lifespan of birds and mammals and their metabolic rates. hiips://www.researchgate.net/figure/An-examination-of-the-rate-of-living-theory-for-mammals-and-birds-The-relationship_fig2_5915827

Several others have made the same observations, and, in 1928, they were developed into the 'Rate of Living Theory' by Raymond Pearl. In this he postulated that '...In general, the duration of life varies inversely as the rate of energy expenditure during its continuance. In short, the length of life depends inversely on the rate of living.' (4)

The theory is quite intuitive; just as longer-lived animals take longer to reach sexual maturity, aging depends on the rate at which life is lived. It also seems to hold up very well with a large amount of empirical evidence showing a very clear trend (see the above figure). However, no actual mechanism for the theory was suggested until the 1950s, when the free radical theory of aging was proposed by Denham Harman (5).

Free radicals (chemical species with unpaired electrons in their valence shell) and oxidants are referred to as reactive oxygen species (ROS), and are very highly reactive. This gives them the capacity to damage biological molecules in the body. They can be produced from interactions with external sources such as UV radiation, but they are also a by-product of normal cellular metabolic processes. The theory dictates that higher metabolisms mean a greater rate of these biochemical reactions taking place in the body, such as cell respiration, and therefore these ROS are produced at a greater rate. The damage that arises from these molecules' interactions with the body will accumulate faster, and produce the symptoms of aging earlier.

Since the theory was first proposed numerous studies have been conducted to support it. It has been demonstrated that ROS and oxidative damage builds up with age (6), and lowering ROS levels extends the lifespan of many model organisms such as fruit flies and mice (7). Some specific visual marks of age, such as wrinkles, are known to be a direct result of this type of damage (8).

However, in more recent past, this theory has been – humourously - declared 'dead'. Many studies have now been undertaken which apparently contradict it. While in some model organisms antioxidant proteins extend lifespan, in others their overexpression seemed to have no significant effect, and in a few cases, decreased lifespan. Similarly, caloric restriction, which has proven effective in some organisms, has been completely ineffective in others (9).

In order to understand why this might be, we must remember that oxidative damage is only one form of damage. Orgel had an idea that there are errors in transcription, translation, and replication, which similarly accumulate over time, and will cause increasing imperfection in protein function or cancers. Every kind of metabolic reaction has imperfections, and over time the side effects of these errors takes its toll on the body (9).

While the rate of these biochemical processes in an organism are, by definition, linked to metabolism, the basal metabolic rate isn't directly proportional to the rate of every chemical reaction taking place in the body – depending on conditions e.g. food ingested, some reactions will be taking place more frequently than others. Some of these damage types may be more destructive than others, or may not have such effective protective mechanisms in place to combat them (many will not be subject to natural selection – see more later in evolution section). Thus in some species oxidative damage may be more relevant as a lifespan-determining factor than in others. For many organisms, aging still occurs in anaerobic conditions where there is little ROS (5).

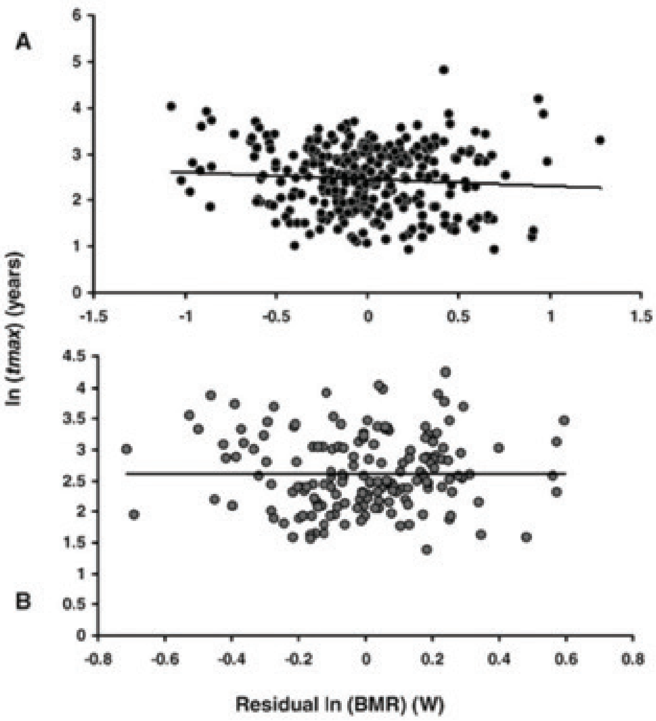


*Figure 3 - Ln-transformed relationship between basal metabolic rate (BMR) residuals and maximum longevity (t_{max}) in mammals (**A**; n =300) and birds (**B**; n = 167). Closed circles: individual mammalian species (A); grey circles: individual bird species (B). hiips://www.ncbi.nlm.nih.gov/pmc/articles/PMC2288695/figure/F2/*

Caloric restriction will limit specific metabolic pathways such as glycolysis, and fatty acid/amino acid metabolism (5), limiting specific forms of damage associated with these pathways, which will have varying effects depending on the defence mechanisms of the species in question.

So, as scientist Steve Austad says of aging, '[metabolism] is thought to be a player. But it's not driving the show.' (10)

And indeed, reanalysis of the relationship between Basal Metabolic Rate and Max Lifespan show no significant correlation when corrected for body mass and phylogeny (11). (Although it is worth noting that resting metabolic rate isn't necessarily a good measure of total metabolism, nor is maximum lifespan a good measure of longevity).

So what therefore explains the clear correlation between body size and tmax? Well, as of yet, there is no evidence of an unknown physiological factor controlling aging in a manner directly proportional to body mass (12). At present, leading scientists consider it more likely that the allometry of lifespan is due to other biological mechanisms discussed later in this essay, which, although they aren't directly related to body mass, have more of an effect in smaller organisms due to evolution. I will look at this later in more detail, but in short, they have higher extrinsic mortality rates due to predation/accidents etc, and therefore aging is less selected against in smaller animals than in larger ones. Thus many of the following potential 'programmed' mechanisms of aging to be discussed in this essay act more substantially in smaller species than in larger ones, and give this correlation.

This hypothesis is supported by the fact that the particular exceptions to the rule, birds and bats, which live far longer than their size and metabolic rates should dictate, are both capable of flight, and have a greater capability to evade predation. Similarly smaller individuals within larger species often live longer in captivity (12) (I will explore this intriguing statistic briefly in the next section of this essay) which does serve as evidence that there may be no direct link between body mass and longevity: correlation without causation.

But there are clearly other factors at work besides those that arise simply as a result of body mass. Even using the maximum hypothesized mass of the Greenland Shark (1400kg) according to the earlier trend of tmax $=5.58M^{0.146}$, its maximum lifespan should be around 44 years. There must be some special adaptations involved in Greenland Sharks that differentiate their lifespan from other animals of a similar size.

DEVELOPMENTAL PERIOD

Although in our day-to-day lives we perhaps consider it all to be the same affair, programmed maturing to adulthood could very feasibly be an entirely different procedure to the gradual senescence we experience beyond sexual maturity. However, in many taxa, time to reach maturity was correlates strongly with maximum lifespan (13), suggesting that perhaps there is a common mechanism. This is no exception with the Greenland Shark, which is thought to take 150 years to reach sexual maturity and grow about 1cm per year (1).

In several model organisms, mutations in certain genes extended both the period of development and the longevity of the organism, suggesting the existence of a general physiological clock for the organism. This has been shown in nematode worms (14), where mutations to clk-1, clk-2, clk-3, and gro-1 genes interact to allow the individuals to live almost five times as long as wild worms, and several other invertebrate models.

However the precise, concrete mechanisms for how development and aging are related still remain unclear. We know that hormones are integral in the developmental process, and it is thus theorised that the endocrine system has a part to play in regulating aging as well.

We also know that the endocrine system in humans changes with age, but whether that drives most other symptoms of aging or not is a matter up for debate.

A particularly popular hormone-based theory of aging is that the brain, which regulates several endocrine changes (e.g. growth hormone production), acts as a centralised pacemaker for aging via this hormonal control – this is known as the neuroendocrine theory of aging (13).

Many studies with different model organisms have been performed in this area with – as seems to be a bit of a theme with gerontology – varying conclusions.

Several experiments associated the insulin/insulin-like pathway with aging. The fact mentioned earlier - that smaller individuals within species such as mice, rats, horses and dogs seem to live longer – could be related to the lower levels of insulin-like growth factor (13). Dwarf mice homozygous for a certain gene – *Pit l* – have lower growth hormone and Insulin-like growth factor levels and live about 40% longer (15).

Similarly, a mutation of the klotho gene, which acts as a circulating hormone, seems to accelerate the aging process, while overexpression of klotho extends lifespan by approximately 30% (16). This gene could be related to insulin/IGF-1 signalling.

Interestingly, caloric restriction appears to induce hormonal alterations – in rodents, decreasing plasma levels of insulin and IGF-1 (17) - so this is potentially an alternative explanation for the effects of caloric restriction beyond the free radical theory of aging.

It has been suggested that the link between time to reach sexual maturity and average lifespan could, again, not be a causal relationship but only correlation. The developmental period could be reflective of the life history of the species and its reproductive strategy (see in the evolution section), which in turn may favour slower or faster rates of aging. As with the correlation to biomass, the developmental period could simply correlate to longevity because of a common link to evolutionary tactics.

IMMUNE SYSTEM AND DISEASE

Disease, although not a cause of aging, does become a more potent external cause of mortality with age, and having a robust immune system is certainly necessary for Greenland Sharks to reach 400 years old, so I thought it worth investigating. In addition, as I will explain later, reducing mortality rates from external factors such as disease can lead indirectly to the evolution of a longer maximum lifespan.

For humans, heart disease is the number one cause of death worldwide (18), with the risk increasing dramatically as individuals reach older age.

However, this seems to be less of a problem for Greenland Sharks. Why might this be?

We know they have an extremely low heart rate: about one beat every ten seconds (19). But scientists at the University of Manchester and Copenhagen are testing various analytical techniques on shark hearts (ethically obtained from unfortunate bycatch) such as mass spectrometry to identify any molecules that may protect the cardiovascular tissue. At present they have found out very little (19).

The same scientists are also analysing bioaccumulation of various organic toxins, which can be a problem for oceanic predators, especially those at high trophic levels. Their findings appear to suggest that this isn't an issue for Greenland Sharks – the toxins apparently don't accumulate with age (19).

However, increasing concentrations of microplastics in the ocean could act as both a source and a vector of toxins, and pose a new serious threat to the shark.

It is equally surprising that cancer isn't more prevalent in these individuals, and killed them off before they can reach such significant ages. In fact, cancer incidences in sharks are so rare that a scientist has published a book with the misleading (or rather, blatantly incorrect) title, 'Sharks Don't Get Cancer'.

Shark cartilage has long been thought to have cancer-resistant properties, and has been marketed as a cancer cure (especially in traditional Chinese medicine). This is unscientific and damaging to conservation efforts. But, although it doesn't work on humans, there is some truth behind it.

Cartilage is avascular, and vascular tissue is necessary for the formation of larger tumours as a blood supply is needed for all of the new cells. Tumours have the ability to get around this problem by angiogenesis, but sharks have evolved defence systems against this too. Relative to a calf, it is thought that sharks have 100,000 times more angiogenesis inhibitory activity on a per animal basis (20) (vastly disproportional even considering the size difference).

Some cartilage components are also thought to be antineoplastic and have, in studies, extended the life of leukemic mice (20).

TELOMERE SHORTENING

Seeing as telomeres seem to be in the forefront of aging research at the moment, I thought that it might be worth investigating them as a potential factor in Greenland Shark lifespan. As far as I can ascertain, nobody has measured the telomeres of Greenland Sharks, but nevertheless it is entirely possible that telomeres play a part in their longevity.

Each time DNA is replicated in a cell, the duplication doesn't occur all the way to the end of the chromosomes. So after each successive replication, some genetic code is lost from the end. The telomeres are sections of repeated base code found at the ends of chromosomes – AGGGTT in vertebrates – which act as a disposable buffer and are lost in place of coding DNA.

As a section of this telomere is lost after each replication, telomeres gradually shorten over time. Once the telomere has been 'worn out' after too many replications (known as the Hayflick Limit – about 60-70 divisions in humans), the cell becomes 'replicatively senescent' and no longer duplicates, in case it should cause harm by missing out vital genes (e.g. causing a cancer). This prevents tissue from being able to repair effectively as gradually more and more of the tissue cells become unable to undergo mitosis over time.

In this way, telomere length controls the rates of cell senescence, and longer telomeres mean the cells can continue to divide and repair body tissue for longer. Several studies have shown significant correlations between telomere length and mortality rates. For example, Cawthon et al in 2003 studied 143 individuals aged 60 and older, and found that individuals with the shortest telomeres for their chronological age had higher rates of mortality (21). It appeared that for those older than 74, there was no significant correlation, however this could be due to the 'survivor effect'. If those individuals with shorter telomeres are more likely to die earlier, then they will not make it into the older age bracket, and thus the over-74s sample group will display less variation in the length of telomeres, subsequently telomere length may differentiate the mortality risk less significantly between the individuals.

This study was of a relatively small sample size, and other studies have thrown up results that both contradict and support the theory. It appears now to be accepted as scientific fact, and is a specific mechanism to which the bat's disproportionately long lifespan in relation to its size has been attributed (22).

Specifically relating to the subject of this essay, a correlation has been shown between fish with slower rates of telomere attrition (and indeed those that elongate telomeres with age –see below) and longevity (23).

As well as having a long telomere to wear down, some cells have the ability to regenerate it. Stem cells reach replicative senescence very slowly indeed – this is due to the action of an enzyme called telomerase, which is a ribonucleoprotein that replenishes the telomere after replication. Consequently, they take very long periods of time to reach the Hayflick Limit and are capable of long term self-renewal.

High telomerase expression has been noted in one of the other longest-lived elasmobranches – The Spiny Dogfish (24) – so I believe that, within all probability, mitigating telomere shortening plays an integral part in slowing the Greenland Shark's aging process.

THE EVOLUTION OF AGING

Taking aging to be at least partly programmed, it seems at first counter-intuitive that aging could have evolved. How could increasing chances of mortality and decreasing reproductive capacity with age be of any evolutionary advantage to an individual?

There have been several hypothesise suggested to account for it.

Initially, in 1891, August Weismann proposed that programmed aging evolved to the advantage of the species rather than the individual, by replacing 'worn out' individuals with younger, fitter ones. This is a form of group selection.

He himself later dropped this theory in favour of a new one, that organisms with somatic and germ cells must allocate resources to the germ cells in order to reproduce, and that detracts from the maintenance of the somatic cells, resulting in aging.

Later a more sophisticated theory was developed by Peter Medwar in 1952. He observed that the force of natural selection decreases with age, as all organisms eventually die of predation/disease/accidents etc. Alleles that are beneficial in early life are thus more important than those that benefit the organism in later life, and are favoured by natural selection – even if at older age they may be of disadvantage to the organism. In contrast, those alleles with negative impacts at extreme ages have very little impact on organisms because very few individuals live to that age, and they will most likely have already reproduced and passed on those alleles by the time they are that old. In this way the genes that cause aging, although they have no evolutionary advantage, are not weeded out of populations by natural selection. Huntington's disease is an extreme example of this phenomenon.

This is the theory that I personally find most convincing, but it still leaves a question unanswered. How come some organisms have evolved to live for hundreds of years, while there exists a mayfly species that can live only 5 minutes in its adult form? It is all to do with the 'life history' of the organism. Some organisms are semelparous, and

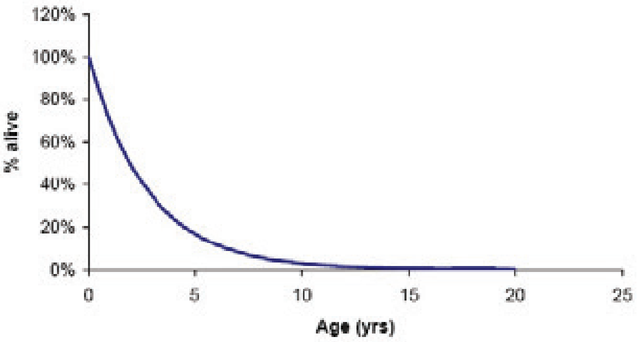


Figure 4 - Survival curve showing the percentage of organisms alive at a given age for a hypothetical population assuming a constant mortality rate across the entire lifespan–i.e., no aging hiip://www.senescence.info/evolution_of_aging.html

die after reproducing only once, while others are iteroparous, and reproduce multiple times during their lifetime. Organisms which adopt the former method usually generate huge numbers of offspring in the single spawning event, and provide little parental care, while in the latter, fewer offspring are produced in each reproductive event, each receiving more attention from the parents. Theoretical models imply that higher chances of mortality in a species are likely to lead to them evolving a semelparous reproductive strategy, as there is less point investing resources in a future reproductive event that is unlikely to occur. This is coupled with the fact that species with lower mortality rates from external factors, as dictated by Medwar's theory above, are more likely to live longer, as that allows natural selection to act more powerfully against negative genes that come into effect at later age. This is why, in general, creatures with higher chances of predation live short periods of time even in captivity, and why, for instance, the evolution of a sophisticated immune system in the Greenland Shark may indirectly lead to the evolution of better anti-aging mechanisms.

In addition, Medwar's theory seemed to me to be a little circular – claiming that aging resulted from genes that act in late life, which surely must have been activated by some physical symptom of later life- aging?.

There are two explanations that I could think of - the late-acting genes that it mentions must either: start their work early in life and have their damage accumulate over time, or, be activated epigenetically later by some somatic factor – in which case there must be some central 'physiological clock' that would cause these genes to be activated.

SUMMARY

In summary, the extraordinary biological feat of surviving to almost 400 years old could be down to any combination of the multitude of proposed factors that delay aging. Personally I find some of these more convincing than others, especially in relation to the Greenland Shark.

While at first the extremely low metabolic rate of the Greenland Shark seemed to me to be an obvious and common-sense explanation for its longevity, the articles I have read lead me to suspect that free radicals and ROS play a relatively small role, and overall metabolism is a significant but by no means the sole factor at work. This viewpoint is shared by scientists studying them (25) The cold temperatures could potentially have other life-extending impacts, activating various anti-aging genes that remain inactive in other species (25).

While, to my knowledge, the telomeres of Greenland Sharks haven't yet been investigated, it would not surprise me, given the results that research has thrown up in other aquatic vertebrates and in the closely related Spiny Dogfish, if it were discovered that the species had some method of delaying cells from reaching senescence, whether through having very long telomeres, expressing high levels of telomerase, or having low rates of telomere attrition.

The correlation between the extremely long developmental period of the Greenland Shark and its longevity also seemed impossible to be coincidence, especially given the strong correlation in other species. Whether this link is direct and regulated by the endocrine system, or indirect and as a product of its reproductive strategies and life history, I am not certain.

Finally, I believe that the resistance to disease and cancers that the shark family displays is certainly essential in reducing extrinsic causes of mortality and allowing individuals to live for such a long time, both directly and indirectly through allowing stronger natural selection against aging genes.

However, these evolutionary tactics could now no longer act in the Greenland Shark's favour. Human influence is causing higher death rates in this species than would be natural, through overfishing, pollution, and potentially climate change. The long developmental period of the shark makes it hard for the species to recover: they are only just starting to increase from overfishing before the Second World War, because any individuals born after that period still have more than 80 years before adulthood.

The delusion of free will and its effect on our perception of culpability

Charlie Buckingham

WHAT IS FREE WILL?

Before attempting to explain why we do not have free will and hence free will is a delusion, I feel it is necessary to define free will, given how diverse interpretations of it have been in history. Free Will is the power one has to think and act unimpeded, to choose between different courses of actions without the constraint of necessity. It is confined to the conscious part of our brain, as it is impossible to make conscious decisions in parts of the brain that simply function like any other organs. We cannot make decisions in the unconscious part of our brain, in the same way that we cannot control when our liver secretes bile and so we cannot claim to have free will in parts of our brain that we are not conscious of.

Upon bringing up the question of whether we have free will, many would take a passive stance and say one of two things. The first is that we don't have the capacity to discuss free will due to an absence of sufficient scientific evidence or understanding about the brain and how it functions in regards to consciousness. The second stance is that free will is immaterial as, whether we do or don't possess free will, it doesn't actually affect how we live our lives and therefore it is consigned to being a purely academic argument. I disagree with the assertion that either of these is true. There is plenty of scientific evidence in the field of neurology, and the argument over free will's existence can change the way we look at the world, particularly in regards to justice and culpability.

In order for us to have free will, two assumptions have to be satisfied. Firstly, our consciousness is the sole source of our actions and secondly, we are free to act differently to how we actually do in any scenario. By this I mean that, for example, we feel as though we want to move our arm and then we do. Our feeling was the sole source of the action, that feeling was something we were conscious of and we could have chosen not to move our arm. I will deal with these two assumptions separately from now on and will discuss how both are false individually.

DO WE HAVE FREE WILL?

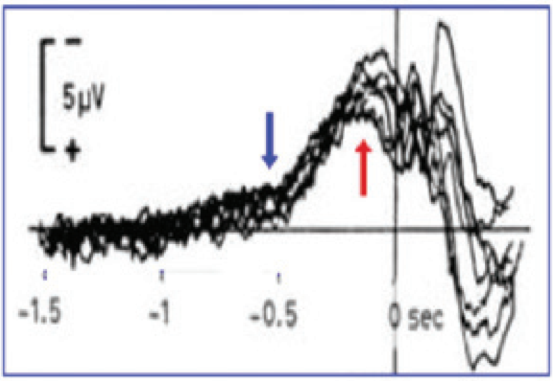
Let us start with the assumption that our consciousness is the sole source of our actions. There is simply no evidence for such an assumption. Consider the question: 'What are you going to think next?' You could not possibly answer; there is an impossibility of knowledge. Think of the statement: 'You do not know what you are going to think next'. Its negation is: 'You do know what you are going to think next'. In order for the negation to be true, you would need to be able to answer the question above and accurately, as in order to choose what to think next, you would need to know your options of what you can think next are. This would mean you would have to think about something, before you actually thought something, which is a contradiction. Therefore, the negation cannot be true and the statement: 'You do not know what you are going to think next' must be true.

Therefore, the earliest that you can know that you are going to think something and when you think it are simultaneous. The thought just emerges into consciousness and you learn of its existence yet you believe that you were the one that chose to think it. For example, if you are confused, you did not choose to become confused. At some point, your brain couldn't keep up with what it was receiving from your sensory neurones and then you became consciously aware of the fact that you were confused. Equally, you cannot choose the opposite, understanding something, as otherwise no one would ever become confused. Hence, we are constantly aware of changes in our thoughts and mood but we are completely unaware of the neurophysiological thought process behind it.

Furthermore, the neurophysiological thought processes that our brain carries out preceding a decision can be detected in science and so the exact moment that our brain has come to a decision can be seen. This has allowed neuroscientists to check to see if there is a time lag between when our brain makes a decision and when we are conscious of making a decision.

In 1964 two scientists, Hans Helmut Kornhuber and Lüder Deecke, investigated the activity of the brain before an action using an electrocephalogram (EEG) and found that about 550ms before an action, an increase in electrical volts can be detected. They called this the 'Bereitshaftspotential' or readiness potential.

The neurobiologist John Eccles then suggested that one must be conscious of the intention to act before the increase in the Bereitshaftspotential (shown by the blue arrow below) and in 1983 Benjamin Libet decided to test this speculation. He measured when people were consciously aware that they had made a decision and found that people became conscious after, not before, the increase in the Bereitshaftspotential.



On the diagram above you can clearly see that the electrical volts required for the action (time=0sec) start to increase at a greater rate at the blue arrow. Before the blue arrow they slightly increase as the brain processes the options and chooses. Then after the blue arrow, they increase significantly as the brain has chosen an action to take and is becoming 'ready' to send the nerve impulse required to make the action happen. The red arrow is the point where the patients reported becoming conscious of the action, about 200ms before the action and 350ms after the change in the rate of increase of the electrical volts. This is a significant lag time in neuroscience and clearly indicates a period of time where the unconscious part of the brain is informing your consciousness of its decision. This clearly shows that our consciousness is not the source of our actions.

In more recent tests using Functional Magnetic Resonance Imaging (fMRI) and direct recordings from the cortices of patients about to undergo surgery, there have been lag times of up to 5 seconds recorded. This is such a long time that if scientists were able to interpret what action was going to be carried out based on the electrical volts detected in the brain preceding the action, they would be able to accurately know your actions before you.

It is clear to see from the scientific evidence started by Kornhuber, Deecke and Libet, which has been continued by scores of other scientists more recently, that our consciousness is not the source of our actions and that there is a decision-making process regarding the said action taking place in the brain well before the thought reaches the consciousness. It can also be seen clearly from the chain of reasoning above regarding when thoughts appear in consciousness that it is simply contradictory for thoughts to appear in the consciousness first before anywhere else. Hence, the first assumption that our consciousness is the sole source of our actions is false and hence it follows that we do not have free will.

In order for us to have free will both assumptions have to be satisfied. Even though we have shown the first assumption does not hold, it would be imprudent to pass over the second assumption without analysing it.

If we are free to act differently from how we do act in each scenario we are present in, it would mean that when information from a sensory neurone enters our brain, the neural pathways present in our head at that exact moment in time would be able to trigger one of multiple different outcomes based on the new piece of information it has acquired. At a glance this seems reasonable. This is because only having a singular output available seems inconceivable, and it is impossible to test whether a different outcome could occur given time is linear.

The closest test to this hypothesis that people can carry out, is to feed the same piece of information but at different moments in time to a person, and of course in such a case, multiple outcomes could occur as the neural pathways in the brain would have altered ever so slightly in between those moments.

However, if the neural pathways haven't altered whatsoever, given they are just a group of complex molecules that have been arranged in a specific way, it would make sense for them to only have one output for every input. This, in philosophy, is called pre-determinism. Seemingly the only way to introduce the construct of having multiple outputs would be to introduce an element of randomness and whilst this would allow us to act differently from how we did act in a scenario, the unconscious part of our brain still wouldn't be free. Introducing a random variable when selecting an output removes the ability to choose just as only having one output does, and with no ability to choose, we cannot have free will.

Consider replacing somebody, atom for atom, to create a new identical person. Their brain, like the rest of their body, wouldn't have changed. It would be impossible to tell the difference between the person prior to the replacement and after the replacement. Avoiding the question of whether they are the same person, would they act in the same way? As we are dealing with an identical brain, even someone who thought that there could be multiple outputs would agree that this identical person would have the same metaphorical 'list' of multiple outputs.

However, to accept that two identical brains would have the same 'list' of outputs is to accept that two people with identical genes (nature) and events in life (nurture) would have the same 'list' of outputs. These two are what cause our brain to develop in the way it does. Our genetics and what our sensory neurones relay to our brain are what forge the neural pathways in our brains and so to accept that it is purely what has happened in the past that shapes the present is to say that we live in a world of cause and effect.

This allows us to present the following argument; as everything that happened in the past cannot be altered, and the past shapes the present, then the present cannot be altered. This only works on the assumption that the past is the sole actor on the present, but, if you consider the event in time directly preceding the present, it is infinitesimally small. This event is the sole actor on the next infinitesimally small event, and that on the next, like a long chain, and that event preceding the present, no matter how small, is the past.

This contradicts the idea that there are multiple outputs that our unconscious brain is free to choose from. This is due to the fact that if the past determines the present, and when the past was the present the present was the future, then the present determines the future, and so the past must also determine the future. This leaves no room for free will, as how can we make a choice in the present when the past has already laid out an infinitely long chain of events stretching out into the future.

Thus, if there are not multiple outputs for our brain to choose from in a given scenario, then we are not free to act differently to how we actually do, and hence the second assumption is also false. Consequently, we do not have free will.

Having shown that the two assumptions required for us to have free will are false, let us look at free will as a whole. Consider a completely open-ended question that you would be able to answer in lots of different ways and one in which your answer bears no consequence to you or anyone else. If we don't have free will when answering this type of question, how can we have it in scenarios where our options are far more limited and our choices do have consequences? There are an unimaginable number of open-ended questions but we are going to use the following: 'What is your favourite film?' It would be expected that whatever film you picked, that it was your choice, and this is true, however it isn't something you consciously chose.

It is unarguable that you weren't free to choose all the films you have never heard of, but were you free to choose all the films you have heard of but couldn't or simply didn't think of? Given that we have established that you cannot know what you're going to think before you think it and that the earliest you can know is simultaneous with when you think it, you cannot know which films you will think of and which you won't. Consequently, you cannot choose which films you will think of and those you won't and so you weren't free to choose any of the films that you didn't think of.

We are left with you being free to choose from any of the films that you did think of, although you weren't free to think of those particular films. In other words, you are seemingly free to choose from a group of films you weren't free to choose. However, we can apply the same reasoning and logic that was applied to why you weren't free to think of certain films to the idea that you aren't free to choose certain films to be your favourite.

When the names of films were flying around your consciousness, thoughts of the odd positive or negative reason for choosing the said film probably came into your head. But you didn't think of every pro and con for every film that came into your head. You won't even be able to remember enough detail from most films to be able to think of every pro and con that you may have thought of when you originally saw the films. Therefore, are you free to think of all the pros and cons needed to make a free decision? The answer is simply no. Using the same logic that you weren't free to think of certain films; you weren't free to think of certain positives and negatives about the films you did think of.

Thus, if only a few opinions came to your head, were you free to choose one film to be your favourite? You couldn't choose whether the memories of the film were positive or negative and so your decision became weighted based on what did emerge into your consciousness and so one of the films was always going to come out on top, but not by your own judgement and so, were you ever free to choose any other film?

This example clearly demonstrates how little, if any freedom we have when making an open-ended decision and this can be directly translated into actions that do have consequences. Not being free to choose films that we didn't think of is the same as not being free to act in a way that didn't come into our consciousness. The inevitability of you always going to have chosen one film is equivalent to you always going to have acted in a certain way. This is further evidence for the absence and ergo, delusion of free will.

THE IMPLICATION OF A LACK OF FREE WILL

One of the main consequences of rejecting the idea of free will is the effect it has on culpability. If a murderer was always going to commit such a crime, then can we blame them for something so seemingly out of their control? If the genetics they were born with, and all the events in their life created a pathological mind-set that allows thoughts and decisions of a criminal nature to occur, and there is an absence of free will, it would appear unjust to punish them on the basis of blame and deservedness. That is not to say that they should not be punished. That would be a misunderstanding of the argument, as retribution is only one of the reasons why we have a judicial system.

Someone who has broken the law should arguably be punished as deterrence to others. Whilst others equally have no free will, events in time still influence their neural pathways, and so the event of another who has committed a crime being punished will dissuade others from committing a crime, regardless of the presence or absence of free will. Similarly, if someone has broken the law then they should be punished as a way to incapacitate them. By this I mean removing them from society as a way to protect society from their actions. The existence of or lack of free will again does not come into question when considering incapacitation as a form of punishment and the same is true for rehabilitation as a reason for punishment.

Imagine being attacked at a zoo by a lion that breaks your leg. Would you blame the lion for doing so, given it is in their genetics and the way they are raised to be aggressive? There is congruity within society to not blame the lion and regularly we even provide an excuse for the animal in the absence of it being able to provide one itself, such as 'It was just protecting its offspring'. We may even go so far as to blame ourselves and say 'I must have angered it'.

Now imagine the same outcome of you breaking a leg, but that this happened when another tourist attacked you at the zoo. There appears to be no question of whether you would deem them responsible. You would be angry at them in a way that you would never be at the lion. The sole difference between the two scenarios is man versus animal and the difference in our reactions lies in the fact that we believe the man could have acted otherwise. However, if the man could not have chosen to act otherwise, then any difference between the circumstances of the scenarios is removed and so any difference in your reactions becomes unjustified.

Therefore, if we treat the tourist in the same way as we treat the lion then we would remove the blame we put upon the tourist. Whilst this sounds quite a drastic measure, it doesn't change whether we punish the tourist or not, as we established above. However, it would remove any anger the victim may direct at the perpetrator on the grounds that they as an individual deserve it, and any thoughts of vengeance would no longer come to fruition if the victim believed wholly in the absence of free will.

Many would find this moral stance incredibly hard and rightly so. We are born into a society that teaches us to blame and it is a natural instinct to retaliate. However, if in court at your trial it had been shown that your attacker had a brain tumour in a specific region of the brain that had caused such a mind-set to arise that it was inevitable that they were going to attack you, would you then blame the attacker? Would you say that they, wholly separate from their actions, deserved the proposed punishment, because they as an individual were responsible? I would argue that in most cases, you would say no and at the very least, your perception of the attacker would change considerably. You would feel less animosity towards the attacker, and may in some cases feel a little empathy towards them, as they have suffered and could not help their actions, and you would wish for them to be treated and rehabilitated.

If scientists were able to understand the brain in much greater depth, then I believe that¬¬¬¬ what they would find in the brain of someone prone to violence, especially someone capable of murder, would be as exculpatory as finding a brain tumour. Given that a brain tumour is purely mutations of cells that haven't divided correctly, and if this can cause such detrimental effects to the point of making them psychopathic, it is conceivable that neural pathways being formed as reactions to events in the past can cause equally detrimental effects. Hence, if you consider the psychopathic brain of a murderer to have damage akin to that of a brain tumour, then not blaming the perpetrator for their actions is seemingly more reasonable.

Therefore, if we cease to believe in free will based on the neurological and philosophical evidence against its existence and the lack of evidence for it, then we must re-evaluate our reaction to events. To refute the concept of free will is not to give everyone the right to avoid responsibility for their actions. They are still the person who carried out the action and they still have to live with the consequences. However, the difference now is that our response to their action should not be carried out because they as an individual deserve the response and should be carried out because the response is in their best interests.

THE ORIGINS OF FREE WILL

Having examined the implications of an absence of free will regarding culpability, we can see why certain societies historically believed in free will. For this we need to look back to when Christianity became the dominant religion in Europe and pagan beliefs disappeared from society. This was in the 4th century and largely down to Constantine the Great, the Emperor of Rome from 306-337 AD. Almost all societies and certainly those that had a fairly structured body of governance exercised the death penalty as a form of punishment in this period.

In pagan societies criminals could be subject to execution due to pagan beliefs that execution was appeasing pagan gods, and pleasing the gods was intrinsically good. This reasoning did not hold in Christian societies as Jesus had taught to "love your enemies, bless those who curse you, do good to those who hate you, and pray for those who mistreat you and persecute you" (Matthew 5:44).

In order to be able to justify using the death penalty, most Christian societies required a greater reason for punishment than deterrence, as if this is the sole reason for using the death penalty, you are not killing a person because they deserve to be killed, but purely to dissuade others from breaking the same law. This in most societies would not be deemed a satisfactory reason as the lawbreaker is being used as a weapon of fear by the justice system and is not being punished for their own actions. Whilst these societies deemed deterrence to be an acceptable reason for punishment, they did not accept it as a reason for capital punishment.

Rehabilitation is clearly not a possible reason for using capital punishment, as you simply cannot make a dead man repent, see the error in their actions or change their behaviour. Furthermore, incapacitation is rarely a justifiable reason for the death penalty as the purpose of a prison is to incapacitate criminals and so it would be excessive to sentence them to death for that reason alone.

Therefore, historically, Christian societies required another reason, which stated that criminals did deserve the death penalty; they had committed such a heinous crime that death was the only option and they were responsible for the crime and so should bear the burden of the crime. This is retribution. Consequently, these societies looked for parts of the Bible that they could use to explain and justify that the criminal should bear the burden of the crime.

As previously stated when we first discussed retribution, it is facilitated by the existence of free will and so, when early Christians found that God had given mankind free will in Genesis 2:16, "You are free to eat from any tree in the garden", they used this to satisfy the people that retributive justice was based on God's will and so could be used as a reason for capital punishment. In addition, in several other books of the Bible humans are described as free to make choices, for example, "You, my brothers and sisters, were called to be free" (Galatians 5:13), and "Anyone who chooses to do the will of God will find out whether my teaching comes from God or whether I speak on my own" (John 7:17).

Therefore, as the Church was where people were educated, taught right from wrong and how to be a morally good person, they were taught that God had given them free will. Thus, they were able to choose to live in accordance with the values they were being taught and that they as a consequence were responsible for their actions. As the people were taught no alternative, the Church was very powerful and they were taught that these were the words of God, they did not question the existence of free will. For centuries, Christian societies unquestioningly believed that they had free will to the point where it became treated as a fact rather than a belief, despite no evidence for its existence outside of the Bible.

Even within the Bible, there are several books in which it is quite clearly indicated that man does not have free will. This contradiction within the Bible is not problematic for the Christian faith, as the Bible is not the direct word of God. The Bible is partly a history book about the Israelites; partly a book of song and prayer; party a book about the Son of God and partly a guide to future believers in God as to how to best live their lives. The Bible was written by many people across many centuries and the majority of the authors did not witness what they had written.

The issue regarding the contradiction within the Bible arises when we look at the fact that the Church largely only taught the verses of the Bible that solidified the idea that we have free will and not the ones that refute it, such as: "For he chose us in him before the creation of the world to be holy and blameless in his sight. In love he predestined us for adoption to sonship through Jesus Christ, in accordance with his pleasure and will" (Ephesians 1:4-5).

Here, it is clearly being said that you need to be holy and blameless to enter heaven and you can obtain these through Jesus Christ. But God has predetermined those who will be sons of Jesus and those who won't and did this before the creation of mankind and so only those that God picked will be allowed access to heaven. Furthermore, as you have to be free of sin to go to heaven, it can be concluded that God has already chosen who is going to sin and abstain from sin and so has already determined our actions. This would mean that we could not have free will.

Additionally, if God 'chose us in him before the creation of the world', then this supports the idea of a chain of events stretching forward from the beginning of time into the future as God must have determined who reached heaven, in the past, the present and the future. This gives a reason for everything to have been predetermined and also provides us with a determiner, which gives the argument even more gravitas.

There are multiple other examples in Ephesians and many more in other books of the Bible. I believe that given that the initial notion for free will was not founded in philosophical thought and was instead founded in the Bible, it cannot be simply accepted as fact in an international society of many faiths and moral beliefs.

CONCLUSION

In conclusion, we have defined free will, have examined the two assumptions that need satisfying for free will to exist under the definition, and have found them to be false. 'Our consciousness is the sole source of our actions' was shown through scientific evidence to be untrue and 'We are free to act differently to how we actually act' did not hold when subject to logical analysis and reasoning.

Having shown that free is a delusion; we proceeded to look at its implications on culpability. With the absence of free will, retribution ceases to stand as a fair reason for justice and punishment. Finally, we looked at how early Christian societies selectively used the Bible to justify punishment that would not be reasonable without an acceptance of free will.

In my opinion, there is ample evidence across neurology, biblical studies and philosophy to suggest that we should re-evaluate the assumption that we have free will. We should consider the weight of evidence against the presence of free will in contrast to the absence of any concrete evidence for the alternative, given how great an effect it has on the justice system. I believe I have clearly shown why we should discuss whether we have free will and why it is much more than a purely academic argument.

It is my view that the only reason that we continue to hold onto the belief that we have free will is that we want to feel 'free' and don't like the idea of not having control of our actions. This is unsurprising, as the alternative can appear scary. However, in many cases people are simply not aware of the possibility of a lack of free will and this in itself is enough of a reason to have the debate. I believe that if the debate were to be had, fewer people would be under the delusion that we have free will.

La Loi de Benford en France et en français

Marcus Hinton

1. INTRODUCTION

Si vous deviez chercher les chiffres significatifs de données individuelles dans un ensemble de données naturelles, vous vous attendriez probablement à trouver que le chiffre significatif aurait la même probabilité d'être l'un des neufs premiers chiffres significatifs (1 à 9). En autre mots, si vous traciez un graphique représentant le nombre de fois où les chiffres de 1 à 9 seraient significatifs, vous pourriez vous attendre à un résultat uniforme. Cependant, en réalité en donnés standards ce n'est souvent pas le cas. En effet, le chiffre 1 est le premier chiffre significatif dans près de 30% des cas[A], et puis les autres chiffres (deux à neuf) suivent une distribution logarithmique ; 2 arrive environ dans 17.5% des cas, puis trois arrive environ dans 12.5% des cas et ainsi de suite. C'est la Loi de Benford[B], une loi mathématique plus applicable lorsque l'on regarde des données couvrant plusieurs ordres de grandeur, et bien que cela semble un peu contre-intuitif au début, cela a beaucoup de sens.

Lors de mes recherches sur la Loi de Benford, j'ai pensé qu'il serait intéressant d'examiner les statistiques de la France en raison de mon intérêt pour le pays et sa culture : la France est le pays le plus visité du monde – à cause de de son éventail d'activités et de lieux à visiter, tels que les Alpes ou le sud de la France – et il existe donc des données importantes sur le tourisme, les populations, les hauteurs de montagne et les longueurs de rivières entre autres. La France est un pays bureaucratique. Cela signifie que les gouvernements locaux et centraux collectent une grande quantité de données sur une grande variété de domaines. Donc, j'ai pensé qu'il serait intéressant de fournir mon analyse sur les raisons pour lesquelles la Loi de Benford fonctionne, et en plus d'explorer divers ensembles de données existant en France et voir si elle supporte la Loi de Benford. Personnellement, je trouve que l'idée de la Loi de Benford est fascinante, même si on peut la trouver absurde au prime abord ; en effet pourquoi les premiers chiffres de toutes les données ne seraient-ils pas distribués de manière uniforme ? Compte tenu de mon utilisation des données françaises pour étudier la Loi de Benford, j'ai décidé d'écrire mon essai en français.

2. POURQUOI LA LOI DE BENFORD FONCTIONNE

Je vais examiner deux manières de démontrer pourquoi la Loi de Benford est vraie.

2.1 LES NOMBRES NATURELS

Tout d'abord, j'examinerai la probabilité d'apparition des premiers chiffres des nombres naturels, au fur et à mesure que les nombres deviennent de plus en plus grands. Si on considère les chiffres naturels (1,2,3,4 etc.), après le premier chiffre, la probabilité que le premier

1. INTRODUCTION

If you were to look at the leading digits of individual data points in a range of naturally occurring data, you would probably expect the leading digit of each data point to have an equal chance of being any of the nine possible digits (the numbers one to nine), meaning that if you were to plot a graph of how many times the numbers one to nine were the leading digits of a data point, you might expect therefore all nine numbers would appear an equal amount of times. In reality however, in naturally occurring data this is often not the case. The number one actually appears as the leading digit of the data about 30% of the time[A], and then then the other numbers (two to nine) follow a logarithmic distribution; two occurs about 17.5% of the time, then three occurs about 12.5% of the time and so on. This is Benford's Law[B], a mathematical law most applicable when looking at data spanning multiple orders of magnitude, and although it seems a bit counter-intuitive at first, it actually makes a lot of sense.

When researching Benford's Law I thought it would be interesting to examine the statistics of France due to my interest in the country and its culture: France is the most visited country in the world – due to its range of activities to do and places to visit, such as the Alps or the South of France – and so there is significant data on tourism as well as populations, mountain heights, river lengths and more. France is a country that enjoys a level of bureaucracy. This does mean that local and central government collects a vast amount of data on a huge variety of areas. Therefore, I thought it would be interesting firstly to provide my analysis of why Benford's Law works, and secondly to explore various data sets that occur in the country of France and see whether they fit Benford's Law. Personally, I find the idea of Benford's Law fascinating, as when you first think about it seems nonsensical: why wouldn't the leading digits of all the data be evenly distributed? Given my use of French data to look for evidence of Benford's Law, I decided to write the essay in French.

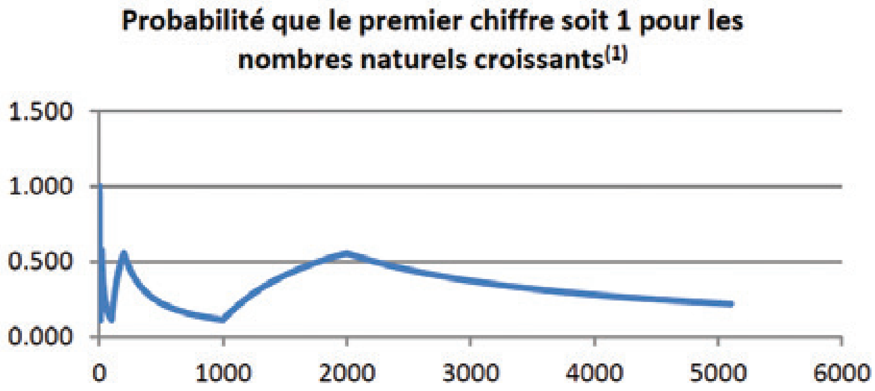
2. SHOWING WHY BENFORD'S LAW WORKS

I am going to consider two ways to demonstrate why Benford's Law is true..

2.1 THE NATURAL NUMBERS

Firstly, I will examine the probability of the occurrence of the leading digits of natural numbers, as the numbers get progressively larger. If we consider the natural numbers (1,2,3,4 etc.), after the first number the probability of the leading digit being a one is one, as there has only been one number and it begins with a one. After the second number, the probability has halved as there are now two numbers and one is chiffre soit 1 est égal à 1, car il n'y a eu qu'un seul chiffre et il commence par 1. Après le deuxième chiffre, la probabilité a été divisée par deux car il y a maintenant deux chiffres, l'un est 1 et l'autre est 2. Après l'introduction du troisième chiffre, la probabilité devient 1/3 (un tiers) et ainsi de suite jusqu'au neuvième chiffre, quand la probabilité est de 1/9 pour obtenir 1. Cependant, quand nous arrivons au nombre dix, la probabilité que le premier chiffre soit un 1 augmente en raison de la série de nombre entre 10 et 19, qui commence tous par 1. Nous pouvons aussi calculer la probabilité que le premier chiffre soit un 1quand les nombres deviennent des milliers et des dizaines de milliers. Nous pouvons ainsi tracer un graphique représentant la probabilité que le premier chiffre de l'un de ces nombres soit un 1Le graphique ressemble à ceci :

a one and the other is a two. After the third number is introduced, the probability becomes a third and so on until the ninth number, when the probability of a one occurring is 1/9. However, once we arrive at the number ten the probability of the leading digit being a one increases, due to the range of numbers between 10 and 19, all of which begin with a one. We can continue to calculate the probability of the leading digit being a one as the numbers increase into the thousands and tens of thousands, and we can plot a graph of the probability of the leading digit of one of the numbers being a one against the number of numbers being considered. The graph looks like this:

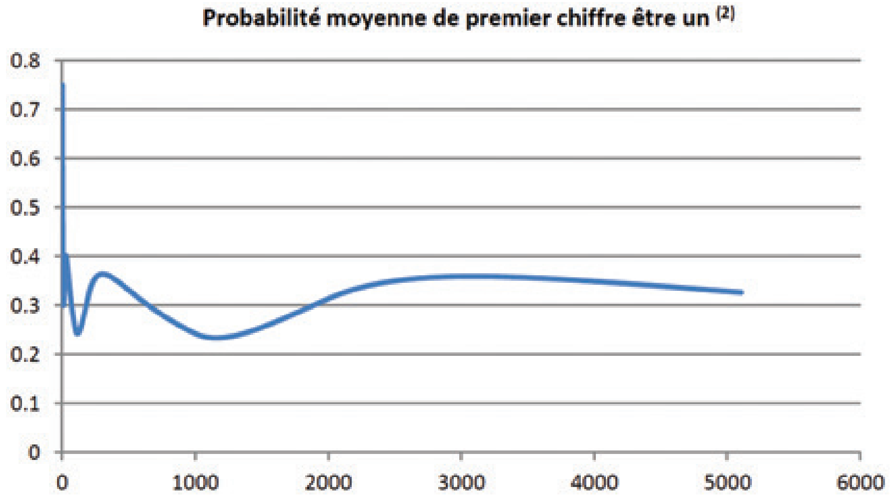


Comme on peut le constater, la probabilité que le premier chiffre d'un nombre aléatoire soit un 1 varie considérablement lorsque le nombre considéré augmente, toutefois il s'ensuit un schéma général d'augmentation rapide de certaines gammes avant de diminuer lentement au cours de la période suivante. Le graphique montre clairement que la probabilité minimale que le premier chiffre d'un nombre choisi au hasard dans une série donnée soit 1 est de 1/9 et la ligne fluctue d'environ 0,11 à 0,56. Une moyenne serait donc quelque part entre ces deux valeurs mais considérablement plus grand que 1/9. Par conséquent, nous pouvons déjà voir que l'idée que les chiffres 1-9 ont une chance égale d'être le premier chiffre est incorrecte, du moins lors de l'analyse de nombres naturels croissants.

Nous pouvons maintenant tracer un graphique de la probabilité moyenne que le 1 soit le premier chiffre (d'un nombre choisi au hasard) en additionnant toutes les probabilités à n et puis en les divisant par n. Par exemple avec le nombre1000, nous pouvons additionner toutes les probabilités où 1 est le chiffre significatif des nombres de 1 à 1000, puis divisons ce total par 1000. Si nous faisons ensuite ce calcul pour chaque point le long de l'axe X et traçons les probabilités moyennes par rapport au nombre de nombres naturels (probabilités) considérés, nous obtenons un graphique qui ressemble à ceci :

As we can see, the probability of the leading digit of a random number being a one varies significantly as the numbers being considered increase, however it follows a general pattern of sharply increasing in certain ranges before slowly decreasing over the following period. It can be clearly seen from the graph that the minimum probability of the leading digit of a randomly selected number in a given range being a one is 1/9, and the line fluctuates from around 1/9 to 0.56, and so an average would be somewhere between the two, and significantly bigger than 1/9. Therefore, we can already see that the idea that the digits 1-9 have an equal chance of being the leading digit is incorrect, at least when analysing increasing natural numbers.

We can now plot a graph of the average probability of a one being the leading digit (of a randomly selected number) by adding up all the probabilities to n and then dividing them by n. i.e. if we were to look at 1000, we would add up all the probabilities of one being the leading digit for the numbers 1 to 1000, and then divide that total by 1000. If we then do that calculation for every point along the x axis and plot the average probabilities against the number of natural numbers (probabilities) being considered, we get a graph that looks like this:



2.2 INVARIANCE D'ÉCHELLE

Encore une fois, ce graphique nous montre quelque chose de très intéressant ; la probabilité moyenne que le premier chiffre soit égal à 1 nous donne une valeur. Cette valeur est log (2) ou 0,301 et représente la valeur, selon La théorie de Benford, que le premier chiffre du point de données aléatoires soit 1, sur de nombreuses sources de données quel

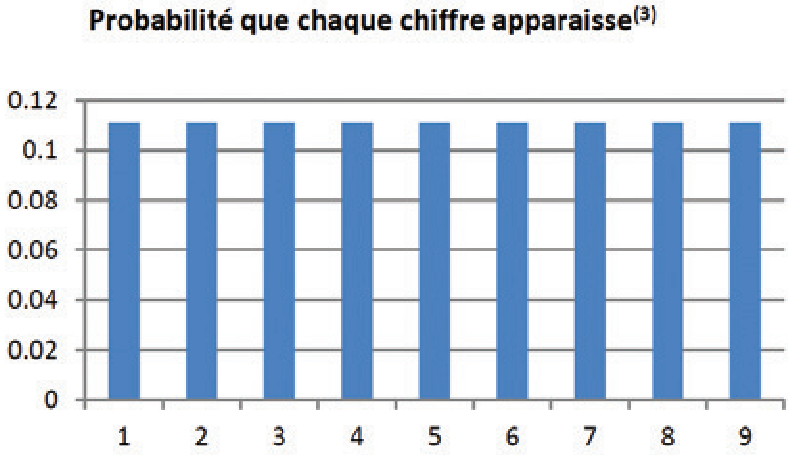
que soit le domaine observé. Par conséquent, nous pouvons constater qu'en examinant simplement le premier chiffre significatif d'un ensemble de valeurs numériques supporte la Loi de Benford et l'on peut comprendre pourquoi elle s'appliquerait sur de nombreuses sources de données ; la probabilité que le premier chiffre soit 1 est beaucoup plus élevé que 1/9.

Deuxièmement, nous pouvons utiliser des unités et des échelles pour montrer pourquoi la Loi de Benford est une distribution logarithmique. Pour commencer, nous devons préciser que cette méthode est basée sur le principe suivant : si la répartition des chiffres significatifs dans un large ensemble de données est structurée, il devrait être insensibles aux changements d'unités de mesure (on peut appeler cela notre premier principe). En effet, il n'y a aucune raison pour qu'un système d'unités arbitraire ait un effet sur le modèle de distribution. En effet, une distribution d'une population est uniforme par rapport au premier chiffre lorsqu'il est mesuré en unités Y, c'est-àdire $P(1) = P(2) = P(3) = P(4) = \cdots = P(9)$. Ceci peut être montré sous forme graphique :

2.2 SCALE INVARIANCE

Again, this graph shows us something very interesting; the average probability of the leading digit being 1 converges to a value. This value is log (2), or 0.301, and is the value that Benford's Law states is the probability of a random data point's leading digit being one, in a set of naturally occurring data that spans multiple orders of magnitude. Therefore, we can see that just by examining the leading digits of natural numbers, it is evident that Benford's Law holds true, and it is clear to see why it would hold true for large data sets; the probability of the leading digit being a one is a lot higher than 1/9.

Secondly, we can use units and scales to show why Benford's Law is a logarithmic distribution. To begin with, we must clarify that this method is based on the principle that if there is a pattern behind the distribution of leading digits in large data sets, then it should be invariant to changes in measurement units (let's call this our first principle). This is because there is no reason why an arbitrary unit system should have an effect on the distribution pattern. Now, a distribution of a population is uniform with respect to the leading digit when measured in units Y, i.e. $P(1) = P(2) = P(3) = P(4) = \cdots = P(9)$. This can be shown in graphical form:



(4)

Valeur des données	Premier chiffre
$1x10^n \leq x < 2x10^n$	1
$2x10^n \leq x < 3x10^n$	2
$3x10^n \leq x < 4x10^n$	3
$4x10^n \leq x < 5x10^n$	4
$5x10^n \leq x < 6x10^n$	5
$6x10^n \leq x < 7x10^n$	6
$7x10^n \leq x < 8x10^n$	7
$8x10^n \leq x < 9x10^n$	8
$9x10^n \leq x < 10x10^n$	9

Supposons maintenant que l'unité de mesure change pour une unité différente F telle que 1Y = 3F. Un exemple de ceci serait les yards et les pieds. Maintenant, un nombre commençant par 1 commencerait par 3,4 ou 5 et un nombre commençant par 2 commencerait par 6,7 ou 8. Cela signifie que la distribution des chiffres significatifs a changé, comme on peut le voir ici :

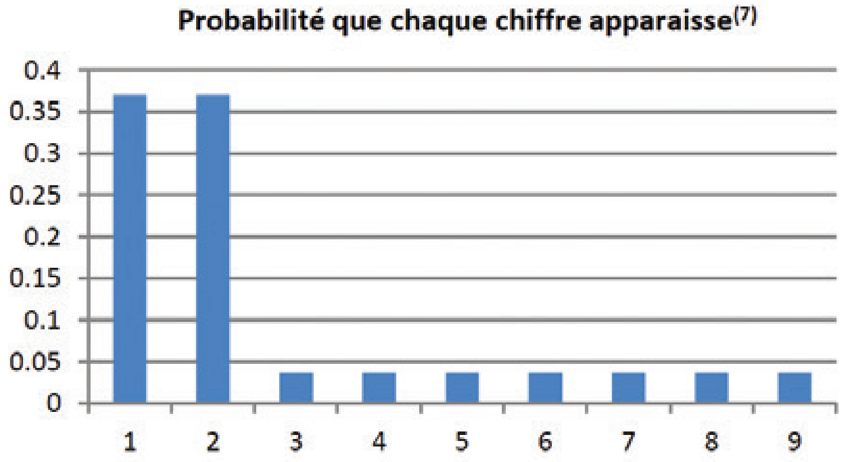
Now suppose that the measurement unit is changed to a different unit F, such that 1Y = 3F. Yards and feet would be an example of this. Now a number that begun with a one would begin with a 3,4 or 5 and a number that began with a 2 would begin with a 6,7 or 8. This means the distribution of the leading digits has changed, as we can see here

(5)

Valeur d'orinine des données	Valeur des données avec de nouvelles unités	Nouveau premier chiffre
$1x10^n \leq x < 2x10^n$	$3x10^n \leq x < 6x10^n$	3 (33.3%), 4 (33.3%), 5 (33.3%)
$2x10^n \leq x < 3x10^n$	$6x10^n \leq x < 9x10^n$	6 (33.3%), 7 (33.3%), 8 (33.3%)
$3x10^n \leq x < 4x10^n$	$9x10^n \leq x < 12x10^n$	9 (33.3%), 1 (66.6%)
$4x10^n \leq x < 5x10^n$	$12x10^n \leq x < 15x10^n$	1 (100%)
$5x10^n \leq x < 6x10^n$	$15x10^n \leq x < 18x10^n$	1 (100%)
$6x10^n \leq x < 7x10^n$	$18x10^n \leq x < 21x10^n$	1 (66.6%), 2 (33.3%)
$7x10^n \leq x < 8x10^n$	$21x10^n \leq x < 24x10^n$	2 (100%)
$8x10^n \leq x < 9x10^n$	$24x10^n \leq x < 27x10^n$	2 (100%)
$9x10^n \leq x < 10x10^n$	$27x10^n \leq x < 30x10^n$	2 (100%)

(6)

Premier chiffre	Probabilité de chiffre	
1	(1/9 x 2/3) + 1/9 + 1/9 + (1/9 x 2/3)	10/27=0.3704
2	(1/9 x 1/3) + 1/9 + 1/9 + 1/9	10/27=0.3704
3	1/9 x 1/3	1/27=0.03704
4	1/9 x 1/3	1/27=0.03704
5	1/9 x 1/3	1/27=0.03704
6	1/9 x 1/3	1/27=0.03704
7	1/9 x 1/3	1/27=0.03704
8	1/9 x 1/3	1/27=0.03704
9	1/9 x 1/3	1/27=0.03704

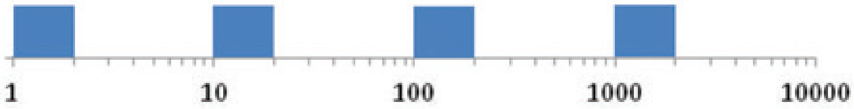


Comme on peut le voir sur le tableau et le diagramme à barres, il est maintenant beaucoup plus probable que le premier chiffre des mêmes données soit 1 ou 2, uniquement à cause du changement d'unités. Cela implique qu'une distribution uniforme des chiffres significatifs est modifiée si l'unité de mesure est modifiée, ce qui est contraire à notre principe selon lequel un changement d'unités ne devrait pas avoir d'effet sur la répartition des premiers chiffres.

Cela démontre que les chiffres 1-9 ne peuvent pas avoir une probabilité égale d'être le chiffre significatif d'un point de données aléatoire, car cette distribution ne reste pas la même lorsque les unités de données changent. Par conséquent, nous devons trouver la distribution qui est insensible aux changements d'échelle.

As we can see from the table and the bar chart, it is now much more likely for the leading digitof the same data to be a one or a two, solely because of the change in units. This implies that a uniform distribution of leading digits is altered if the unit of measurement is altered, which is contrary to our principle saying a change in units should have no effect on the distribution of the leading digits.

This shows us that the leading digits 1-9 cannot have an equal probability of being the leading digit of a random data point, as this distribution does not stay the same when the units of the data change. Therefore we must find the distribution which is invariant to scale changes.



Considérons maintenant cet axe logarithmique; choisissez un nombre au hasard le long de l'axe et regardez le premier chiffre significatif . La section bleue représente les nombres commençant par 1 ; par conséquent, la probabilité que le premier chiffre significatif soit 1 est obtenu selon la proportion de l'axe qui est bleue. Si vous choisissez au hasard un nombre compris entre 1 et 10 et un autre entre 1000 et 10 000, les deux nombres ont une probabilité égale de commencer avec un 1. Cela signifie que cette distribution logarithmique est insensible au changement d'échelle et n'est donc pas affectée par un changement d'unités, ce qui signifie qu'elle satisfait notre principe. Si nous examinons cet axe de plus près, nous pouvons voir que :

Now consider this logarithmic axis; pick a random number along the axis and look at its leading digit. The blue section represents numbers beginning with 1, and therefore the probability of the leading digit being a one can be found from the proportion of the axis that is blue. If you pick a random number between 1 and 10, and another between 1000 and 10000, both have equal probability to start with a one. This means that this logarithmic distribution is invariant to scale change and therefore not effected by a change of units, meaning it satisfies our principle. If we examine this axis more closely, we can see that:

$$P(1) = \frac{\log 2 - \log 1}{\log 10 - \log 1} = 0.301$$

$$P(1) = \frac{\log 2000 - \log 1000}{\log 10000 - \log 1000} = 0.301$$

En effet, lorsque nous modifions les unités d'une distribution, nous la multiplions par un facteur d'échelle. Cependant, lorsque nous appliquons cette méthode à une distribution logarithmique uniforme, nous ne faisons que nous déplacer sur l'axe, car :

This is because when we change the units of a distribution we multiply it by a scale factor. However, when we do this to a uniform logarithmic distribution, all we do is shift along the axis, since:

$$\log(kx) = \log(x) + \log(k)$$

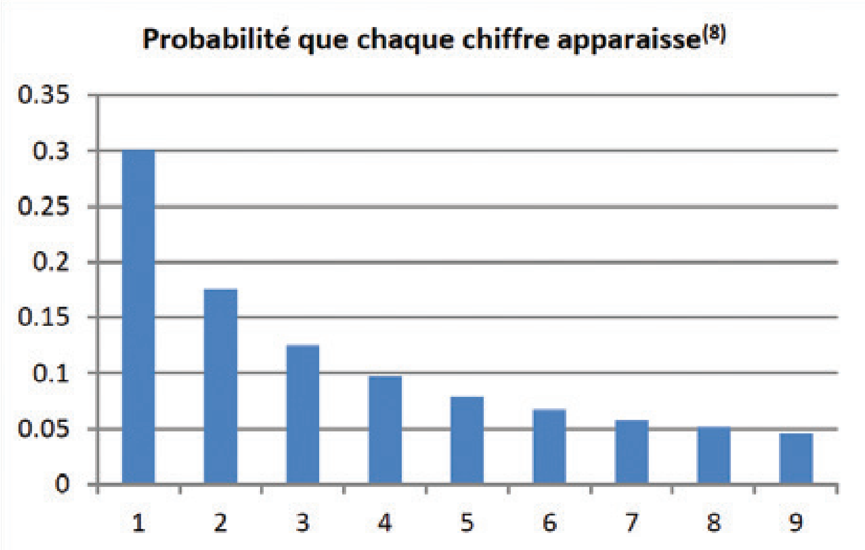
Par conséquent, pour une distribution logarithmique uniforme, la probabilité que le premier chiffre significatif soit compris entre 1 et 9 est constante pour chaque chiffre (entre 1 et 9), quelle que soit l'échelle que vous utilisez. Nous pouvons également utiliser cette échelle pour calculer la probabilité que le premier chiffre significatif soit n, pour des données qui obéissent parfaitement à la Loi de Benford :

Therefore, for a uniform logarithmic distribution, the probability of the leading digit being a certain number between 1 and 9 is constant for that number, regardless of what scale you are at. We can also use this scale to calculate the probability of the leading digit being n, for data that perfectly obeys Benford's Law:

$$P(n) = \log(n+1) - \log(n) = \log(1+\frac{1}{n})$$

En utilisant cette formule, nous pouvons maintenant tracer un graphique de la distribution logarithmique des chiffres significatif de ces données :

Using this formula, we can now plot a graph of the logarithmic distribution of the leading numbers of the data:


Probabilité que chaque chiffre apparaisse[8]

C'est la distribution parfaite de la Loi de Benford ; si vous tracez le nombre de fois où chaque chiffre de 1 à 9 est le premier chiffre significatif de (nombreuses) données couvrant plusieurs ordres de grandeur, vous trouverez cette distribution. Cette échelle de distribution est complètement insensible, ce qui signifie que peu importe les unités que vous utilisez, vous obtiendrez ce graphique. La distribution ressemble beaucoup au graphique de y = 1 / x, ce qui signifie que les probabilités de chaque chiffre (trouvées si nous intégrons la courbe) sont proportionnelles à $log_e x$, ce qui signifie que les probabilités sont trouvées en utilisant les logarithmes (voir l'équation 2 ci-dessous).

Après avoir montré comment l'invariance d'échelle implique la Loi de Benford, nous pouvons aller encore plus loin et utiliser le calcul pour montrer pourquoi une distribution insensible à l'échelle doit suivre la Loi de Benford et son modèle logarithmique. Si nous considérons que la distribution de probabilité P (x) du premier chiffre est l'un des chiffres de

This is the perfect Benford's Law distribution; if you plot the number of times each number 1-9 is the leading digit of (lots of) data spanning multiple orders of magnitude, this is the distribution you will see. This distribution is completely scale invariant, meaning regardless of what units you use you will get this graph. The distribution looks very similar to the graph of y=1/x, meaning the probabilities of each digit occurring (found if we integrate the curve) is ln(x), or $log_e x$, meaning the probabilities are logarithmic.

Having shown how scale invariance implies Benford's Law, we can go one step further and use calculus to show why a scale-invariant distribution must follow Benford's Law and follow the logarithmic pattern. If we consider the probability distribution P(x) of the first digit being one of the numbers 1 to 9, and the distribution must be scale invariant meaning:

1 à 9, la distribution doit être invariante à l'échelle,alors :

$$P(kx) = CP(x) \text{ (Équation 1)}$$

Où « C » est une fonction de k (c'est-à-dire un facteur général qui dépend de k). Puisque P (x) est une distribution de probabilité, si nous intégrons toute la distribution, elle doit être égale à un, car c'est la probabilité maximale possible qu'un événement se produise.

Where "C" is some function of k (i.e. some general factor which depends on k). Since P(x) is a probability distribution, if we integrate the entire distribution it must be equal to one, because that is the total probability of the events occurring.

$$\int P(x).\, dx = 1$$

Alors

So

$$\int P(kx).\, dx = \int \frac{P(y)}{k}.\, dy = 1/k \int P(y).\, dy = 1/k \qquad \text{(Où y = kx et dy = kdx)}$$

Et (à partir de l'équation 1) :

And (from equation 1):

$$\int CP(x).\, dx = 1/k$$

Donc

Therefore

$$C \int P(x).\, dx = 1/k$$

Donc

Therefore

$$C(k) = 1/k$$

Si nous retournons alors à l'équation 1 P (kx) = CP (x) et différencions :

If we then return to equation 1 P(kx)=CP(x) and differentiate:

$$\frac{d}{dk} P(kx) = \frac{dC}{dk} P(x)$$

Donc

Therefore

$$x \times \frac{dP(kx)}{d(kx)} \times \frac{d(kx)}{dk} = \frac{d}{dK}\left(\frac{1}{k}\right) \times P(x) = (-1/k^2)\; P(x)$$

Donc si on laisse maintenant k = 1 on obtient :

So if we now let k=1 we get:

$$x \frac{d}{dx} P(x) = -P(x)$$

Ceci est une équation différentielle. Résolution :

This is a differential equation. Solving:

$$\frac{1}{P} dP = -\frac{1}{x} dx$$

Donc

Therefore

$$\int 1/P \,.\, dP = -\int \frac{1}{x} dx$$

Donc

Therefore

$$\ln(P) = -1(\mathrm{x}) = \ln(x^{-1})$$

(Car nous ne sommes préoccupés que par les chiffres 1-9, nous pouvons ignorer le fait que $\int_0^\infty 1/x.dx$ n'est pas bien défini à 0 ou ∞)

(Because we are only worried about the digits 1-9 we can ignore the fact that $\int_0^\infty 1/x.dx$ is not well defined at 0 or ∞)

$$P = 1/x \text{ (Équation 2)}$$

Laisse D= un nombre aléatoire de 1 à 9.

Donc P (le premier chiffre est D) =

Let D=a random number 1 to 9.

Therefore P(first decimal is D)=

$$\frac{\int_D^{D+1} P(x).dx}{\int_1^{10} \mathrm{P(x)}.dx} = \frac{[\ln(x)]_D^{D+1}}{[\ln(\mathrm{x})]_1^{10}} = \frac{[\log(x)]_D^{D+1}}{[\log(\mathrm{x})]_1^{10}} = \frac{\log(D+1) - \log(D)}{1} = \log(1+\frac{1}{D})$$

(car log(x)=ln(x)(log$e$))

(since log(x)=ln(x)(log$e$))

Maintenant nous pouvons voir pourquoi la distribution est logarithmique et car les chiffres qui suivent la distribution ; c'est La Loi de Benford.

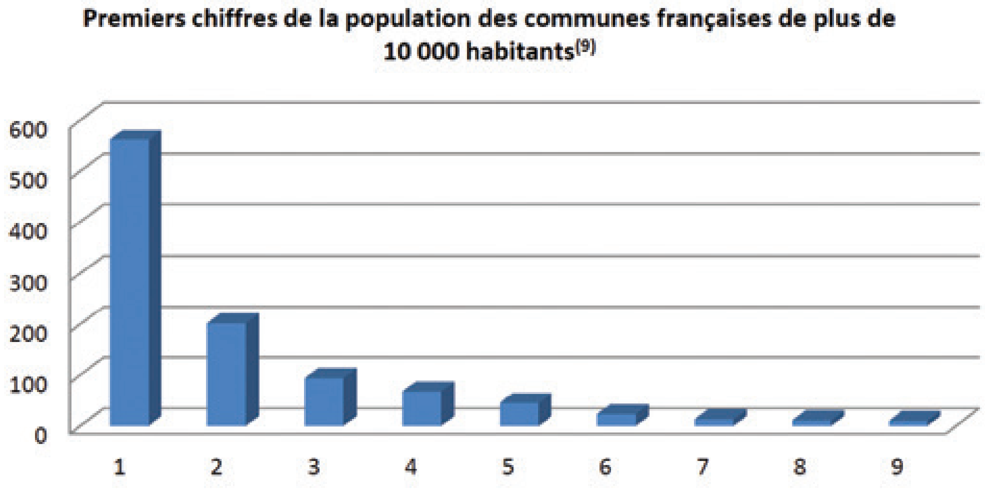
Now we can see why the distribution is logarithmic, and why the digits follow the distribution that they do; this is Benford's Law.

3. JEUX DE DONNÉES EN FRANCE

Maintenant que nous avons démontré que la Loi existe, nous pouvons regarder à quelques exemples réels de la Loi de Benford et les examiner pour voir dans quelle mesure ils correspondent à la distribution idéale. Premièrement, nous pouvons examiner les populations des communes françaises de plus de 10 000 habitants[C]. En France, une commune est le nom donné à la plus petite division administrative ; une commune est composée de toutes les parties d'une ville ou d'un village du même nom. Les seules exceptions à cette règle sont Paris, Lyon et Marseille, qui sont divisés en arrondissements, chacun représentant l'équivalent d'une grande commune. Les arrondissements sont utilisés pour diviser les grandes villes en de plus petits secteurs et Paris compte vingt arrondissements, dont certains ont une population équivalente à une petite ville. Quoi qu'il en soit, il y a eu un recensement en France en 2013 pour connaître la population des différentes communes et à partir des données de ce recensement, nous pouvons examiner la taille de la population de toutes les communes de France comptant plus de 10 000 habitants, puis tracer un graphique du nombre de fois où le chiffre de 1-9 est le premier chiffre significatif de l'une de ces populations, et cela nous donne ce graphique :

3. DATA SETS IN FRANCE

Now that we have shown why the law works, we can try and look at some real-life examples of Benford's Law and examine them to see how well they fit the ideal distribution. Firstly, wecan examine the populations of French communes above 10,000 people[C]. In France, a commune is the name given to the lowest administrative division; a commune is made up of all the parts of a town or a village under the same name. The only exception to this is Paris, Lyon and Marseille, which are split into arrondissements, each of which is the equivalent of a large commune. Arrondissements are used to split particularly large cities into smaller sectors, and Paris has twenty arrondissements, some of which have the population equivalent to a small city. In 2013 there was a census in France to determine the populations of the different communes, and using the data from this census we can look at the population sizes of all the communes in France with a population of more than 10,000 people, and then plot a graph of the number of times each number 1-9 is the leading digit of one of the populations, we are met with this graph:

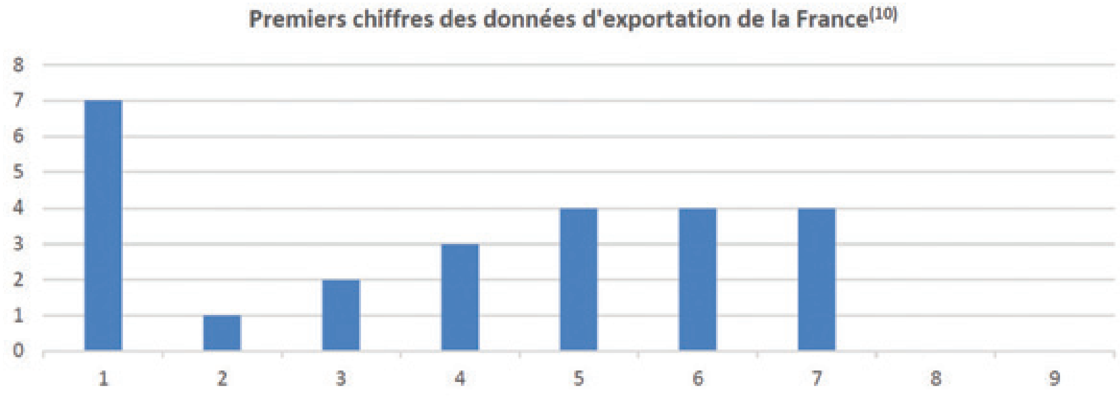


Ce graphique montre clairement que la population de ces communes suit à peu près la Loi de Benford, car on voit que le chiffre 1 apparaît plus de deux fois plus souvent que le chiffre 2, puis le chiffre 2, deux fois plus que le chiffre 3, etc. Cela est probablement dû au fait que la taille de la population dans ces communes va de 10 003 (Pélissanne, une commune du sud de la France proche de Salon-de-Provence) à 466 000 (Toulouse, capitale de la région Occitanie, en France) et que ces données existent sans que personne n'a modifié ces chiffres ou arrondi les populations.

C'est un exemple qui suit exceptionnellement bien la Loi de Benford, mais il existe certains types de données qui, bien qu'elles suivent globalement la Loi de Benford, ne le démontre pas aussi efficacement. Par exemple, si nous examinons les exportations économiques de la France en 2017[D], nous pouvons tracer un graphique similaire des premiers chiffres significatifs par rapport au nombre de fois où ils apparaissent :

This graph clearly shows that the populations of these communes follows broadly Benford's Law, as we can see that the number one appears more than twice as often as the number two, and then the number two twice as much as the number three and so on. This is likely due to the fact that the population sizes in these communes go from 10,003 (Pélissanne, a commune in the south of France near to Salon-de-Provence) to 466,000 (Toulouse, the capital city of France's Occitanie region) and the fact that the data is naturally occurring; no-one has tampered with the numbers or rounded populations up or down.

This is an example that follows Benford's Law exceptionally well, but there are some types of data that, although they broadly follow Benford's Law, they do not show it anywhere near as effectively. For example, if we look at France's economic exports of 2017[D] we can plot a similar graph of leading digits against the number of times they appear:



Comme le montre ce graphique, ces données ne suivent pas la Loi de Benford aussi clairement que l'exemple précédent. Cela est dû au fait que les données sont insuffisantes : les principales exportations françaises sont réparties en 24 catégories (par exemple, les aéronefs et les véhicules spatiaux - l'une des exportations les plus importantes de la France avec 51,48 milliards de dollars en 2017 - et les plastiques, par exemple), alors c'est plus difficile de voir une distribution uniforme. De plus, les principales exportations françaises ne varient pas beaucoup en valeur ; le plus bas est pour les meubles, les enseignes lumineuses et les bâtiments préfabriqués, d'une valeur de 3,99 milliards de dollars en 2017 et le plus élevé pour les machines, les réacteurs nucléaires et les chaudières, d'une valeur de 60,79 milliards en 2017. Comme nous pouvons le constater, il n'y a pas de grande différence entre les valeurs les plus élevées et les plus basses de cet ensemble de données, ce qui souligne la nécessité de disposer de large base de données pour démontrer la Loi de Benford. Les hauteurs des montagnes seraient un autre exemple qui ne fonctionnerait pas aussi bien. En France, elles vont de 4 810 mètres (Mont Blanc) à environ 1 000 mètres, ce qui, encore une fois, n'est pas suffisant pour que la Loi de Benford soit vérifiée.

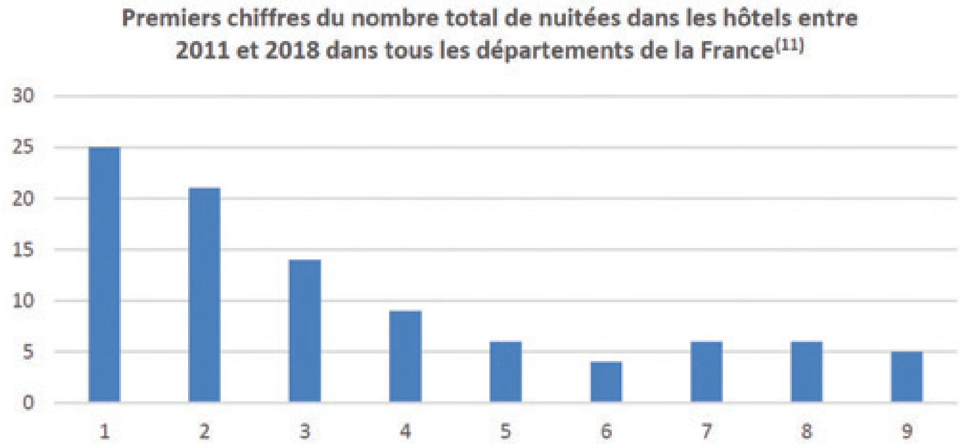
Nous avons donc établi que pour que la Loi de Benford soit évidente, nous avons besoin de large base de données. Comme je l'ai mentionné précédemment, la France est le pays le plus visité au monde et rassemble une quantité considérable de données sur le tourisme. La France est composée de 101 départements, qui sont des sections plus petites des grandes régions (il y en a 22). Par exemple, le département de la Savoie fait partie de la plus grande région Auvergne-Rhône-Alpes dans les Alpes françaises. Cependant, la Haute-Savoie, l'Ardèche, le Rhône et bien

As this graph shows, this data does not follow Benford's law anywhere near as well as the previous example. This is due to the fact that there are insufficient data points: France's main exports are broken down into 24 categories (e.g. aircraft and spacecraft – one of France's highest valued exports at $51.48 billion in 2017 – and plastics is another), meaning it is much harder for the distribution to be seen. Furthermore, France's main exports do not vary much in their value; the lowest is furniture, lighting signs and prefabricated buildings which had a value of $3.99 billion in 2017, and their highest is machinery, nuclear reactors and boilers, which had a value of $60.79 billion in 2017. As we can see from this there is not a large difference between the highest and lowest values in this set of data, emphasising the need for data to span several orders of magnitude in order to obey Benford's Law. Another example that would not work as well would be mountain heights, as in France they range from 4,810 metres (Mont Blanc) down to around 1,000 metres, which is again not a sufficient spread for Benford's Law to be apparent.

So, we have established that in order for Benford's Law to be evident we need data that is varied across multiple orders of magnitude. As I mentioned earlier, France is the most visited country in the world, and collects a huge amount of data on tourism. France is made up of 101 departments, which are smaller sections of the larger regions (of which there are 22). For example, the department Savoie is a part of the larger region of the Auvergne-Rhône-Alps in the French Alps. However, Haute-Savoie, Ardèche, Rhône and others all also all departments of the Auvergne-Rhône-Alps region. France attracts tourists for a number of reasons; whether it is fine wine, skiing, beautiful cities or luxurious d'autres sont tous les départements de la région Auvergne-Rhône-Alpes. La France attire les touristes pour plusieurs raisons. Que ce soit pour le bon vin, le ski, les belles villes ou les plages luxueuses que vous essayez de trouver pendant vos vacances, l'une des régions ou Départements de France saura satisfaire vos besoins. Cependant, il y a bien sûr des régions de la France qui ne sont pas aussi attrayantes pour les touristes, car il y a moins à faire ou à voir et il y a donc une large distribution de touristes pour les différentes régions de France. Si nous examinons le nombre total de nuits passées dans des hôtels de chaque département entre 2011 et 2018[E], les variations sont, par exemple, beaucoup plus importantes que pour les hauteurs de montagne. Le plus grand de ces chiffres est (sans surprise ici) Paris, avec un total de 287 083 100 nuitées dans des hôtels entre 2011 et 2018. Paris est la région la plus visitée de France et la troisième ville la plus visitée au monde en raison de ses nombreux monuments historiques (tels que la cathédrale Notre-Dame, le Sacré-Coeur, le Louvre, l'Arc de Triomphe, la Tour Eiffel, etc.), sa cuisine de renommée mondiale et sa belle architecture. Le département de la France avec le plus petit nombre total de nuitées est l'Ariège (avec seulement 1 937 180), un département essentiellement agricole de la région Occitanie. L'Ariège compte seulement 153 000 habitants (en 2016), ce qui en fait le septième plus petit département en France. Bien qu'ayant quelques-uns monuments historiques, elle n'est pas aussi attrayante pour les touristes que la plupart des autres régions du pays. En comparant ces deux valeurs (287 083 100 et 1 937 180), on constate que le nombre pour Paris est près de 150 fois supérieur à celui de l'Ariège, ce qui montre non seulement l'ampleur de la différence de nombre de touristes en France, mais également l'application de la Loi de Benford qui est plus facile avec ces données qu'avec les données précédentes.

beaches that you are trying to find on your holiday, one of the regions or departments of France will satisfy your needs. However, there are of course areas of France which are not as attractive to tourists, and so there is a large range in tourist numbers for the regions of France. If we look at the total number of overnight stays in hotels in each department between 2011 and 2018[E], there is a much larger variation in numbers than for mountain heights, for example. The largest of these numbers is (no surprises here) Paris, with a total of 287,083,100 overnight stays in Hotels between 2011 and 2018. Paris is the most visited part of France and the third most visited city in the world, due to its array of historic monuments (such as Notre Dame Cathedral, Sacré-Coeur, the Louvre, the Arc de Triumphe, the Eiffel Tower and more), world-renowned food and its beautiful architecture. The department of France with the smallest number of total overnight hotel stays in France is Ariège (with only 1,937,180), a largely rural department in the Occitanie region. Ariège has a population of only 153,000 (as of 2016), making it the 7th smallest department in France, and although it does have a few of its own historical monuments, it is not as attractive to tourists as most of the other areas of the country. When we compare these two values (287,083,100 and 1,937,180) we can see that the number for Paris is nearly 150 times larger than that for Ariège, which not only shows how large the difference in numbers of tourists there are across France, but also that Benford's Law should be much more evident in this data than in the previous data.

This graph clearly shows Benford's Law in action, although since the data does not span multiple orders of magnitude it is still not perfect.


Premiers chiffres du nombre total de nuitées dans les hôtels entre 2011 et 2018 dans tous les départements de la France[11]

Ce graphique démontre clairement la Loi de Benford, malgré une base de données limitée et donc imparfaites. Cependant, je trouve toujours cet exemple particulièrement efficace, car il n'y aurait aucune raison évidente de penser que 1 serait si souvent le premier chiffre significatif de ces données.

Maintenant que nous avons examiné quelques exemples plus ou moins fort de cette loi, nous pouvons déterminer comment appliquer la Loi de Benford. Bien que cette loi ait été simplement une découverte intéressante, la Loi de Benford a été utilisée plusieurs fois dans la pratique et en particulier dans le domaine financier. Par exemple, après avoir découvert que la Grèce modifiait les données financières qu'elle envoyait

However, I still find this example particularly effective as there would be no obvious reason to think that one would be the leading digit of this data so often.

Now that we have looked at a couple of compelling examples, and another example where the law was not as obvious, we have established when Benford's Law is most apparent. Although this law does seem to merely be an interesting discovery, Benford's Law has been used a number of times practically, especially when dealing with finance. For example, after it was discovered that Greece had been altering its financial returns that they were sending to the EU, some mathematicians wrote a paper examining Greece's financial records à l'UE, des mathématiciens ont écrit un article examinant plus en détail les états financiers de la Grèce en utilisant la Loi de Benford, afin d'essayer de déterminer depuis combien de temps la Grèce dénaturait ses chiffres. L'un des aspects importants de la Loi de Benford est qu'elle a besoin de données naturelles ; la manipulation des statistiques et des chiffres influence le résultat de la loi Bendford ou peut même de totalement en fausser le résultat final. À la fin des années 2000, les chiffres présentés par la Grèce à l'Union européenne étaient suspects, mais en 2011, quatre mathématiciens et économistes allemands ont rédigé un document examinant tous les rapports économiques de la Grèce avant l'an 2000[F]. Le document a révélé que les données de la Grèce présentaient la plus grande variation statistique de distribution parmi tous les pays de l'UE (la Belgique, la Roumanie et la Lettonie avaient également une grande variation dans la distribution attendue, alors que l'Espagne, le Portugal et l'Italie avaient tous une distribution presque parfaite dans l'analyse de leurs données). Le document a également révélé que l'année où les données de la Grèce étaient les plus éloignées de la Loi de Benford était l'année 2000, peu avant l'adhésion de la Grèce à l'UE. Malheureusement, l'UE n'a pas donné suite aux découvertes de ce document, ce qui soulève la question importante de savoir si on peut utiliser la Loi de Benford pour vérifier des données. Bien entendu, la distribution de Benford est basée sur la probabilité ; la probabilité que le premier chiffre d'un nombre dans les données de la Grèce soit 1 est de 0,3. Cela ne signifie pas nécessairement que si l'on a moins de 30% de chiffres significatifs qui soient 1, les données sont fausses. Cependant il est peu probable que sur plusieurs années les données ne correspondent pas à la Loi Benford a moins d'avoir été falsifiées. Pour cette raison, la Loi de Benford a été utilisée comme preuve dans des affaires précédentes ; dans la campagne électorale iranienne de 2009 (les votes avaient été falsifiés et le vainqueur supposé avait truqué les votes) la fraude a été détectée grâce à la Loi de Benford[G] (les chiffres significatifs du nombre de votes pour chaque candidat dans chaque région d'Iran avaient été analysés), bien que cela n'ait finalement eu aucun effet sur le résultat final des élections. De plus, aux États-Unis, la Loi Benford est utilisée comme preuve pour détecter la fraude fiscale dans les déclarations de revenus[H]. Tout cela nous montre que la Loi de Benford a sa place dans le système juridique, mais je pense que cette loi étant basée sur les probabilités, la plupart des gens y voient plutôt une possibilité de fraude / altération des données, plutôt qu'une preuve solide, ce qui signifie qu'elle n'est pas utilisée globalement en tant que preuve.

RÉSUMÉ

Pour résumer, nous avons constaté que la Loi de Benford a besoin de données couvrant plusieurs ordres de grandeur et comportant de nombreux points de références. Nous avons également vu pourquoi la distribution se base sur un modèle logarithmique et mathématique. Et avons examiné divers exemples de données illustrant la distribution de Benford.

Enfin, nous avons analysés l'utilisation de la Loi Benford dans le monde réel et si elle devait être plus utilisée comme preuve de fraude. En écrivant cet essai, j'ai personnellement beaucoup appris sur les mathématiques, mais aussi sur la France. J'ai aussi appris beaucoup de nouveau mots français, en particulier concernant les probabilités mathématiques et statistiques, ce qui faisait partie du défi intéressant d'écrire cet essai en français.

more closely using Benford's Law, in order to try and work out how long Greece had been misrepresenting its numbers. One of the important aspects of Benford's Law is that it is a naturally occurring phenomenon; editing statistics and numbers is likely to result in Benford's Law not being as evident in the data, or it could even result in Benford's Law not being evident at all. During the late 2000s Greece's numbers being submitted to the EU were found to be suspicious, but in 2011 four German mathematicians and economists wrote a paper examining all of Greece's economic reports since before the year 2000[F]. The paper found that Greece's data had the largest variation from Benford's distribution out of any country in the EU (Belgium, Romania and Latvia also had a large variation from the expected distribution, whereas Spain, Portugal and Italy all had nearly perfect distributions in their data). The paper also found that the year when Greece's data was furthest from Benford's Law was 2000, soon before Greece joined the EU. Unfortunately, the EU has not acted on the discoveries made by this paper, which raises the important question about whether Benford's Law can or should be used to verify data. Of course, the Benford distribution is based on probability; the probability that the leading digit of any number in Greece's data is a one is 0.3, but it does not necessarily mean that if less than 30% of the leading digits of the data are ones that the data is fake, however it would be very unlikely for data to not follow Benford's Law multiple years in a row if it had not been tampered with. In fact, Benford's Law has been used as evidence in cases before: in the 2009 Iranian election voter fraud (the votes were tampered with and the supposed winner had actually rigged the votes) was detected using Benford's Law[G] (the leading digits of the number of votes for each candidate in each region of Iran were analysed), although this ultimately had no effect on the outcome of the election. Furthermore, in the US Benford's Law can be used as evidence and is used to detect fraud in tax returns[H]. This all shows us that Benford's Law has a place in the legal system, however I believe that due to the law being based on probability most people view it as a suggestion of fraud/tampered data, rather than solid evidence, meaning it is not used globally as a form of evidence.

SUMMARY

To summarise, we have seen that Benford's Law is most evident in data that spans multiple orders of magnitude and that has many data points. We have also seen why the distribution must show a logarithmic pattern, and we have demonstrated the maths behind it, and looked at various examples of data to see whether they show Benford's distribution.

Finally, we discussed Benford's Law's uses in the real world and whether or not it should be more heavily relied on as evidence. In writing this essay I personally learnt a great deal about not just the maths involved, but also about France as a country. I also learnt much new French vocabulary, particularly relating to mathematical and statistical probability, which has been part of the interesting challenge of writing this paper in French.

Three neurobiological routes to religious experience

Samuel Cherry

ABSTRACT

This paper seeks to collate and understand various neuroscientific explanations of religious experiences, with a focus on three pathways: the consumption of psychedelic drugs, temporal lobe epilepsy and psychotic conditions (such as schizophrenia). An attempt has been made to identify the neural correlates of religious experiences, understand how these neural states might result in a religious experience and see if there are any common pathways between causes. Current hypotheses of neurobiological pathways of religious experience, including the Limbic Marker Hypothesis and the Temporal Lobe Hypothesis are analysed through the three routes. A new, unified hypothesis of the neurobiological basis of religious experiences is presented alongside a brief discussion of what constitutes a religious experience and the possible theological implications of these hypotheses.

INTRODUCTION: WHAT IS A RELIGIOUS EXPERIENCE?

To understand which neural states are correlated with religious experiences, we must first understand what a religious experience is[1]. Various attempts have been made by philosophers, theologians, mystics, researchers and clinicians to define, qualitatively or quantitatively, what constitutes a religious experience. The purpose of this section is to explore two philosophical/theological attempts to characterise religious experiences, and see how these characteristics are incorporated into various scales or questionnaires used in research.

The most well-known attempt is perhaps William James' in his work *The Varieties of Religious Experience* (1). James defines religious experiences as being ineffable (indescribable), noetic (insightful or revealing truth), transient (being impermanent - coming and then going) and passive (happening to an individual, uncaused and uncontrollable). The qualities identified by James are also used at least in part in the Revised Mystical Experiences Questionnaire, an internally reliable[2] 30-item scale designed for measuring 'hallucinogen-occasioned mystical experiences'; questions are asked on factors such as ineffability, transcendence of space and time (this might be likened to the otherness of the experience identified by James) and feelings of sacredness and/or a noetic quality (2). This questionnaire allows researchers to attempt to quantitatively measure the intensity and predominant features (via using four separate sub-categories) of a religious experience.

In a similar vein to the Revised Mystical Experiences Questionnaire is the INSPIRIT (Index of Core Spiritual Experiences) scale, another attempt to quantitatively measure the intensity of a religious experience (3). INSPIRIT is made up of eleven questions, three of which measure the qualities of specific religious experiences. These include: '[feeling] very close to a powerful spiritual force'; overwhelming experiences of love, 'profound inner peace' or joy; miracles and Divine healing; 'a feeling of unity with the earth and all living beings'; encountering spiritual figures, angels or the deceased; and a near death or afterlife experience. The other questions assess 'behaviours and attitudes that would be present in a person who feels close to God', i.e. they provide measurements that could suggest whether or not the person was genuinely moved and convinced by their religious experience.

I think it is also important to consider the nature of a religious experience from the perspective of Rudolf Otto (4). His description is commonly paraphrased as 'mysterium tremendum et fascinans': 'mysterium' being something wholly other; 'tremendum' referring to something terrifying, majestic and overpowering; 'fascinans' meaning alluring or entrancing. Otto's religious experience is therefore something that is both repelling in its otherness and power, yet nonetheless attracting. This conceptualisation, I believe, is particularly useful for understanding why people who have religious experiences after ingesting psychedelic drugs often take them multiple times (5). This indicates an attraction (fascinans) to the subjective effects, as they are often taken expressly for the purpose of achieving a mystical experience (mysterium tremendum) (5). (This study looked at the use of psilocybin-containing mushrooms by college age students and found that while the mode number of uses was 1, the mean was 3.4, indicating that many individuals took the substance multiple times. The top three listed reasons for doing so were 'curiosity', 'to achieve a mystical experience' and 'introspection'.)

Otto also saw religious experiences as 'self-authenticating', i.e. an individual who has had a genuine religious experience is convinced of its truth immediately and overwhelmingly. If an individual is indeed convinced of the truth and significance of their experience, we might expect that the experience would influence their religious behaviour and perhaps reinforce their faith. Thus scales like INSPIRIT that seek to also measure the strength of religious convictions, or the Duke Religion Index (6) which measures the type and frequency of religious behaviour might also be useful in ascertaining the strength and significance to an individual of a religious experience.

We have thus seen how philosophical and theological views about what constitute a religious experience are incorporated, whether intentionally or not, into quantitative scales commonly used for studying religious experiences. Do we now have a scientific toolkit for investigating religious experiences which has a philosophical and theological grounding?

There is one key factor in 'genuine' religious experiences that cannot be overlooked: that they are caused or generated by an External Divine Being. Whether or not the events categorised as a 'religious experience' by the aforementioned questionnaire actually involve an External Divine Being cannot be empirically determined. Thus when we are discussing religious experiences 'caused' by psychedelic drugs or temporal lobe excitation etc. what we actually mean is that we are causing (or simulating) experiences that might be subjectively categorised as religious. It is impossible to know, either from the perspective of the person having the experience or from the external vantage point of the researcher, if any of these experiences are truly 'religious' in the sense that they are caused by the Divine; this significant epistemological limitation must be acknowledged.

Moreover, given the frequency of individuals reporting that such experiences are 'ineffable', it seems somewhat contradictory to attempt to create an index or questionnaire reliant on individuals being able to describe aspects of their phenomenology. However, the validity of such tools might be defended, from a pragmatic angle. Simply, these indices and questionnaires are the only way that we can even begin to estimate what a religious experience feels like, through what many individuals who describe their experience as ineffable will consider imperfect and imprecise language. Somewhat similarly, we should note that it is a widely held position in the philosophy of mind that we can never know what another person's subjective mental state is like simply from an external evidential analysis of their brain (7). Therefore, due to the inability of a neuroscientific analysis of the brain to shed light on the subjective phenomenology of a religious experience, and despite the major limitation of the self-reported ineffability of many of the experiences, indices and questionnaires are all we have to work with to attempt to understand evidentially what it is like to have a religious experience.

To conclude, a religious experience is a subjective psychological state which can manifest itself in various ways, but is often characterised by feelings of profundity, ineffability, sacredness and the presence of an overwhelming force. For a religious person, a genuine religious experience must be caused by The Divine. There are widely used indices and questionnaires that seek to characterise, both qualitatively and quantitatively, the subjective nature of experiences characterised as 'religious', but they have some significant limitations. Moreover, it is impossible to tell if any experience is genuinely the product of an External Divine Being, meaning that at best all we are able to do is investigate experiences that are characterised as religious, without knowing if they actually are religious experiences. When discussing the neuroscience of these experiences, we must recognise these substantial theological limitations.

1. PSYCHEDELIC DRUGS:

Throughout the globe, various different cultures have and continue to use psychedelic drugs to cause religious experiences. In these settings, these substances are often referred to by researchers as 'entheogens'. Notable examples include the traditional use of peyote (8) (containing mescaline (9)), and psilocybin[3]-containing mushrooms (10) in Native American tribes, and the use of ayahuasca (containing DMT and MAOIs that allow the DMT to be orally active (11)) by tribes in the Amazonian basin (12). There are also reports of the use of various psychedelic drugs within Western subcultures for the purpose of producing profound and often mystical states (5) (13).

Recent research into the effects of consuming synthetic psilocybin has found that it can reliably produce 'mystical' or 'spiritual' experiences - i.e. experiences that are ineffable, transcend space and time, have the qualities of tranquillity, ecstasy, amazement, profound sacredness or holiness, and a feeling of encountering 'ultimate reality', 'eternity', 'infinity', 'oneness' and 'unity… with what was felt to be greater than your personal self' (14). (This result was determined using the aforementioned Revised Mystical Experiences Questionnaire.)

All these drugs, like many other psychedelics, function as $5HT_{2A}$ receptor agonists – i.e. they activate $5HT_{2A}$ receptors (15) (16) (17). $5HT_{2A}$ receptors are a subtype of serotonin receptor, and their activation is thought to be responsible for the effects of psychedelic or hallucinogenic drugs (18), and possibly also the psychotic symptoms in individuals with conditions like schizophrenia or bipolar disorder (evidenced by the efficacy of $5HT_{2A}$ antagonist drugs as antipsychotics (19)).

The dopamine hypothesis of psychosis posits that psychotic symptoms are the result of dysregulated dopamine release, primarily the over-activation of D_2 receptor (a subtype of dopamine receptor) (20). It is also hypothesised that $5HT_{2A}$ receptors modulate dopamine release, as $5HT_{2A}$ receptor activation has been shown to result in the excitation of D_2 receptors in the ventral tegmental area and medial prefrontal cortex (21). A further explanation and discussion of how abnormal dopamine release may lead to religious experiences is included in the psychosis section.

$5HT_{2A}$ agonism has also been shown to increase spontaneous glutamatergic activity in the prefrontal cortex (22). While abnormalities in glutamate levels in the brain are thought to play a role in schizophrenic symptoms (23) the evidence for elevated glutamatergic transmission resulting in psychosis is mixed: most studies show no link between elevated regional glutamate levels and positive schizophrenic symptoms (like delusions and hallucinations) (24). Generally, the evidence points to schizophrenic symptoms (both positive and negative) being associated with decreased glutamate levels (23) (25). This being said, it has been also been observed that some individuals with treatment resistant schizophrenia have elevated glutamate levels (24); $5HT_{2A}$ mediated increases in glutamatergic activity in the visual cortex may result in the visual hallucinations observed in individuals who have taken psychedelic drugs (26). The precise role of glutamate function in the aetiology of psychosis is not fully understood, and requires further research (25). That being said, it is possible that glutamatergic modulation as a result of $5HT_{2A}$ agonism might be responsible for some of the subjective effects of psychedelic drugs which manifest as religious experiences.

It has been hypothesised that some of the subjective and behavioural effects of psychedelic drugs are due to an alteration in cortical and prefrontal cortex function (22). Muramoto's hypothesis of the role of the medial prefrontal cortex in human religious activity provides a possible explanation of how this might be possible (27). It has been noted that metabolic activity in the posterior superior parietal lobe decreases during meditation. (This was determined using SPECT[4] scanning of regional cerebral blood flow in experienced practitioners of Tibetan Buddhist meditation (28).) Decreased metabolism in the posterior superior parietal lobe causes an increase in metabolism in the medial prefrontal cortex. It is hypothesised by Muramoto that, inversely, excitation of the medial prefrontal cortex would decrease metabolism in the posterior superior parietal lobe, indicating its deafferation (loss of sensory input). As the posterior superior parietal lobe is implicated in the perception of spatial relations, including the relation of the self to the external world, deafferation of the posterior superior parietal lobe could explain the subjective effects of 'transcendent' or 'out of body' experiences commonly seen in religious experiences caused by psychedelic drugs. This loss of perception of the relationship between the self and the external world may be interpreted by the individual as 'merging' with an external divine being.

Muramoto admits that his hypothesis is limited by the fact that there are no reports of individuals with parietal lobe lesions experiencing these effects (27), suggesting that the impairment of the posterior superior parietal lobe is not the only factor at play. More research is also needed to determine if the excitation of the medial prefrontal cortex does indeed deafferentiate the posterior superior parietal lobe. As it stands Muramoto's hypothesis, in combination with reports of increased glutamatergic activity in the prefrontal cortex from $5HT_{2A}$ agonism, could be a compelling explanation of some of the subjective effects of psychedelic drugs and their ability to induce religious experiences, but there is simply not enough evidence at this point to confidently accept this explanation.

In conclusion, the ability of psychedelic drugs to produce, at the very least, subjective effects that simulate religious experiences is reliable, and seems to be due to modulation of dopaminergic activity. The effect of psychedelics on glutamatergic activity in the prefrontal cortex may also be relevant, but more evidence is needed to confirm this hypothesis.

2. TEMPORAL LOBE EPILEPSY:

Temporal lobe epilepsy is a condition defined by recurrent focal seizures that start in the temporal lobe. It has been suggested that seizures or micro-seizures in the temporal lobe may be the cause of religious experiences (29), including very theologically important ones such as the appearance of Christ to Saul on the road to Damascus (30), or the visions of the Prophet Mohammed (31). (We should however be somewhat sceptical of attempts to retroactively diagnose individuals with medical conditions based on their subjectively reported symptoms.)

The link between temporal lobe epilepsy and religiosity (i.e. the quality of being religious) has been noted before. A 2015 study found that having mesial temporal lobe epilepsy with hippocampal sclerosis[5] was predictive of high organisational religiosity (i.e. regularly attending formal group religious services) (32). It also found that only mesial temporal lobe epilepsy with hippocampal sclerosis was predictive of high intrinsic religiosity (i.e. subjective feelings of the presence of the Divine, alongside religion being seen a significant influence in an individual's life). mesial temporal lobe epilepsy with hippocampal sclerosis was associated with higher non-organisational religiosity (i.e. regular private prayer, meditation or scriptural study) (32). These factors were measured using the Duke Religion Index, and the values for individuals with mesial temporal lobe epilepsy and hippocampal sclerosis were compared against non-epileptic controls. It should be noted that none of patients in the 2015 study had interictal[6] psychosis, suggesting that the link between religiosity and mesial temporal lobe epilepsy with hippocampal sclerosis was not due to any sort of religious delusion or hallucinations. That being said, evidence suggests that 'hyper-religious' temporal lobe epilepsy patients are much more likely to have postictal psychosis, though which precedes the other is unknown (33). Overall, this evidence does seem to suggest that intrinsic religiosity, both organisational and non-organisational, is higher in patients with abnormal temporal lobe function, compared to the general public. It could therefore be reasonably posited that temporal lobe function has some effect on religious attitudes, behaviour and belief – one which may or may not influence religious experiences.

Somewhat unexpectedly, many studies that measure the percentage of individuals with temporal lobe epilepsy who have also had religious experiences have consistently reported a far lower rate of incidence compared with the wider public (1-2.3% versus c.20-60%) (34). It should however be noted that one study, which compared the intensity (using the INSPIRIT scale) of religious experiences between a group of patients with temporal lobe epilepsy and non-epileptic regular churchgoers, found that the intensity of the religious experience was greater in the temporal lobe epilepsy group (33). These experiences were characterised by 'awareness of an external being, evil or great spiritual presence, feelings of death and dying and overwhelming fear'. The same study also interestingly found those patients incorporated elements of their religious experiences into their religious beliefs (33), indicating that the patients thought that their experiences had genuinely religious validity and significance.

Attempts to induce religious experiences by the external stimulation of an individual's temporal lobes have had mixed results. Persinger et al. reported successfully being able to induce religious experiences (either 'out of body' experiences or the 'sensed presence' of 'a "Sentient Being" that was beyond the experiment but associated with such personal significance and relevance that emotional responses were common') in individuals via 'milligaus intensity extremely low frequency magnetic stimulation' (35), using a device colloquially known as the 'God Helmet'. Their results however have not stood up to peer review, with other researchers failing to replicate results (36). Moreover, there is insufficient evidence for the how milligaus intensity extremely low frequency magnetic stimulation can actually interact with the temporal lobe, with it being argued that the magnetic field strength used in Persinger's study (5000 times weaker than the magnetic fields used in transcranial magnetic stimulation[7]) is too weak to have any effect on the brain (36). However, the failure of the 'God Helmet' to induce religious experiences does not suggest that temporal lobe is not involved in religious experiences, rather that the 'God Helmet' is unable to induce them.

The controversy surrounding failed attempts to induce religious experiences with the 'God Helmet' have undermined Persinger's position, and his hypothesis has subsequently been the subject of considerable criticism (34) (38). Researchers who attempted to replicate his study with proper blinding failed to obtain similar results, finding instead that reporting a 'religious experience' was far more likely to be predicted by the participant's suggestibility, rather than whether the God Helmet was switched on (36). It should be noted that there were disputes between Persinger and the authors of the blinded replication study over the similarities of their method (34). Further research is therefore required to explain fully what role the temporal lobes have in religious experiences, though there is a lack of compelling evidence for milligaus intensity extremely low frequency magnetic stimulation as the primary cause of these experiences over placebo, given its practical implausibility and the conflicting results from studies using 'God Helmets'.

But how might altered temporal lobe function induce a religious experience? In his 1983 paper *Religious and Mystical Experiences as Artifacts [sic.] of Temporal Lobe Function: A General Hypothesis*, Persinger lays out an explanation of how various defining features of religious experiences might be caused by dysfunction in parts of the temporal lobes, in particular the amygdala and the hippocampus (29). Emotional aspects are attributed to stimulation of the amygdala, as the amygdala is thought to be responsible for the affect (pleasure versus pain) of an experience, as are out of body experiences. Time and spatial distortion, features that might be interpreted by the individual as 'transcendence', could be due to altered hippocampal regulation of memories (making old memories appear real and immediate). Disruptions of the temporal lobe in serious cases are caused by temporal lobe epilepsy, but Persinger also postulates that they can be caused by weak magnetic fields from tectonic activity. Subsequent studies have found that psychedelic drugs, like LSD (which functions as a $5HT_{2A}$ agonist (18)) and mescaline increase activity in the medial temporal lobes (37), potentially suggesting that the mechanism by which psychedelic drugs produce religious experiences has some involvement with the temporal lobes.

Another important fact to consider in the relationship between temporal lobe epilepsy and religious experiences is postictal or interictal[8] psychosis. It has been noted that in certain forms of epilepsy, particularly temporal lobe epilepsy, schizophrenic-like psychosis can occur (39). The prevalence rate of postictal psychosis in patients with temporal lobe epilepsy is thought to be around 7% (40), which, while uncommon, is far higher than the rate in the general public at around 1% (39). Somewhat expectedly, one study found that the percentage of temporal lobe epilepsy patients who have had religious experiences is much higher in those who also have postictal psychosis (at 27.3%) compared to those who do not experience psychosis (at 1.3%)[9] (41). The incidence rate of religious experiences in the postictal psychosis cohort may or may not be greater than that of the general public, at around 20-60% (34).

It is currently thought that postictal psychosis is not caused by epileptic seizures, but rather that both the epileptic and the psychotic condition share a common neurophysiological cause (39). These psychotic episodes are responsive to antipsychotic medication (39), suggesting that postictal psychosis also has at its basis as a disruption of dopaminergic pathways. This was further affirmed by a study that showed increased L-DOPA[10] metabolism in patients with mesial temporal lobe epilepsy and psychosis, also suggesting increased dopaminergic activity is the underlying pathological cause (42). It has been hypothesised that this increase in dopamine levels is due to the upregulation of enzymes responsible for dopamine metabolism in response to abnormal dopamine regulation by the amygdala and hippocampus (42), two areas of the brain where activity is increased during psychotic episodes in patients with temporal lobe epilepsy (43). An exploration of how psychosis might cause religious experiences is included in the psychosis section.

Overall, while a hypothesis based around temporal lobe function as a neurobiological basis for religious experience on its own has not been sufficiently validated, it is not entirely without merit. The psychotic condition associated with temporal lobe epilepsy does seem to plausibly lead to religious experiences through its effect on dopaminergic activity. However, the precise link between aberrant behaviour in the temporal lobe leading to dopaminergic dysfunction, given the low percentage of patients with postictal psychosis, warrants further investigation, as does the seemingly lower proportion of individuals with temporal lobe epilepsy having religious experiences compared to the general public. There is some evidence, and there are some plausible hypotheses, to suggest that the temporal lobe has a role to play in religious experiences, but the conclusions as to what its specific role is remain unclear.

3. PSYCHOSIS:

Psychotic disorders, like schizophrenia, are characterised by delusions and auditory and visual hallucinations, among other symptoms (44). Though the DSM-5[11] notes that delusions may be of a religious nature, it also notes that 'hallucinations may be a normal part of religious experience in certain cultural contexts' (44). From a clinical standpoint, it makes sense to exclude transient hallucinations or pseudo-hallucinations that might be culturally normal or expected (the example given in the DSM-5 is of a patient reporting hearing God's voice at a religious service), as they likely do not meet the burden of causing 'significant distress or disability', and therefore don't meet the necessary DSM-5 guidelines to constitute a mental disorder. Nonetheless, for the purpose of this section, whether or not religiously-themed delusions or hallucinations meet the diagnostic criteria for psychotic disorders from the point of view of a clinician is irrelevant. It is clear that most religious experiences meet the DSM-5 criteria of hallucinations: they lack an external stimulus, are 'vivid and clear', involuntary and 'perceived as distinct from the individual's own thoughts' (44). The focus is on understanding the possible aetiology of religious experiences though the neurobiological pathways of psychosis, rather than any question of clinical implication or relevancy, so it is acceptable to treat religious experiences as having the same aetiology as other psychotic states.

It is unclear how many people with psychotic conditions have religious experiences, but one study found that 24% of hospitalised schizophrenics had delusions of a religious nature (45). (This may or may not be a higher rate of incidence than in the general public, whose rate of incidence is between 20-60% (34).) Religiously themed delusions in psychotic individuals are found across all cultures (46). These two pieces of epidemiological evidence suggest that there is a potentially biological link between psychosis and religious delusions. However cultural factors may be relevant in the variation of religious experiences through history and between social groups; it has been noted that, in an increasingly secular world, the incidence of religiously themed delusions has appeared to decrease (47). The specific role of religious belief in the development of psychosis is unclear, with different studies suggesting that religious people develop psychosis at lower, higher or similar rates to non-religious people (47). Therefore, while the precise link between culture and religious psychosis is unclear, we should not assume it to be negligible – more research is needed in this field.

As has been noted earlier in this paper, both $5HT_{2A}$ agonists and temporal lobe epilepsy can seem to cause psychotic states that resemble those of schizophrenia. A full explanation of the effects of these two conditions on religious experiences must also include an explanation of how psychosis might result in religious experiences.

The most widely accepted hypothesis for the cause of psychotic symptoms is the dopamine hypothesis of psychosis (also known as the dopamine hypothesis of schizophrenia) (20) (23). The hypothesis posits that the delusions and hallucinations seen in psychotic conditions is a result of over-activity at D_2 receptors. Of particular importance is the disruption of mesolimbic dopaminergic pathways.

The mesolimbic pathway is a dopaminergic system made up of neuronal cells that project from the ventral tegmental area in the midbrain to the ventral striatum in the forebrain (48). It is part of the larger limbic system, which is thought to mediate emotional, behavioural, memory and learning func-tions (48). Hyperactivity of D_2 receptors in the ventral striatum from the ventral tegmental area is thought to be the cause of auditory hallucinations and delusions (26), two key features of psychosis. Lesions in the limbic systems of animal are thought to impair the ability to filter out multiple stimuli (48), which might explain the aetiology of psychosis as a sort of sensory overload.

As we saw in the section on psychedelic drugs, $5HT_{2A}$ agonists are thought to increase dopamine release in the ventral tegmental area. The dopamine hypothesis of psychosis can therefore explain why $5HT_{2A}$ agonists cause individuals to experience subjective effects like auditory hallucinations or delusions.

Although $5HT_{2A}$ agonism also seems to increase glutamatergic activity, the NMDA[12] receptor hypofunction hypothesis of psychosis suggests that psychotic symptoms are caused by a decrease in glutamatergic activity at NMDA receptors, resulting in a subsequent increase in activity at D_2 receptors in the mesolimbic pathway (26). NMDA receptor hypofunction is thought to result in the hypofunction of GABAergic[13] inhibitory interneurons in the prefrontal cortex. The reduction in GABAergic inhibition results in an increase of glutamatergic activity in the ventral tegmental area, thus again resulting in increased activity in the limbic system.

We have seen how various different neural events might alter the limbic system and cause psychosis, but how can this lead to a religious experience? In their 1997 paper *Neural Substrates of Religious Experience*, Saver and Rabin propose the 'Limbic Marker Hypothesis' of religious experiences (49). The limbic system is responsible for both assigning positive or negative emotional values to stimuli and to determining the importance or significance of a new stimulus. Saver and Rabin hypothesise that over-activity of the limbic system could result in experiences being marked as 'crucially important or self-referent' (with regards to significance) and 'harmonious' or 'ecstatic' (with regards to emotional value).

Within social and cultural contexts, if this overwhelmingly emotional or profound experience is explained and interpreted within a religious framework, this can lead to a religious experience (this notion is called attribution theory (50)). Individuals who have had a profound, joyous and novel experience want to understand a cause for what they have experienced. Due to the 'otherness' of a religious experience, secular or naturalistic explanations often feel inadequate, so individuals are more likely to turn to a religious explanation. (My explanation here is partially based on the 1985 paper *A General Attribution Theory for Religion* by Spilka et al. (51).)

This is affirmed by the fact that 'set and setting' effects the interpretation of experiences from hallucinogens (49). As we have established prior, there is conflicting data as to whether people who are religious before epileptic activity are more likely to have a religious experience, which might suggest that attribution theory is incorrect. Similarly, there are several reported cases of individuals with temporal lobe epilepsy converting after religious experiences (52) (53). However, while these individuals might not have been religious prior, they almost certainly would have been exposed to religious ideas, and thus knowledge of these concepts could affect their interpretations of ecstatic or profound experiences. Attribution theory is thus not undermined by conversions or the conflicting data as to the effect of religious belief on the experiences of epileptics.

To summarise Dopaminergic or glutamatergic disruption resulting in limbic over-activity can explain why individuals with psychosis erroneously attribute such powerful emotions or significance to ordinary sensory inputs. Attribution theory can explain why these individuals, subject to certain cultural and social factors, might attribute these experiences to Divine source. Overall then, it is clear that all the routes to subjectively defined religious experiences that have been studied in this paper ultimately result in over-activity of limbic system, mediated by dopamine (in some cases among other neurobiological changes). Saver and Rabin's hypothesis, in combination with attribution theory, provide a reasonably compelling explanation as to how limbic over-activity is able to result in an experience that can be characterised as religious.

DISCUSSION AND CONCLUSION:

NEW UNIFIED HYPOTHESIS FOR RELIGIOUS EXPERIENCES:

It would seem clear from the evidence laid out that there are some strong and consistent neural correlates of religious experiences, and several hypotheses to explain how these correlates might result in a religious experience.

The hypothesis proposed by Saver and Rabin in their 1997 paper *Neural Substrates of Religious Experience* suggests that religious experiences are the result of the imposition of religious beliefs onto ordinary sensations, facilitated by abnormal sensory interpretation by the limbic system (49). Psychosis, a result of $5HT_{2A}$ agonism from psychedelic drugs, or some forms of temporal lobe epilepsy, or another genetic or developmental psychiatric disorder like schizophrenia, can ultimately modulate the limbic system and result in erroneous sensory processing. Subsequently using attribution theory, we can explain why these experiences of profundity or ecstasy are likely to be interpreted as religious. We thus have a unifying hypothesis of the neurobiological basis of religious experience for all the three routes that have been analysed.

Persinger's hypothesis, focused entirely on temporal lobe function, has merit in its ability to explain how dysfunction in parts of the temporal lobe might lead to religious experience, but is not sufficiently backed by the evidence to be accepted. The temporal lobe may indeed be implicated in religious experiences, but the precise pathway is not yet understood.

I would therefore like to posit a new unified hypothesis of the neurological basis of religious experiences, based around abnormal neurological functioning as a result of aberrant glutamatergic and dopaminergic activity in specific parts of the brain producing a set of subjective effects that are interpreted by the individual as having a Divine origin. This hypothesis is summarised in Fig.1.

There are some limitations to this hypothesis. First, the link between temporal lobe epilepsy and elevated dopamine levels, resulting in a psychotic disorder, is not established. Further research is needed to

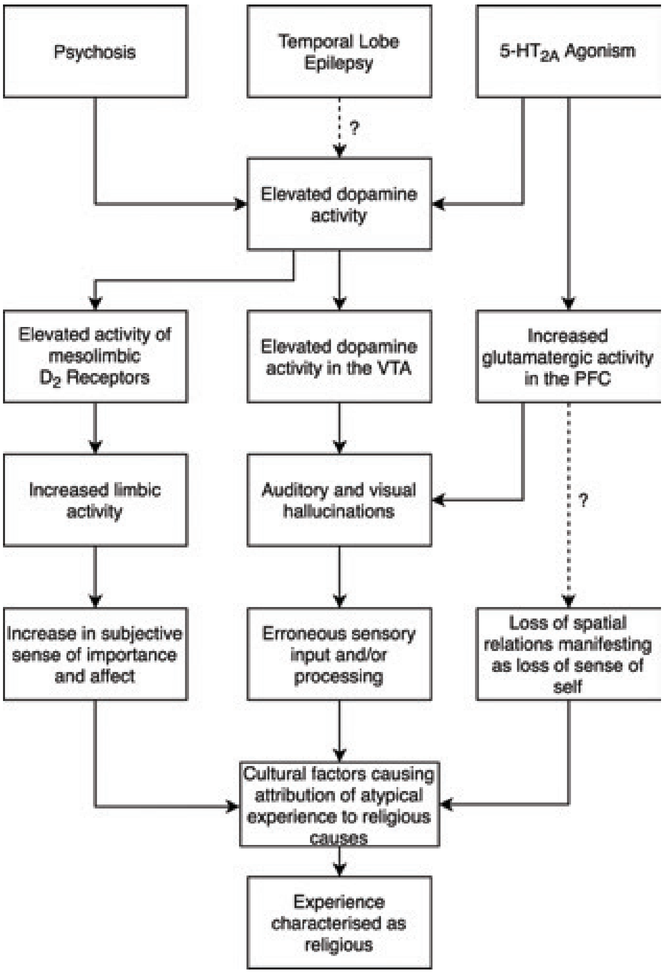

Psychosis
Temporal Lobe Epilepsy
5-HT2A Agonism
?
Elevated dopamine activity
Elevated activity of mesolimbic D2 Receptors
Elevated dopamine activity in the VTA
Increased glutamatergic activity in the PFC
Increased limbic activity
Auditory and visual hallucinations
?
Increase in subjective sense of importance and affect
Erroneous sensory input and/or processing
Loss of spatial relations manifesting as loss of sense of self
Cultural factors causing attribution of atypical experience to religious causes
Experience characterised as religious

Figure 1: Flowchart to show a new, unified hypothesis of the neurological basis of religious experiences.

explain both postictal psychosis, and the unusual nature of non-psychotic religious experiences in people with temporal lobe epilepsy (i.e. that they seem to occur in a smaller percentage of individuals, but are stronger). Similarly, the link between increased glutamatergic activity in the PFC and changes in spatial reasoning processing is simply a postulation.

Second, this hypothesis does not currently explain other possible causes of religious experiences such as mediation, prayer or breath-work, though they are not necessarily incompatible with an updated version of this model.

PHILOSOPHICAL PROBLEMS AND IMPLICATIONS:

It would be remiss not to include at least a brief discussion on some potential philosophical and theological implications of the suggestion that religious experiences have a basis in neural activity. While such a position could be seen as an attempt to provide a naturalistic explanation for religious experiences, i.e. one that removes the need for a God or Divine Being to cause these phenomena, to suggest this would require taking evidence further than we are able to. At best, we can only ever correlate neural states to subjective religious experiences, even if the seeming cause of those neural states is from a man-made object, like a synthetic hallucinogenic drug. Direct causation cannot be demonstrated. It would be coherent, from a theistic point of view, to posit that God communicates with us through $5HT_{2A}$ or D_2 receptor agonism, or through aberrant electrical behaviour in the temporal lobe. Thus an acceptance of a neurobiological basis for understanding religious experiences does not invalidate the theological truth or significance of them, and is subsequently not at odds with holding a theistic or spiritual worldview.

I am therefore inclined to agree with Felicity Ng, who writes in her paper *The Interface Between Religion and Psychosis*: ‘In our increasingly secular world, the complexity of religion may be compromised by reductionistic trends. In psychiatry, this may be apparent in excessive pathologising of religiosity… such knowledge should… deepen rather than restrict the forms of our conceptualisation of religion by helping to unveil the neurological mechanism that allow humans the capacity to have religious and spiritual experiences… It neither negates the existence of Divinity nor decreases the importance of faith…’ (47).

There are other factors that must be considered, particularly regarding the nature of how religious experiences are frequently studied neuro-scientifically. The environmental conditions under which ‘natural’ religious experiences occur are evidently not comparable to being in an fMRI or SPECT scanner. There are always going to be uncontrolled variables between religious experiences in a laboratory setting and a non-laboratory setting. I cannot currently see how it might be possible to conduct a study into the neural correlates of religious experiences where the investigatory method is sufficiently controlled for through blinding, without practical problems or ethical concerns. While this does not invalidate the current evidence, especially if the issue is unavoidable, it is a limitation that must be acknowledged.

Moreover, it would be overly reductive to conclude that genuine religious experiences are ‘caused’ exclusively by these conditions, or that they can be induced through material processes in laboratory conditions. As discussed in the introduction, the defining aspect of a religious experience for a religious person is that it is genuinely caused by God. For any of these ‘induced’ religious experiences, such as those caused by taking synthetic psilocybin, there are three distinct possibilities. First, that they are not genuine religious experiences (i.e. ones caused by the Divine), as they are just products of the brain, and so are only experiences that simulate genuine religious experiences, even if they are phenomenologically similar or even identical to genuine religious experiences. (A variant on this position would also be acceptable to the atheist, who, though accepting of the fact that some people have experiences characterised as religious, would not believe that these experiences have their genus in an External Divine Being, or God.) Second, that some sort of change in neural activity facilitates the causation of a genuine religious experience by the Divine (such a pathway would be metaphysical, and so it is not worth attempting to postulate an example in a scientific paper), or vice versa, whereby the Divine being causes a change in the neural state of the individual that causes them to have a religious experience. Finally, it could be posited that there is no real causal link between the change in neural activity observed during religious experiences and the experience itself, with a Divinely caused experience simply correlating to such a neural state. The verification or falsification of all of these positions, however, is beyond the realms of science, as it necessarily regards some mechanism of action between the Transcendent and Immaterial Divine (as usually conceived), and the material brain, or indeed an immaterial mind.

This question will therefore never be solved by neuroscience; it must be left to the philosophers and theologians. For my own part, as a philosopher and theologian, I am not particularly worried about the specificities of the mechanism between an External Divine Being and a religious experience for two reasons.

First, I don't think that it is possible to determine, either through the subjective phenomenology of the experience or through a neuroscientific analysis, whether or not the experience was genuinely caused by an External Divine Being. It is a basic premise of philosophy (since Descartes) to recognise that our minds can deceive us even if we are convinced of the nature of an experience (54). Contrary to what Rudolf Otto might posit, I do not think that any experience can be 'self-authenticating'. Abandoning phenomenological analysis to determine the validity of a religious experience, we are left with material neuroscience, which cannot analyse the behaviour or influence purportedly Immaterial Thing. Thus it seems impossible epistemologically to establish whether or not a religious experience is genuine.

Second, I think that, given the impossibility of determining which of the pathways is true over the others, the one that people are most convinced by is usually a reflection of their pre-existing worldview, rather than the product of impartial rational process. The atheist will prefer the explanation that does away with God as the cause of 'religious experiences', whereas a religious person convinced by the importance of a soul might reject the notion that there is any material basis for their religious experience, and claim that it is entirely caused immaterially by God. Moreover, it is entirely consistent to move between different explanations based on one's religious beliefs: e.g. a Christian might claim that a person who claims to have seen Christ has had a genuine, Divinely caused experience, but dismiss the religious experiences of Hindu or a Muslim as simply erroneous products of their mind.

For these reasons, it is clear to me that establishing the actual role of God in religious experience, although obviously of huge theological significance, is epistemologically difficult, if not impossible, and is certainly beyond the realm of current neuroscience. Deepening our understanding of the neuroscience of religious experiences therefore does not tie us to a particular theological worldview, theistic or atheistic.

CODA:

Overall, it is clear that religious experiences are a rich area for research to better understand not just the neuroscience of disorders in consciousness and perception, but also to investigate the link between philosophy, neuroscience and aspects of theology in an attempt to holistically understand the human brain and mind.

‘Creativity is seeing what everyone else sees,
but then thinking a new thought
that has never been thought before
and expressing it somehow.’

Neil deGrasse Tyson

www rgs-guildford co.uk

@RGSGuildford

RGS Guildford

Registered Charity No. 312028